LIFE IN THE COSMOS

From Biosignatures to Technosignatures

Manasvi Lingam | Avi Loeb

Harvard University Press

Cambridge, Massachusetts | London, England 2021

Printed in the United States of America

First printing

Cover design by Jill Breitbarth
Cover photo: sdecoret / Shutterstock

9780674259942 (EPUB)
9780674259959 (PDF)

The Library of Congress has Cataloged the Printed Edition as Follows:

Names: Lingam, Manasvi, 1987– author. | Loeb, Avi, author.
Title: Life in the cosmos : from biosignatures to technosignatures /
Manasvi Lingam and Avi Loeb.
Description: Cambridge, Massachusetts : Harvard University Press, 2021. |
Includes bibliographical references and index.
Identifiers: LCCN 2020048895 | ISBN 9780674987579 (cloth)
Subjects: LCSH: Exobiology. | Life on other planets. |
Life—Origin. | Habitable planets.
Classification: LCC QH326 .L564 2021 | DDC 576.8/39—dc23
LC record available at https://lccn.loc.gov/2020048895

To our cherished families, friends, and the myriad denizens
weaving their lives amidst the *Sternenzelt* of the Cosmos

CONTENTS

PREFACE

Over the span of human history, the prevailing dogma was that the Earth comprised the center of the physical Universe and that human beings occupied a privileged place within the Cosmos. The Copernican revolution, with its distinct antecedents in ancient Greece, India, and the Islamic Golden Age, inter alia, ushered in a new era that gradually displaced the Earth from its privileged position in space and thus engendered due recognition of a vast (conceivably infinite) Universe populated by innumerable galaxies, stars, planets, and moons. This paradigm shift was chiefly driven by the development of sophisticated instruments and observational techniques that facilitated the empirical validation of the heliocentric model. In contrast, however, the enigma surrounding our presumed location at the center of the biological Universe—namely, the presence of extraterrestrial life, colloquially exemplified by the archetypal cliché "Are we alone?"—has hitherto lacked a definitive resolution due to humankind's limited scientific and technical capabilities.

In the pre-Copernican ages, visionaries such as Epicurus, Lucretius, Muhammad al-Baqir, Moses Maimonides, Fakhr al-Din al-Razi, Albertus Magnus, Hasdai Crescas, and Nicholas of Cusa (and countless others unjustly lost to history) envisioned a plurality of worlds teeming with multifarious lifeforms. In addition, the presence of life on other worlds was regarded as either possible, plausible, or indisputable by sundry religious traditions, including the major non-Abrahamic creeds of South Asia—to wit, Hinduism, Buddhism, Jainism, and Sikhism. Of the post-Copernican philosophers and scientists who advocated the existence of alien life, the

first and most renowned among them is arguably Giordano Bruno, who averred that other stars are similar to the Sun in terms of hosting planets with life. What unites this diverse group of thinkers, drawn from various continents and distinct periods of history, is their shared belief in extraterrestrial life. The word *belief* is rendered crucial in this context because there was no empirical evidence to support their conjectures at that stage; indeed, for many millennia, debates relating to life in the Cosmos were confined to theoretical musings for this very reason.

However, the tide is palpably changing, as we stand on the cusp of a new epoch when we might settle this question in a rigorous fashion. Over the past couple of decades, we have witnessed many pivotal breakthroughs in our understanding of the origin and evolution of life on Earth. This progress has been facilitated by great strides in realms as diverse as genetics, paleontology, and biogeochemistry, to name a few. In tandem, the discovery of planets beyond our Solar system (exoplanets) has witnessed spectacular and explosive growth since the 1990s, with the number of confirmed exoplanets having exceeded several thousands and continuing apace. It comes as no surprise, therefore, that one-half of the 2019 Nobel Prize in Physics was awarded to Michael Mayor and Didier Queloz "for the discovery of an exoplanet orbiting a solar-type star."[1] To these two broad spheres of knowledge, one must add the sustained exploration of the Solar system, which has yielded new wonders such as the discoveries of hydrothermal vents (potential sites for the origin of life) in the subsurface oceans of Saturn's small moon Enceladus; complex organic molecules on the surface of Mars and briny lakes in its subsurface; water ice in cold traps scattered across our Moon's surface; and the enigmatic traces of phosphine (a possible metabolic product of life) in the atmosphere of Venus.

Each exciting development has spawned fresh questions and quandaries, thereby opening up new vistas for exploration and contemplation. Scientists are currently grappling with major conundrums such as "Can (and should) we create synthetic life in the laboratory or simulate it on the computer?", "What habitat(s) gave rise to the origin of life on Earth, and when (or where) did it first emerge here and elsewhere?", and "Where and how do we look for signatures of life beyond Earth?" Remarkably, we find ourselves poised

1. *The Nobel Prize in Physics 2019,* Nobel Media AB 2020, Oct. 14, 2020, https://www.nobelprize.org/prizes/physics/2019/summary/

on the threshold of a unique era in human history when we might, under optimal circumstances, proceed to answer some of these profound mysteries in the upcoming decades by virtue of the rapid pace of advancements in the aforementioned fields. Ipso facto, humans are closer than ever before to determining whether we are alone or otherwise in the Universe.

With this backdrop in mind, we came to the conclusion that the time was truly felicitous to author this book on the life *out there* that constitutes the bedrock of the burgeoning domain of astrobiology. The definition of *astrobiology* espoused by the NASA Astrobiology Institute (NAI) warrants reproduction,[2] inasmuch as our perspective of this discipline dovetails with the NAI:

> Astrobiology is the study of the origins, evolution, distribution, and future of life in the universe. This interdisciplinary field requires a comprehensive, integrated understanding of biological, geological, planetary, and cosmic phenomena. Astrobiology encompasses the search for habitable environments in our Solar System and on planets around other stars; the search for evidence of prebiotic chemistry or life on Solar System bodies such as Mars, Jupiter's moon Europa, and Saturn's moon Titan; and research into the origin, early evolution, and diversity of life on Earth.

A number of excellent books centered on astrobiology have been written, ranging from the 1950s to the current day. Why did we choose to write one more?

The year 1966 witnessed the publication of the pioneering treatise *Intelligent Life in the Universe* by Iosif Samuilovich Shklovskii and Carl Sagan, which itself built on an earlier (1962) monograph by Shklovskii. A number of first-rate textbooks and review papers have appeared in the ensuing decades, but the classic tome by Shklovskii and Sagan remains virtually unique in two different respects. Our goals therefore were to preserve those two facets, which are described hereafter, while concurrently embarking on a more comprehensive, quantitative, and modern exposition of astrobiology. First and foremost, as our title suggests, we have chosen to explicitly engage with the prospects for intelligent life in the Cosmos.[3] In contrast,

---

2. About NAI, NASA Astrobio Institute, July 24, 2018. https://nai.nasa.gov/about/,

3. In actuality, *intelligence* instantiates a vexatious and nebulous concept, owing to which it must be understood henceforth that we deploy this word in lieu of *technological intelligence*.

most books on astrobiology have opted to maintain a discrete or conspicuous silence, dismiss outright the prospects for extraterrestrial technological intelligences (ETIs), or restrict themselves to cursory renditions of this topic.

In our opinion, this approach is misguided for a number of reasons. It must, of course, be acknowledged at the very outset that the transition to ETIs cannot happen spontaneously from non life—that is, technological entities should originate, at some point down the line, from non technological species. Thus, we do not deny that the frequency of ETIs is much lower than that of non technological extraterrestrial life in all likelihood. In spite of these caveats, however, several benefits accrue from envisaging and thence pursuing the Search for Extraterrestrial Technological Intelligence (SETI) as a legitimate scientific endeavor. Before enumerating them, Karl Popper's (2002) apothegm from *The Logic of Scientific Discovery,* which is perceived by many scholars as his magnum opus, strikes us as being particularly apposite:

> Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature: our only organon, our only instrument … Those among us who are unwilling to expose their ideas to the hazard of refutation do not take part in the scientific game. (p. 280)

First, SETI offers us an alternative avenue to search for life in the Cosmos by looking for signatures of extraterrestrial technology. These *technosignatures* are ostensibly less susceptible to misinterpretation compared to those generated by non technological organisms (*biosignatures*) because the latter may readily arise from abiotic mechanisms; technosignatures are also potentially discernible over larger distances. Second, in the admittedly fortuitous event that we discover technosignatures, the impact on human societies is likely to be correspondingly larger, if for no other reason than sheer anthropocentrism—to wit, humans are likely to accord more weight to the discovery of an intelligent species in comparison to the detection of microbial life. Third, our current technologies for identifying ETIs encompass a much larger search volume than that associated with non technological lifeforms, which might therefore partly compensate for the lower probability of occurrence of ETIs. Lastly, even if we fail to find evidence for ETIs, this pursuit could promote the development of new empirical and theoretical methods in the natural and social sciences, consequently enriching them.

A couple of general points are worth bearing in mind with regard to the SETI endeavor before moving ahead. The transition from an assorted collection of prebiotic chemicals to the smallest known bacterium (or virion) might conceivably entail a much steeper increase in chemical and physical complexity than the transition from bacteria to humans. Yet, this has not stopped us—and indeed it ought not deter us prima facie—from attempting to effectuate the transition from non life to life in the laboratory. Furthermore, much of the prejudice against SETI is attributable to a "giggle factor" that is being continually exacerbated by the jejune portrayal of ETIs in the media as well as the notoriety ensuing from the so-called unidentified flying objects (UFOs) and other subjects of the same ilk. It is a strategic mistake to self-impose blinders on our eyes (telescopes) and to avoid searching for ETIs because of the reasons adumbrated earlier. If astronomers refuse to seriously contemplate the existence of technosignatures, they will almost certainly never be discovered, even if they are writ large across the sky.

However, it would be a gross mistake to assume that this book, or Shklovskii and Sagan (1966) for that matter, is exclusively concerned with technosignatures. As noted earlier, all technological species (which encompasses post-biological intelligences) must have emerged at some stage, either directly or indirectly, from non technological organisms. Thus, it is arguably of the utmost importance that we are cognizant of the major evolutionary events in Earth's history such as the origins of life, multicellularity, eukaryotes, and intelligence among others. Likewise, it is crucial to categorize and quantify biological signatures that may result from extraterrestrial biospheres, mainly because the likelihood of detecting biosignatures is possibly orders of magnitude higher than finding technosignatures, as elucidated in Chapter 9. On account of these reasons, the majority of this monograph is devoted to examining and addressing the prospects for non technological alien life.

The second major aspect in which our book differs from the majority of its predecessors pertains to the choice of writing style. We strove to arrive at the right balance between conveying a substantial amount of technical information and expressing ourselves in an engaging and insightful fashion. We are of the opinion that scientific writing these days tends to emphasize the former aspect at the latter's expense, thus reducing many textbooks to dreary or intimidating compendiums of facts, figures, and equations. In our view, as one of the cornerstones underpinning humanity's ongoing quest for knowledge and self-actualization, science is principally a framework for

asking the right questions and gaining a deeper appreciation of the Cosmos during the process of seeking answers. Hence, we have opted to eschew an orthodox approach that is exclusively oriented toward the explication of technical details and endeavored instead to enkindle and inculcate a genuine passion for the subject by enhancing the readability of this tome; our perspective is akin to that espoused by the French writer Antoine de Saint-Exupéry (1948) in his posthumously published work *Citadelle*:

> Building a boat isn't about weaving canvas, forging nails, or reading the sky. It's about giving a shared taste for the sea. (p. 687)

A number of quotations and epigraphs are consequently interwoven throughout the book to enliven the accompanying discussion while preserving didacticism. The book is organized into three major parts, each of which may be read independently of one another in the main. Charity, as the popular aphorism goes, begins at home. In a similar vein, we must first endeavor to gain an in-depth understanding of the origin and evolution of life on Earth before seeking to comprehend life elsewhere, albeit with the explicit proviso that extraterrestrial life is by no means guaranteed to follow the same biochemistry and evolutionary dynamics. In Chapter 1, we introduce some of the salient properties of life and sketch a few alternative biochemistries. In Part 1, composed of Chapters 2 and 3, we investigate the origin of life on Earth as well as the major evolutionary events that eventually led to the advent of technological intelligence (*Homo sapiens*) on our planet.

In Part 2, we cover some of the vital determinants underlying the search for non technological life beyond Earth. In contradistinction to the widespread scheme of focusing on specific worlds (e.g., Mars and Europa), we opt to delve into a plethora of major generic phenomena presumably responsible for regulating habitability—that is, conditions conducive to the emergence and sustenance of life—in Chapters 4 and 5. Habitability is governed not only by obvious factors such as the temperature and the presence of liquid water but also by a number of stellar processes like stellar winds and flares. Equipped with an understanding of habitability, we tackle the central issue of detecting biological signatures of extraterrestrial life in Chapter 6. This part closes with the stand-alone Chapter 7, which deals with life in subsurface oceans underneath icy envelopes. We have opted to treat these

worlds separately as they differ from Earth in many crucial respects, further motivated by the fact that several prominent astrobiological targets within our Solar system (e.g., Europa and Enceladus) are quintessential members of this category.

Part 3 deals almost wholly with ETIs. Two crucial unknowns that arise in this context are the total number of extant ETIs in the Milky Way and why we have not established any contact with them thus far. Hence, both of these topics are addressed in Chapter 8. We follow this discourse with an exploration of potential technosignatures that may originate from the activities of ETIs, irrespective of whether they are extinct or not, in Chapter 9. The terminal chapter of the book (Chapter 10) is rather heterodox in nature, since it investigates the likelihood of life spreading from one world to another, either via microbes hitchhiking in rocks or ETIs—and perhaps even the descendants of human beings in the distant future—undertaking interstellar travel by means of fast-moving spacecraft.

In writing a book of this magnitude, it is virtually inevitable that a substantial number of inaccuracies will have crept into it. All mistakes intrinsic to this book are solely products of our own cognitive biases and limitations, by reason of which we take this opportunity to solicit and welcome corrections from the readers. Moreover, due to the highly transdisciplinary nature of astrobiology, a comprehensive treatment of every germane subfield is well-nigh impossible. A veritable multitude of pertinent topics have been duly omitted despite their indubitable significance. For example, in grappling with habitability, we have largely set aside gravitational dynamics, although it is manifestly evident that variables such as the eccentricity and obliquity of a given world will influence its short- and long-term climate. This book primarily deals with astrobiological targets outside our Solar system, although the likes of Mars, Venus, and the icy moons of Jupiter and Saturn do make cameo appearances. Motivated by this decision, we focus almost exclusively on techniques for detecting signs of life via telescopes as opposed to in situ investigations. However, by no means whatsoever are we positing that on-site methodologies and missions are rendered unimportant. To reiterate, our approach is calibrated so as to preserve thematic continuity with the rest of the book.

Yet another key facet of our monograph that calls for clarification is the bibliography. As most of the fields comprising astrobiology are witnessing unprecedented growth, the most recent literature as of today may well

become obsolete in a couple of decades. Hence, in lieu of assembling a definitive bibliography, one that would invariably run into the thousands, we have chosen our references selectively, owing to which we apologize for the inadvertent exclusion of cardinal publications. Instead, we have furnished a number of pivotal papers, reviews, and books that ought to be consulted by the reader for additional information. The exceptions to our modus operandi are Chapters 9 and 10, primarily because comprehensive state-of-the-art reviews of these disciplines are surprisingly few in number.

Our expectation is that this work will be employed as a textbook by graduate and possibly advanced undergraduate students, while also functioning as a point of reference for professional scientists whose research interests overlap with astrobiology. On account of the monograph's size and scope, we recommend that the contents herein are best covered over a two- or three-semester sequence, although it is also well suited for specialized courses in SETI and space exploration (Chapters 8–10), the origin and evolution of life on Earth (Chapters 1–3), and the habitability and biosignatures of extrasolar planets and moons (Chapters 4–7).[4] To all those readers who embark on this exhilarating voyage of discovery, we hope that they shall derive as much wonder, enthusiasm, and knowledge as we did while writing the book and that they will get to behold the momentous breakthrough of detecting unambiguous signatures of life on other worlds in their lifetimes.

4. We have not included any assignments at the end of each chapter, since astrobiology is a rapidly growing field wherein developing tractable and meaningful exercises is not only challenging and incommodious but also involves a high degree of subjectivity.

LIFE IN THE COSMOS

Chapter 1

SOME INTRINSIC PROPERTIES OF LIFE

> What is the meaning of life? That was all—a simple question; one that tended to close in on one with years, the great revelation had never come. The great revelation perhaps never did come. Instead, there were little daily miracles, illuminations, matches struck unexpectedly in the dark; here was one.
>
> —Virginia Woolf, *To the Lighthouse*

The principal aim of this book is to address the prospects for extraterrestrial life. There are two words at play here: *extraterrestrial* and *life*. The former is fairly self-explanatory, as it refers to environments beyond the confines of Earth. Hence, this region includes not only worlds within our Solar system but also those outside its limits. On the other hand, the underlying meaning of the second word is riddled with ambiguities. Thus, we must ask ourselves: *What is life?*

Now several levels of ambivalence are at play. First, we must address the question of defining *life*. Second, we should inquire as to whether this endeavor is even possible or necessary in the first place. Finally, even if we do arrive at a sufficiently accurate definition of *life,* it is important to distinguish between *life-as-we-know-it,* which is founded on similar chemical principles as life on Earth, and *life* in a generic sense; the latter can potentially encompass many versions of "weird life" that display a very limited degree of commonality with life on our planet. In this chapter, we shall carry out a brief exploration of these issues. In the spirit of the above quotation from *To the Lighthouse* by Virginia Woolf, it behooves us to acknowledge and accept the possibility that "great revelations" concerning such topics will not manifest all of a sudden. Instead, we may have to rely on a steady stream

of smaller-scale breakthroughs to augment our understanding of the origin and evolution of life on Earth and other worlds; this point was echoed in the recent review by J. P. Nelson (2020).

Before embarking on this study, a seemingly valid objection can be raised—namely, that attempting to define *life* belongs to the provenance of philosophers, where it has indeed provided much grist for the mill. However, a number of reasons have been advanced to justify the importance of defining *life*. Among other reasons, it has been claimed that having a precise definition of *life* could enable us to demarcate the transitional point(s) at which chemistry became biology, i.e., when the origin(s) of life occurred. Second, identifying the salient traits of life may assist us during in situ searches for extraterrestrial life to determine whether a given sample contains life (either extinct or alive). Last, a rigorous definition of *life* might aid in designing appropriate life-detection experiments and clearly communicating results to the public in the event that life is discovered elsewhere.

The last point is subtle and therefore warrants further explication. In 1976, the *Viking* missions carried out a series of famous life-detection experiments to gauge the existence of Martian life. Among the various tests, the most promising results were arguably obtained from the aptly named *labeled release* (LR) experiment. In the LR experiment, a nutrient solution containing radioactive carbon was added to Martian soil, with the subsequent mixture continuously monitored for the release of radioactive gases (e.g., carbon dioxide). The central premise of the LR experiment was that metabolism—to wit, chemical reactions contributing to the sustenance of life—constitutes a universal attribute of life. In the event that life existed on Mars, it was anticipated that radioactive gases would be duly generated.

Remarkably, radioactive gas *was* emitted shortly after the addition of the nutrient solution. In contrast, if the soil was heated to 433 K—namely beyond the thermal limits of Earth-based organisms—no such gas was detected, along expected lines. Initially, this finding was interpreted as strong evidence for the existence of life on Mars. There was, however, an implicit assumption that life on Mars utilized similar organic compounds for metabolic purposes as its counterparts on Earth. Despite this promising sign, there exists widespread consensus today that the *Viking* lander experiments did not yield robust evidence for life on Mars; a detailed contrasting viewpoint can be found, however, in G. V. Levin and Straat (2016). The reader may well wonder why this is the case. The reason is that parallel

experiments conducted using gas chromatography–mass spectrometry failed to detect organic molecules in the Martian soil and atmosphere.[1] The positive results of the LR experiment were explained away as emerging from the presence of unidentified oxidants in the soil. Nevertheless, in light of the limited and ambiguous data available to date, the many subtle and glaring uncertainties and gaps in our knowledge of both abiogenesis and the Martian environment, it is premature to altogether dismiss the notion that the *Viking* experiments discerned life on Mars.

This saga reveals at least two illuminating points. First, the choice and design of life-detection experiments are heavily inspired by life on our planet. This has its share of upsides and downsides. On the one hand, it is easier to detect life when we have a precise idea of what we are seeking. On the other, we run the genuine risk of missing out on unearthing anomalies that might be indicative of biological activity (Cleland 2019). Second, the preceding narrative suggests that scientists accorded a higher priority to the premise that life is founded on organic chemistry than they did to the idea of metabolism being central to life (Benner 2010). Whether this program is correct or not does not concern us here, as our chief purpose is to simply illustrate how humanity's conceptions of extraterrestrial life have patently shaped the multifarious ways in which we seek them.

1.1 DEFINING LIFE: DOES IT MATTER?

Human beings have pondered over the definition of *life,* and what it means to be alive, for millennia (Dick 1996; Vakoch 2013; Weintraub 2014). Our historical journey commences, as many others do, with ancient Egypt: while its peoples might not have recorded a concise definition of life, it becomes fairly apparent from their writings that they had cultivated a sophisticated mythos imbued with certain indelible qualities to demarcate living entities from their nonliving counterparts; the *ka,* for instance, purportedly embodied the spiritual essence of humans and other living beings. Jumping ahead in time, one of the first explicit statements concerning the nature of life was elucidated by Aristotle (1907) in his remarkable *De Anima* as follows:

---

1. The identification of complex kerogen-like organic molecules by the *Curiosity* rover, preserved in mudstones of age 3 Gyr at the Gale crater, has imparted a new and fascinating twist of sorts to this narrative (Eigenbrode et al. 2018).

> Of natural bodies some possess life and some do not: where by life we mean the power of self-nourishment and of independent growth and decay. Consequently every natural body possessed of life must be substance, and substance of the composite order. (p. 49)

Over the subsequent centuries, some of the world's greatest intellectuals weighed in with their thoughts in defining what life is. As the examples are too numerous, complex, and varied, we will desist from a thorough historical exploration of this issue, despite its indubitable importance; the reader may consult Tirard et al. (2010), Mariscal et al. (2019), Cornish-Bowden and Cárdenas (2020), and K. C. Smith and Mariscal (2020) for lucid synopses of this intricate field.

The latter half of the twentieth century was distinguished by a rapid proliferation of attempts to define *life*. A number of these treatises sought to identify a definitive set of properties that, viewed collectively, were purportedly unique to life. One such example is the classification proposed by Koshland (2002), wherein the "seven pillars" of life were identified as (1) program (i.e., analog of DNA), (2) improvisation (roughly analogous to adaptation), (3) compartmentalization, (4) energy, (5) regeneration (self-maintenance), (6) adaptability (flexible behavioral responses to changing environments), and (7) seclusion (preventing mix-up of metabolic reactions). In this schema, it is important to recognize that not all of the seven pillars are independent of one another.

Perhaps the most famous and hotly debated conceptualization of life is none other than the working definition adopted by NASA's astrobiology division ever since the 1990s (Joyce 1994, p. xi):

> Life is a self-sustaining chemical system capable of Darwinian evolution.

While the majority of the words in this definition are fairly straightforward and self-explanatory, the phrase "Darwinian evolution" is oddly specific prima facie. The status of species that deploy *artificial* selection—that is, deliberately engineering their equivalents of genes to enhance fitness—is problematic. Let us push the boundary further by contemplating machines endowed with artificial intelligence that lack the capacity for Darwinian processes; this is not idle speculation as the possibility of superintelligence looms large in the upcoming decades (Bostrom 2014). As per the above

definition, they are not strictly alive, although their biological ancestors were presumably so. Yet, as per common sense, we would acknowledge them as lifeforms, thereby serving to illustrate the limitations of definition-oriented perspectives on life. And what's more, from a practical viewpoint, it is hard to envision a robust detector for Darwinian evolution, especially if the mission under question does not have the luxury of carrying out comprehensive on-site experiments.

Broadly speaking, at least nine major defining terms ostensibly emblematic of life crop up in the literature (Trifonov 2011). Some of the notable ones are self-reproduction, evolution, information, and metabolism. By carrying out a statistical analysis of 123 definitions of life, Trifonov concluded that the two groups self-reproduction and evolution constituted a minimal set, from which life could be defined as "self-reproduction with variations." Interestingly, this definition is fairly close to the preceding definition embraced by NASA. One of the major issues with traditional definition-based approaches along the lines encountered heretofore is that several counterexamples exist; another is the ambiguity introduced by gray areas such as viruses and synthetic lifeforms.[2] Viruses, in particular, are conventionally classified in the non-alive category, but recent developments in our understanding of viral evolution coupled to a growing appreciation of the contradistinction between viruses and virions (virus particles) pose challenges to the status quo (Forterre 2016).

To illustrate this point further, it is instructive to tackle a few concrete examples. Metabolism is often perceived as one of life's most distinctive and enduring characteristics. When distilled to its essentials, metabolism is "merely" a network of chemical reactions entailing electron transfer that consequently yields usable energy. However, similar properties are evinced by luminescent minerals that absorb photons of a particular energy, undergo electronic excitation and deexcitation, and subsequently emit radiation at a different wavelength after the initial source of photons is removed. Clay minerals are also known to be capable of storing, transmuting, and transferring energy. In fact, they also exhibit a limited degree of adaptation to their environment à la life. In view of these functional similarities, it is

2. Instead of using the phrase "robots with artificial intelligence," it is easier to work with "synthetic lifeforms," inspired by the famous *Mass Effect* science-fiction video game series.

not surprising that clay minerals were proposed by Cairns-Smith (1982) as candidates for proto-life (see Section 2.4.3).

As the preceding examples illustrate, it is possible to discover counterexamples if life's seemingly unique traits are viewed in isolation. Despite this caveat, it is helpful to delve into some of the key concepts invoked in connection with defining *life*. The first that we shall tackle pertains to the thermodynamic aspects of life. An offshoot of the rapid development of thermodynamics in the nineteenth century was that the relatively high degree of internal order associated with living systems was cast in thermodynamic terms. In popular parlance, disorder is conventionally linked to entropy. One of the earliest expositions of life, as interpreted through a thermodynamic lens, can be found in the remarkably prescient lecture delivered by Ludwig Boltzmann in 1886 to the Imperial Academy of Science in Vienna (Boltzmann 1974, p. 24):

> The general struggle for existence of animate beings is not a struggle for raw materials—these, for organisms, are air, water and soil, all abundantly available—nor for energy, which exists in plenty in any body in the form of heat, but of a struggle for entropy, which becomes available through the transition of energy from the hot sun to the cold earth.

Several decades later, Erwin Schrödinger (1944) espoused similar thoughts in his equally seminal work *What Is Life?* which paved the way for modern developments in molecular biology and biophysics):

> . . . everything that is going on in Nature means an increase of the entropy of the part of the world where it is going on. Thus a living organism continually increases its entropy—or, as you may say, produces positive entropy—and thus tends to approach the dangerous state of maximum entropy, which is death. It can only keep aloof from it, i.e., alive, by continually drawing from its environment negative entropy—which is something very positive as we shall immediately see. What an organism feeds upon is negative entropy. Or, to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive. (pp. 71–72).

Entropy is a subtle concept that has witnessed much misuse, overuse, and abuse in both scientific circles and other realms of human knowledge.

Schrödinger was equally aware of this sensitive issue, owing to which he admitted in 1948 that it would have been more appropriate to employ the term *free energy* in lieu of *entropy*. As opposed to invoking terms such as *free energy* and *entropy* due to their propensity for misinterpretation, life can be envisioned in terms of *thermodynamic disequilibrium,* as summarized succinctly by Branscomb et al. (2017, p. 3):

> Living systems inherently depend on a host of endergonic, thermodynamically "up-hill," reactions each of which must therefore be forced, or driven, by being coupled to a thermodynamically larger down-hill (exergonic) reaction … Such coupling processes effect a conversion of thermodynamic disequilibria, creating one by dissipating another.

In qualitative terms, living organisms harvest energy from a number of sources ranging from light and chemical energy to feeding on other organisms. The energy thus generated is utilized either for carrying out activities (i.e., work) or for maintaining their high internal order. The propensity for self-maintenance is one of the essential features of living systems; it forms the bedrock of the *autopoiesis* paradigm introduced by Humberto Maturana and Francisco Varela in 1972 to explain the maintenance of living cells (Maturana & Varela 1980; Fleischaker 1990).

Two other major properties of living systems are replication and evolution. The importance of replication stems not only from its obvious relevance to producing offspring but also in the maintenance of organisms—that is, when cells die, they must be replaced by newly minted cells. As one may expect, replication requires a very high degree of fidelity, as otherwise there would be a breakdown of organism functions. The issue of evolution is closely connected to replication for Earth-based organisms. If replication was completely perfect, there would be no variation in the offspring. As a result, in theory, natural selection cannot select those organisms with a higher fitness. The NASA definition of life presumably relies on the phrase "Darwinian evolution" to embody the existence of imperfect replicators that can nevertheless replicate with a high degree of fidelity, so as to pass on their imperfections to their descendants (Benner 2010).

Thus far, we have not explicitly addressed the question of whether we are dealing with life in the *singular* or in the *plural*. This question is an important one that goes beyond phraseology, as it has practical consequences of import in the search for life. On the one hand, we can seek to identify

what properties of a given entity make it alive; the list includes metabolism, replication, and so on. On the other hand, we may envision life as a global (perhaps even planetary-scale) and collective process. If we adopt the latter perspective, it would be natural to treat ecosystems or biospheres as the units or entities in question, with complex cycles of matter and energy exchange occurring between them and other components of the planet—namely, its atmosphere, hydrosphere, and lithosphere.

Even though this apposite picture has garnered traction in recent times, it must be noted that certain well-defined criteria (e.g., replication and evolution) used regularly in defining life vis-à-vis individual organisms do not have exact analogs when it comes to ecosystems and biospheres. Before moving ahead, we observe that the individual and collective aspects of living systems are not mutually exclusive. In fact, they are mutually intertwined, owing to which it may be necessary to synthesize both these facets to yield a definition of life. One such definition was espoused in Ruiz-Mirazo et al. (2010, p. 339), with life defined as

> a complex network of self-reproducing autonomous agents whose basic organization is instructed by material records generated through the open-ended, historical process in which that collective network evolves.

In a similar vein, life could be understood in terms of the acquisition and promulgation of adaptive information by living systems. The adaptive information at play acquires an overall strategic value in case the interacting autonomous agents engender an emergent ecology.

As we have seen, models that seek to understand life through definitions face a number of obstacles. In fact, it is unclear as to whether universal and rigorous definitions of *life* are even possible at this stage, as they require a thorough theoretical understanding of the necessary and sufficient properties that characterize life. As a result, there has been a distinct movement in the twenty-first century to step away from definitions of life (Tsokolov 2009). For instance, Cleland and Chyba (2002) argued that definitions of life will be fraught with controversies in the absence of a theory of living systems. Edouard Machery (2012) has contended that, while scientific definitions of *life* may be possible, scientists studying life across multiple disciplines will end up with divergent definitions, in essence rendering this enterprise pointless. From a specific technical perspective, Nobel laureate Jack Szostak (2012a) has asserted that the endeavor to define *life* and

distinguish it from nonlife is rendered irrelevant from the standpoint of understanding how the transition from chemistry to biology occurred.

Despite the array of negative critiques, several avenues can be pursued in a gainful fashion. The first, and most promising, avenue is to move past definitions of *life* and seek to understand life in terms of its observable manifestations. For example, life exhibits distinctive spatiotemporal structure and dynamical behavior, a high degree of selectivity in its usage of organic molecules, and niche construction (modifying its local environment). This approach not only sidesteps thorny philosophical issues but also opens up avenues for designing life-detection experiments (McKay 2004; Bains 2014). As current life-detection experiments are based, to a substantial extent, on what we know of life on Earth, the search for novel life should ideally endeavor to prioritize the detection and in-depth investigation of anomalies, that is to say, phenomena that are not readily classifiable as living or nonliving (Cleland 2019).

Although definition-based approaches toward life have come in for strong critique, other commentators have defended their utility and usage. Instead of being fixated on an all-embracing definition of *life,* it seems eminently possible to develop specialized conceptualizations for different fields that embody the fundamental traits of Earth-based life, albeit with the proviso understanding that extraterrestrial life need not resemble our planet's lifeforms in all respects (Bich & Green 2018; Mariscal & Doolittle 2020). In a recent study, Mix (2015) introduced the terms "Darwin life" and "Haldane life" to identify organisms that exhibit Darwinian evolution and metabolism / maintenance, respectively. In addition, thinking critically about definitions of *life* might be useful in the context of developing clear and precise scientific practices and communication.

1.2 THE REQUIREMENTS FOR LIFE

Hitherto, we have not specified the actual biochemistry on which life is founded. Throughout the rest of this book, by and large, we will use *life* as shorthand notation for the more cumbersome (albeit accurate) term *life-as-we-know-it*. The latter comprises carbon-based molecules, water as the solvent, and key bioessential elements such as nitrogen, sulfur, and phosphorus. Of this list of constraints, we will not explicitly delve into the properties of nitrogen and sulfur for the most part, while the relevance of phosphorus from a chemical standpoint is addressed in Section 2.2.5. This leaves us

with carbon compounds and water. We will therefore briefly explore the advantages accruing from the use of carbon as the chemical bedrock and water as the solvent.

However, by no means does the emphasis on carbon and H_2O imply that alternative biochemistries are improbable. Our decision to restrict ourselves to life-as-we-know-it in this book is motivated by two key reasons. First, from a theoretical standpoint, there is much that remains unknown with regard to the origin and evolution of life-as-we-know-it on our planet. Hence, attempting to trace evolutionary trajectories on other worlds that are based on heterodox biochemistries is well-nigh impossible. Second, the identification of putative signatures of life (biosignatures) for life-as-we-know-it is easier because we have at least one data point—namely, life on Earth. However, when it comes to life that diverges from life-as-we-know-it, knowing what to look for is comparatively uncertain. Thus, faute de mieux, we focus almost exclusively on life-as-we-know-it because we possess a better understanding of its origin and evolution.

1.2.1 Carbon and water

We will sketch some of the widely accepted reasons for justifying carbon- and water-based biochemistry; the reader may consult N. R. Pace (2001) and Schulze-Makuch and Irwin (2018) for reviews of this subject.

Let us first consider carbon. It is widely accepted that life must comprise large molecules, as the latter are advantageous for numerous reasons ranging from replication to catalysis. Carbon and silicon are considered the best candidates for constituting the backbone of large molecules. Among the two, carbon is known to form complex and stable molecules with many elements, consequently giving rise to millions of compounds. It is not only the number of molecules that singles out carbon but also the varieties of structures that can be formed. On the one hand, carbon atoms can link up easily to produce linear chains. On the other hand, they are equally capable of forming rings, with a classic example being the aromatic hydrocarbons, of which the simplest is benzene.

The variation in structures of carbon molecules enables them to fulfill different functions; for example, the three-dimensional nature of enzymes is essential in enabling catalysis. Apart from structure, another crucial aspect of carbon is its capability to form double (C=C) and triple (C≡C) bonds. The importance of these bonds stems from the fact that the electrons are

delocalized relative to single bonds. This feature facilitates the polarization of certain organic molecules, thereby enhancing the diversity of intermolecular interactions and permitting actions such as binding and catalysis. Another aspect of carbon chemistry worth mentioning is that the resultant molecules are stable over a fairly broad thermal range.

The following are some of the other characteristics of carbon that make it advantageous from a biological perspective:

- Carbon readily forms long polymer chains endowed with the dual properties of being stable while permitting the inclusion of a vast array of functional groups. To offer a couple of examples, carbon comprises the backbone of both proteins and nucleic acids.
- A substantial fraction of reactions in metabolism involve electron transfer with an element transitioning from one oxidation state to another. It is relatively easy for carbon to be transformed from its most reduced (CH_4) to most oxidized (CO_2) variants, with oxidation states of -4 and $+4$, respectively. Both of these gases are available in sufficient abundance on Earth, in addition to being the end products of certain metabolic pathways.
- Carbon and water (as solvent) are ideally suited for one another. The reason is that carbon forms strong bonds with the constituent elements of water—namely, hydrogen and oxygen. The bond strengths of C−H and C−O are ~ 435 and ~ 360 kJ/mol, respectively.

Next, let us direct our attention toward liquid water, which shall be dubbed *water* hereafter for simplicity. There are a number of unique properties possessed by water that make it an ideal solvent for life on Earth. An important point to recognize is that water is not a passive solvent, but rather an active participant in cell biology. It has complex structural and dynamical qualities that are still being uncovered, owing to which our knowledge of the "matrix of life," to borrow the eloquent phrase coined by Nobel laureate Albert Szent-Györgyi to describe water, remains incomplete (P. Ball 2017). We will therefore explore a few of the essential qualities of water from the perspective of uncovering its utility as a solvent.

First, however, we must explicate the necessity for a solvent. Although life on Earth is liquid-based, we cannot be confident at this stage that solid- or gas-based life is altogether impossible. A fair number of benefits originate from a liquid medium that combine the pros of solids and gases. Certain

molecules, for example, are stabilized in the solvent, whereas others undergo dissolution and enable the exchange of chemical compounds and energy. Likewise, the liquid medium imposes constraints on the concentration and dilution of reactants; instead, if we had a gaseous medium, dilution could occur rapidly and almost indefinitely. As a matter of fact, the unsung property of liquids to undergo evaporation provides an excellent mechanism for the concentration of reactants and products, thus powering prebiotic synthesis as chronicled in Section 2.4.1. Lastly, solvents enable the dissolution of some molecules and not others, thereby producing new interfaces and boundaries.

Now we shall address the reasons why water is often perceived as an excellent solvent, especially in the context of carbon-based lifeforms (Pohorille & Pratt 2012; Westall & Brack 2018). One of the most prominent among them has to do with its polar nature—that is, it has a finite dipole moment. The intrinsic dipole moment of water molecules gives rise to a finite dielectric constant ε in accordance with basic electromagnetic theory. The polarity of water enables the molecules to conglomerate together via hydrogen bonding; in crude terms, hydrogen bonding describes the force exerted by one dipole (i.e., water molecule) on another. This hydrogen bonding not only makes water polymerlike in some respects but also increases its freezing and boiling points relative to a hypothetical nonpolar solvent with the same molecular weight.

The thermal range over which liquid water exists at an atmospheric pressure of 1 atm is higher than many other liquids at the same pressure. To offer a case in point, let us compare water, ammonia (NH_3), and hydrogen cyanide (HCN): while H_2O evinces a thermal range of 100 K, the corresponding temperature interval for NH_3 is 44.4 K and that of HCN is 39.3 K. If we consider planets with surface temperatures and pressures close to Earth, we see that the triple point of water is not far removed from the average temperatures on the surface. Hence, this attribute permits the transformation of water from one phase to another; in other words, ice and water habitats can exist along with water vapor in the atmosphere.

As noted previously, water has a high dipole moment. It is thus easy for polar or ionic molecules such as salts, amino acids, and sugars to dissolve in water; note that each of these compounds plays vital biological roles on Earth. Theoretical models indicate that high values of ε are vital for optimizing electromagnetic interactions in the solvent; water belongs

to this category (with $\varepsilon \approx 80$). On the other hand, as nonpolar organic molecules are stable in water, they constitute the basic components of cell membranes that serve to separate the interiors and exteriors of cells. Hitherto, the polar nature of water molecules comes across as beneficial, but there are disadvantages as well. A number of organic molecules are degraded in the presence of water (hydrolysis) due to its high reactivity, whereas other reactions involving the removal of water (dehydration) are suppressed.

Apart from the wide thermal range over which liquid water can exist, its high specific heat capacity and enthalpy (latent heat) of vaporization are two more advantageous features worth highlighting. The former ensures that raising or lowering the temperature of a given mass of water is relatively difficult, thereby ensuring that oceans can serve as an effective buffer against climate fluctuations. The latter constitutes a common strategy used by organisms to cool down, as evaporation transports away excess heat. At the standard atmospheric pressure, the enthalpy of vaporization for water is 40.7 kJ/mol, whereas the associated values for ammonia and hydrogen cyanide are 23.3 kJ/mol and 25.2 kJ/mol, respectively.

Water is unusual because its density actually decreases at temperatures below 277 K, where its maximum density is attained. An immediate consequence of this fact is that water ice floats on the surface, consequently serving as an insulator and preventing the lower layers of liquid water from freezing over under select circumstances. This property is not only of paramount importance for aquatic life in cold habitats on Earth (Henderson 1913) but also for worlds with subsurface oceans of water underneath icy shells; see Chapter 7 for an exposition of this topic. It is currently suspected that many of water's anomalous traits (e.g., density variation with temperature) are traceable to the existence of weak van der Waals forces or at least two distinct molecular structures (Brini et al. 2017).

To properly assess the benefits of water as a solvent, it should be jointly evaluated with the potential biomolecules deployed in building putative organisms. Consider, for instance, extraterrestrial lifeforms where proteins play an indispensable role. It is well-known that proteins fold into complex three-dimensional structures, which exhibit slightly enhanced solubility and stability in water; this feature is presumably hard to replicate in other solvents. Last, but by no means the least, any solvent must be sufficiently abundant to permit the origin and sustenance of life. Water, in its many phases, is widely distributed not only on our planet but also in our Solar

system and the Universe as a whole. This fact is along expected lines, seeing as how both hydrogen and oxygen are among the most abundant elements in the cosmos.

1.2.2 Alternatives to carbon and water

We will briefly elucidate some promising alternatives to carbon and water chemistry that have been explored. The reader is referred to Sagan (1973), Bains (2004), Baross et al. (2007), and Schulze-Makuch and Irwin (2018) for further details.

1.2.2.1 Looking beyond carbon

As we have seen, one of the primary advantages in employing carbon as the backbone stems from its ability to form long chains via C−C bonds. It must, however, be recognized that chains are not obligated to comprise the same element throughout; instead, alternating atoms could form the backbone, with B−N and Si−O representing two such examples.

When contemplating alternatives to carbon, silicon is the first element that comes to mind, since it belongs to the same group as carbon in the periodic table. However, due to the smaller relative size of the carbon atom, its electrons are located in closer proximity to the nucleus, thus permitting the formation of stronger bonds with lighter elements like hydrogen. The larger size of silicon's electron cloud also increases the degree of Debye shielding and leads to the formation of weaker Si−Si bonds ($\sim$ 300 kJ/mol) compared to their C−C counterparts ($\sim$ 400 kJ/mol). One of silicon's chief limitations is that the formation of double and triple bonds, along the lines of carbon, is much more difficult because they are disfavored on thermodynamic grounds (Petkowski et al. 2020), partly due to the larger atomic sizes and bond angles involved. It is worth recalling here that double and triple bonds play a major role in facilitating polarization and "storing" energy, implying that their possible near absence may pose issues.

In general, silicon compounds are more reactive than their carbon counterparts. First, as Si−X bonds typically exhibit lower bond strengths than C−X bonds, less energy is required to break them. Second, as silicon tends to form more polarized compounds with nonmetals on account of its ability to donate electrons more readily, these compounds are more susceptible to attack by electron donors and acceptors. Third, the access to low-lying

d-orbitals permits certain reaction pathways to occur at lower energies relative to carbon, which lacks this feature. As a result of the high bond strength of Si—O, silicon readily undergoes oxidation to yield compounds like SiO_2. Hence, in environments with oxygen or water (which contains O), the formation of other silicon compounds is typically suppressed.

Stable polymers of silicon involving either Si—Si (silanes), Si—O (silicones), or Si—C backbones are known to exist. Silanes have the general chemical formula Si_nH_{2n+2} and represent the direct analogs of hydrocarbons. However, with increasing length, the stability of silanes decreases. A potential avenue for bypassing this conundrum is to replace hydrogen with organic functional groups that enhance the stability of the ensuing molecules. Silicones, also called polysiloxanes, are based on Si—O—Si linkages and exhibit a number of useful properties including lower susceptibility to oxidation, high dielectric constants, greater thermal stability, and resistance to ultraviolet (UV) radiation. Apart from Si—Si and Si—O polymers, silicon polymers with nitrogen (i.e., polymers with Si—N bonds) are known to exist.

Silicon has multiple valence states and forms stable compounds with a number of elements. These compounds span a diverse array of structures, with some notable examples: (1) branched and unbranched chains, (2) rings (e.g., cyclohexasilanes), and (3) cage-like molecules (e.g., silsesquioxanes). It is believed that aromatic compounds such as benzene, which play important roles in biology on Earth, lack silicon counterparts on account of silicon's difficulty in forming double bonds.[3] However, features such as electron delocalization and light-activated electronic effects are achievable through other avenues in silicon compounds, as laboratory experiments involving polysilanes (silicon analogs of polythenes) have demonstrated. We round out our list by observing that cell membranes, one of the key constituents of cells, comprise complex molecules known as lipids. Polysilanes containing up to 26 Si—Si bonds are known to possess many of the same desirable characteristics as lipids.

If we are interested in silicon-based biochemistry, we must ask ourselves what geochemical conditions would optimize its chances. In order to carry out this thought experiment (gedankenexperiment), let us tackle life based

3. More precisely, it has to do with the inability of silanes to form π-conjugated systems, which are necessary to enable the synthesis of aromatic compounds.

on silanes as they represent the analogs of hydrocarbons. For starters, the atmosphere should be mostly devoid of oxygen, as otherwise the silanes undergo oxidation to yield silicates. Likewise, the use of liquid water as solvent is problematic due to silicon's high affinity for oxygen. The temperatures should be sufficiently low and / or the pressures sufficiently high to ensure that the high reactivity of silanes is suppressed via the slowdown of reaction rates. Ideally, carbon should exist in limited quantities, as otherwise it might outcompete silicon to produce complex molecules. In a similar vein, it is possible to identify conditions under which life based on silicones is feasible: a combination of high temperatures and pressures, low carbon availability, reducing atmosphere, and solvents other than water (cf. Petkowski et al. 2020).

Apart from silicon, boron and nitrogen have been studied as potential candidates for the building blocks of life. Several advantages arise naturally from boron-nitrogen compounds. As boron-nitrogen bonds resemble carbon-carbon bonds, it is not surprising that boron-nitrogen compounds resemble hydrocarbons in many respects, except with higher melting and boiling points. Boron-nitrogen compounds have adequate thermal stability and are capable of existing in the temperature range where ammonia is a liquid; the latter is important because these compounds exhibit high affinity for ammonia as solvent. However, the major strike against boron-based life is that the abundance of boron is very low on Earth ($\sim$ 10 ppm in the crust). However, on worlds endowed with higher boron abundances or those with effective pathways for locally enriching this element, it is conceivable that boron-based chemistry might exist. Aside from boron-nitrogen compounds, sulfur and phosphorus have also been investigated, but they form a relatively limited number of organic analogs and exist over a narrower thermal range, owing to which they may be discounted.

1.2.2.2 Looking beyond water

In contrast to carbon-based chemistry, where not many viable alternatives seem to exist, it is easier to identify potential candidates that can substitute for liquid water as the solvent. In order for a particular compound to be deemed a potential replacement for water, it should exist in sufficient abundance. Moreover, due to the advantages conferred by water's polar nature, a nonzero dipole moment is beneficial but not strictly necessary. In Table 1.1, many of the solvents commonly perceived as alternatives

Table 1.1 Physical and chemical properties of potential liquid solvents for life

Solvent	T_M (in K)	T_B (in K)	Dipole moment (in D)	ΔH_V (in kJ / mol)	ε
H_2O	273	373	1.85	40.7	80.1
NH_3	195	240	1.47	23.3	16.6
HCN	260	299	2.99	25.2	115
HF	190	293	1.83	30.3	83.6
H_2SO_4	283	610	~2.7	~56	~100
H_2O_2	273	423	~2	~52	~84
CH_3OH	175	338	~1.7	~38	~35
$HCONH_2$	275	483	~3.7	~62	~110
CH_4	91	112	0	8.2	1.7
C_2H_6	90	185	0	14.7	~1.9
N_2	63	77	0	~5.6	~1.4

Notes: T_M and T_B are the melting and boiling points of the liquid at 1 atm; the liquidity range can be determined by calculating $T_B - T_M$. The dipole moment is measured in units of debyes (D), where $1\,D = 3.33 \times 10^{-30}$ C·m. The abbreviations ΔH_V and ε represent the enthalpy of vaporization and the dielectric constant of the solvent, respectively. Most of the data in this table is adapted from Schulze-Makuch and Irwin (2018). (*Data source:* D. Schulze-Makuch and L. N. Irwin [2018], *Life in the Universe: Expectations and Constraints,* 3rd ed. [Berlin: Springer Verlag].)

to liquid water are presented along with their basic physical and chemical characteristics.

Among all candidates, perhaps none has received the same degree of attention as ammonia. On Earth, life utilizes compounds containing the carbonyl (C=O) unit in metabolism, but theoretical studies indicate that the C=N bond could be used instead on worlds with liquid ammonia. Liquid ammonia is known to be analogous to water in several respects, with certain chemical reactions (ammoniation reactions) in ammonia playing the same functional role as hydration reactions in water. Ammonia also dissolves a number of organic molecules akin to water, but the latter is arguably a more powerful solvent because liquid ammonia has a lower enthalpy of vaporization, dipole moment, dielectric constant, and liquid range (at 1 atm) relative to water. In addition, liquid ammonia is less robust to dissociation by UV radiation and does not concentrate macromolecules with the same ease as water. However, against these putative concerns, it should be noted that liquid ammonia is well suited to being a solvent at lower temperatures, either in its pure form or mixed with water. On such worlds, instead of phosphates (addressed in Section 2.2.5), it is plausible that

ammonium-phosphate compounds might play analogous roles in energy transactions.

Hydrocyanic acid (HCN) exhibits a number of potential advantages relative to ammonia, and even to water in some instances. It has a higher dipole moment than both these liquids and offers thermal buffering comparable to water and better than ammonia. Liquid HCN also offers shielding against UV and even forms amino acids in alkaline solutions, as evinced by the following reaction:

$$3\mathrm{HCN} + 2\mathrm{H_2O} + h\nu \rightarrow \mathrm{C_2H_5NO_2} + \mathrm{CN_2H_2}, \tag{1.1}$$

where the third term on the left-hand side denotes UV radiation, while the first and second terms on the right-hand side are glycine and cyanamide, respectively. The nucleobase adenine, which we encounter in Chapter 2, is a pentamer of HCN. Against these positives, some of the cons should be listed. HCN has a limited range over which it exists as a liquid, it does not dissolve many biologically relevant molecules employed on Earth, and it probably has a lower abundance in liquid form with respect to water.

The situation with regard to hydrofluoric acid (HF) is rather interesting. Its solvent properties are similar to water, and fluorination can effectively substitute for oxidation in metabolic reactions and release even more energy in the process. However, a major stumbling block with this solvent is that the abundance of fluorine, and hydrofluoric acid by extension, is very low. Note that the abundance of fluorine is $\sim 10^3$ times smaller compared to carbon and oxygen in our Solar system. Sulfuric acid (H_2SO_4) is another intriguing solvent that possesses a number of desirable traits including a high dipole moment, liquidity range, and dielectric constant. If life exists in the Venusian atmosphere (see Section 5.7), it might utilize liquid H_2SO_4 as solvent, in view of the fact that this compound has been detected in the clouds of Venus and the pros listed above. Metabolism involving sulfuric acid could potentially use C=C bonds as the reactive units in metabolism. Before moving on to a different class of solvents, we note that hydrogen peroxide (H_2O_2) also has promise as a solvent, especially if it exists as a mixture with liquid water.

In attempting to identify alternatives to liquid water, there is a tendency to search for candidates from the same class (i.e., which are inorganic and polar). It is necessary to appreciate that other types of liquids might also constitute viable solvents. For example, liquid nitrogen is a nonpolar

inorganic solvent that exists at very low temperatures (< 80 K) and might permit silicon-based chemistry. Organic solvents are often nonpolar, but the inclusion of the hydroxyl group ($-$OH) endows the ensuing molecules with a finite dipole moment. Among polar organic solvents, two promising examples are methanol (CH_3OH) and formamide ($HCONH_2$).

Methanol is appealing in several respects: it has a dipole moment higher than ammonia, a broad liquid range, a high dielectric constant, and enthalpy of vaporization. However, a potential issue with methanol as solvent is that its abundance may be low; none of the missions to objects in our Solar system have detected widespread methanol. In some respects, formamide is an ideal solvent: it possesses a higher dipole moment, dielectric constant, and liquidity range than water. In addition, the synthesis of a number of valuable prebiotic compounds occurs spontaneously in formamide, whereas the same molecules are very unstable in water; other pros associated with formamide are specified in Section 2.3.4. Despite the manifold benefits accruing from using formamide as a solvent, its abundance is probably low on most habitable worlds. With that said, the existence of formamide pools and lakes in specialized environments is conceivable.

Hitherto, we have focused only on polar solvents. One of the reasons for doing so has to do with cell membranes. The molecules that make up these membranes possess polar "heads" and nonpolar "tails," thus providing a definite orientation; this is because the heads are oriented toward the solvent and vice versa. Hence, at first glimpse, it would appear as though the same scheme cannot function in nonpolar solvents, but unorthodox membrane alternatives do exist (Stevenson et al. 2015). By and large, there are sufficient empirical grounds for supposing that the reactivity of organic molecules in nonpolar solvents is not far removed from that in water. Hydrogen bonding and phase separation of liquids are two other phenomena that are realizable in nonpolar solvents. In addition, the high reactivity of water poses nearly insuperable barriers for the synthesis of certain types of biomolecules (e.g., RNA and DNA) that are presumably nullified when it comes to nonpolar solvents. Finally, the hydrocarbons comprising the putative solvents readily form hazes in the atmosphere that serve as effective UV screening compounds and thereby mitigate the damage wrought by UV radiation.

Hence, in principle, seas and lakes made up of nonpolar hydrocarbon solvents are seemingly conducive for the origin and evolution of life. At the standard pressure of 1 atm, methane and ethane have boiling points of ~ 112 K and ~ 185 K, respectively. As one moves toward longer

hydrocarbons, the boiling point increases accordingly. As the surface temperature of Saturn's moon Titan is ~ 94 K, it can support oceans comprising liquid methane and ethane; observations by the *Voyager* and *Cassini-Huygens* missions have confirmed this fact. In the hydrocarbon-rich environments of Titan, a wide range of exotic metabolic pathways are possible. For instance, microbes may reduce acetylene (C_2H_2) with atmospheric hydrogen to yield methane, analogous to microbes known as methanogens on Earth. The net reaction is expressible as

$$C_2H_2 + 3H_2 \rightarrow 2CH_4. \tag{1.2}$$

Apart from potential metabolic pathways, the synthesis of a wide range of the functional equivalents of sugars, proteins, and lipids is also conceivable. On the whole, Titan fulfills many (and perhaps all) of the standard requirements for life: suitable solvents,[4] thermodynamic disequilibrium, and a high abundance and diverse spectrum of organic and nonorganic molecules, to name a few (Hörst 2017). As such, it represents one of the most promising sites within our Solar system for discovering novel life dissimilar to life-as-we-know-it.

The last category of solvents that merit a mention are supercritical fluids. On the phase diagram of a given substance (e.g., water), there exist a critical temperature and pressure beyond which the distinct partitioning into liquid and gas phases breaks down. In this regime, the supercritical fluid thus formed is neither a liquid nor a gas; instead, its properties are intermediate between gases and liquids. Supercritical fluids have garnered interest in recent times as they exhibit high solubility, diffusion rates, and stabilization of molecules as well as the ability to synthesize complex compounds that are not easy to produce in water. Of the various fluids in this category, supercritical CO_2 is of particular interest because carbon dioxide is found in several worlds of our Solar system, thus suggesting that supercritical CO_2 oceans might exist in some worlds.

If we postulate that life requires a liquid solvent and that the emergence of life is possible in different solvents, determining the relative abundances of various solvents on habitable worlds in the Universe is of marked importance. By gauging the primary components of extraterrestrial oceans and

4. Setting aside the presence of liquid hydrocarbons on the surface, Titan is known to also host liquid water beneath the surface.

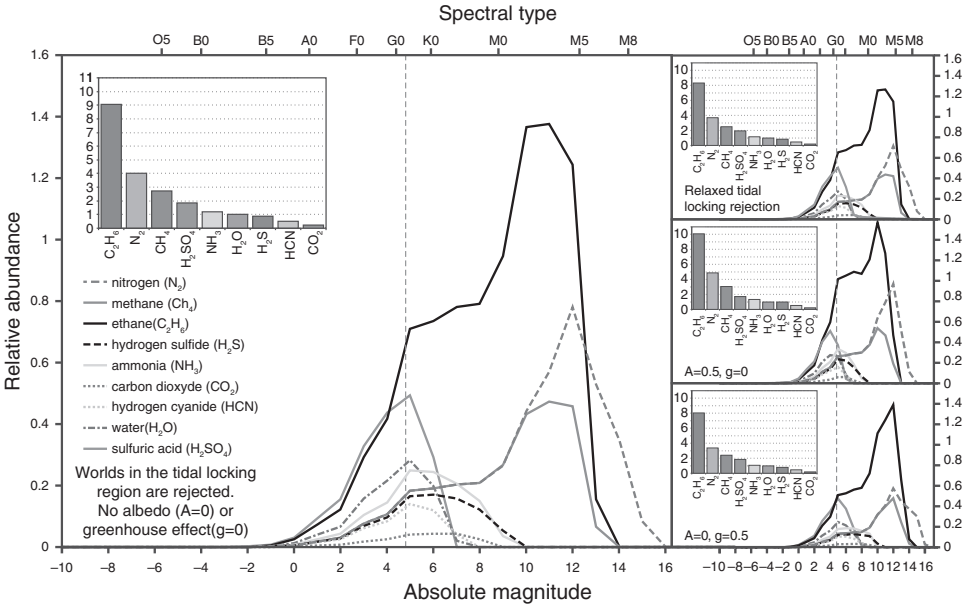


Figure 1.1 The average likelihood of different solvents relative to water is displayed as a function of the stellar spectral type. In the main panel, the planet's albedo is set to zero and the greenhouse effect is neglected. In the inset, the histogram depicts the cumulative abundances of solvents in our Galaxy (found by computing the integrals of the corresponding average likelihood curves); by construction, the global abundance of liquid water is unity. In the right-top panel, worlds with tidal locking (i.e., rotation period equals the orbital period) are included, thus taking Jupiter-sized planets with orbital periods of < 10 days into account. In the right-middle panel, the albedo is held fixed at 0.5 and the greenhouse effect is neglected. In the right-bottom panel, the albedo is zero but a finite greenhouse effect is incorporated (© Mary Ann Liebert, Inc. *Source:* F. J. Ballesteros, A. Fernandez-Soto, and V. J. Martinez [2019], Diving into exoplanets: Are water seas the most common?" *Astrobiology* 19[5]: 642–654, fig. 6.).

their prevalence, we can endeavor to identify possible biochemistries that are compatible with the solvents under discussion as well as predict the ensuing biosignatures and the likelihood of detecting them. In order to estimate the global relative abundances of solvents, the spatial extent of the zones over which the liquid solvents could exist must be combined with (1) the frequency of planets and moons above a certain size threshold in these zones and (2) the frequency of stars belonging to a particular spectral type.

This analysis was carried out by Bains (2004), which was recently updated by Ballesteros et al. (2019). An inspection of Figure 1.1 suggests

that ethane seas might be nearly ten times more common than water seas. A crucial point to bear in mind concerning models of this ilk is that they tend to yield *upper limits* on the number of worlds capable of harboring oceans, which can diverge significantly from the actual values. For instance, the surfaces of Mars and Venus may have featured widespread liquid water on their surfaces around 4 billion years ago, but neither of them hosts perennial water bodies in the current epoch.

1.3 THE ANNA KARENINA PRINCIPLE

> Happy families are all alike; every unhappy family is unhappy in its own way.
>
> —Leo Tolstoy, *Anna Karenina*

In 1994, Jared Diamond introduced the Anna Karenina principle (AKP), which was subsequently refined, and expanded on, in his influential (albeit controversial) treatise *Guns, Germs, and Steel* (1997). The gist of Diamond's interpretation of *Anna Karenina* is summarized as follows. A family must fulfill a number of pivotal criteria collectively in order to be deemed a happy one. If even one single crucial ingredient, or a bevy of them, is missing, the family may experience unhappiness. To wit, unhappy families differ from one another as the factors responsible for their unhappiness are not always the same.

Diamond (1997) posited that a similar line of reasoning could be duly brought to bear on other realms of science. It is instructive to contemplate what he had to say concerning this topic:

> This principle can be extended to understanding much else about life besides marriage. We tend to seek easy, single-factor explanations of success. For most important things, though, success actually requires avoiding many separate possible causes of failure. The Anna Karenina principle explains a feature of animal domestication that had heavy consequences for human history—namely, that so many seemingly suitable big wild mammal species, such as zebras and peccaries, have never been domesticated. (p. 157)

Thus, Diamond invoked the AKP to explain why certain animals have been domesticated by human beings while others have not: the ones belonging to the latter category fail to meet one or more desiderata. In studies of consumer psychology, the conjunctive rule—which states that a product

must exceed a certain minimum threshold on all counts to be deemed acceptable—bears some resemblance to the AKP. The AKP has been utilized in fields as diverse as ecology, marketing research, and bibliometrics; the reader may consult Bornmann and Marx (2012) for an overview of this field.

We can, therefore, ask ourselves whether the AKP is applicable to the quest for extraterrestrial life. In order for the AKP to possess practical value, a necessary condition is that we must have knowledge of the essential criteria that must be fulfilled in order to permit the origin and evolution of life. It is immediately apparent that resolving this subject is a highly complex endeavor. In fact, it may very well be that no universal criteria exist for the emergence of biological systems; that is, the core factors vary from one world to another. However, it appears reasonable to contend that certain basic physicochemical conditions should be met in order to permit the existence of extraterrestrial life. Needless to say, identifying these conditions is fraught with difficulties and ambiguities.

A major obstacle encountered in this realm has to do with our use of the word *life*. The minimal set of desiderata will evidently differ greatly depending on whether we are interested in (1) all versions of life, (2) life-as-we-know-it (both "simple" and "complex"), (3) complex multicellular life-as-we-know-it, and (4) humans. On account of the pragmatic and theoretical grounds justified in Section 1.2, we opt to focus henceforth on item (2). In other words, we tackle lifeforms that are based on carbon-based chemistry with water serving as the solvent. Despite our knowledge of the origin and evolution of life on Earth being far from complete, a list of key criteria is adumbrated below based on Benner et al. (2004), with the proviso understanding that it ought not be regarded as definitive.

- Thermodynamic disequilibrium, as we elucidated in Section 1.1, is arguably one of the most fundamental requirements. Loosely speaking, it is tantamount to the existence of free energy sources that accordingly furnish the minimum energy "quanta" necessary for lifeforms (Lever et al. 2015).
- Temperatures and pressures conducive to the presence of liquid water, which in turn facilitates different types of chemical bond formation and cleavage. Naturally, by imposing the constraint that water serve as the solvent, we require molecules taking part in biological functions (e.g., metabolism) to exhibit sufficient solubility in water.

- Sufficient reserves of carbon are desirable because it constitutes the scaffolding for life-as-we-know-it. The availability of hydrogen and oxygen, in the form of water, is also self-evident. In addition, if we confine ourselves to strict analogs of life on Earth, nitrogen, sulfur, and phosphorus are also indispensable. In the early Universe, the abundance of these bioessential elements was, on average, lower than today. However, even in the early stages, carbon-rich and metal-poor planets did exist (Mashian & Loeb 2016) and might have provided tenable nurseries for life.
- If we consider the counterparts of nucleic acids, the presence of repeating backbone charges is advantageous in several respects. They enhance solubility, safeguard the central backbone from disruptive inter-strand interactions, confer structural stability, and permit Darwinian evolution to operate. On Earth, the backbone charges are contributed by phosphate groups, thereby underscoring the significance of phosphorus.
- Unlike nucleic acids with their repeating charges (monopoles), the repeating elements of polypeptides (used for building proteins) are dipoles. As the positive charge of one dipole can interact with the negative charge of another, this feature is well suited for protein folding as well as functionally similar biopolymers. The importance of folding stems from its ability to drive efficient catalysis.
- One of the major aspects of life is that it requires compartmentalization to protect fragile biopolymers and allow metabolic reactions to occur unhindered by the external environment. Hence, in order to synthesize the analogs of cell membranes in aqueous media, the existence of amphiphilic molecules is beneficial. These molecules should resemble lipids in that they possess polar heads and nonpolar tails as described in Section 1.2.2, thus permitting self-assembly in water to yield compartments.

We reiterate that this list is by no means exhaustive, owing to which it merely accentuates a set of potentially necessary, but *not* sufficient, criteria. The reader is directed to Bartlett and Wong (2020) for a thoughtful exposition that encompasses other salient characteristics such as autocatalysis and learning. On this note, other studies have posited autocatalytic feedback and self-organization as one of the chief hallmarks of living entities (Tsokolov

2010; Vitas & Dobovišek 2019). Likewise, theoretical modeling and empirical results are pointing toward deep—but scarcely understood—connections between outwardly disparate paradigms such as entropy maximization, free energy, Bayesian inference, reinforcement learning, cognition, Darwinian evolution, and the origins of life (Friston 2010; Wissner-Gross & Freer 2013; Kirchhoff et al. 2018; Ramstead et al. 2018). Although an exhaustive treatment of the aforementioned topics is not practical, we will briefly encounter autocatalysis and learning in Sections 2.6.1 and 3.7, respectively.

The checklist delineated above constitutes a useful heuristic for determining what targets are optimal in searching for life beyond Earth that is broadly carbon- and water-based. For instance, as we shall see in Chapter 4, the concept of the habitable zone serves to define the region around the host star where liquid water could theoretically exist on the surface. Furthermore, the catalog has the potentiality to stimulate additional research in quantifying the minimal set of necessary conditions that ought to be fulfilled for the emergence of life-as-we-know-it as well as weird life—namely, life based on alternative biochemistries—in the Universe.

PART 1

THE ORIGIN AND EVOLUTION OF LIFE ON EARTH

Chapter 2

THE PATHWAYS TO THE ORIGIN OF LIFE ON EARTH

> But, after all, who knows, and who can say
> whence it all came, and how creation happened?
> The gods themselves are later than creation,
> so who knows truly whence it has arisen?
>
> —From the hymn "Nasadiya Sukta," *Rigveda* 10:129

Needless to say, the origin of life (abiogenesis) on our planet remains one of the greatest mysteries of modern science and has attracted all manner of conjectures starting with ancient creation myths, a classic example of which is the passage from the *Popol Vuh* quoted at the end of this chapter. The above quotation from the *Rigveda* constitutes a striking and concise encapsulation of the inherent ambiguities concerning the origin of living beings on Earth. Abiogenesis is, moreover, a conundrum that has attendant scientific (and philosophical) consequences for the evolution of *Homo sapiens* because technological intelligence can only evolve when life has originated in the first place. Hence, as we shall discuss below, origin-of-life research has important practical consequences in the search for extraterrestrial life.

Broadly speaking, virtually all scientific theories of abiogenesis subscribe to a continuity principle between nonlife and life, implying that the latter must have emerged from the former after some process of complexification starting at the molecular level. What remains unclear, however, is how and when this process occurred. One could also ask why life originated, i.e., why does something (life in particular) exist rather than nothing? While this observation can be explained to some degree by selection bias—life must exist in the first place before it can observe itself—it remains unclear as to whether this line of reasoning constitutes the complete answer. The hows

and whys of abiogenesis naturally lead to another fundamental question that has been the subject of fierce debates between opposing camps.

Is abiogenesis, on Earth and elsewhere, a deterministic process or a matter of chance? This question is sometimes embodied in a different form: Is the origin of life a highly probable event or an extremely unlikely one? Although these two versions are conflated or equated with one another, a deterministic process does not automatically translate to a high probability of occurrence. For instance, in mathematical terms, chaos is a feature of deterministic dynamical systems and yet appears on the surface as though it were random chance. Advocates of determinism propose that there exist general principles that govern the probability of abiogenesis in a systematic and precise fashion. In contrast, proponents from the opposing camp have argued that life arose from a specific and intricate sequence of multitudinous chance events, implying that altering the ordering or contents of this sequence may have led to its nonemergence (thereby making accurate predictions unattainable).

It is readily evident that this issue is a highly complex one, and the two viewpoints espoused above are not even wholly contradictory. For instance, each individual event may be deterministic unto itself, but the agglomeration and ordering of these factors could lead to an effectively random process. As noted in the previous paragraph, chaos is deterministic and yet, loosely speaking, comes across as being stochastic in nature. Furthermore, deterministic processes are not synonymous with events that have a high probability of occurrence. Apart from these two camps, a plethora of viewpoints that seek to navigate the shifting terrain between determinism and randomness have been expounded. The interplay between chance and necessity will not be elaborated on herein; instead, we refer the reader to Fry (2000) for an in-depth discussion of this intricate topic.

Although the prior discourse may come across as intriguing, it also seems abstract prima facie. However, important practical consequences stem from answering the aforementioned questions. If life was indeed deterministic and quite probable, this would increase our chances of detecting extraterrestrial life, regardless of whether it is simple or complex in nature. On the other hand, if life entailed a succession of random events and the overall outcome was highly improbable, the likelihood of finding extraterrestrial life becomes infinitesimally small. Thus, learning more about the origin of life is arguably imperative not only from a scientific perspective but also from the standpoint of ensuring that the appropriate amount of federal and private funding is allocated to it.

When framed this way, understanding how the origin of life took place on our planet serves as the natural starting point for studying extraterrestrial life. If we can determine the processes by which abiogenesis occurred on Earth, we might be able to extract some conclusions about how often life could emerge on other worlds. Moreover, we may even be able to answer an equally profound question: Is the origin of life attainable through a unique and fixed route or via multiple pathways? Hence, in this chapter, we shall delve into the complex and rapidly evolving field of origin-of-life research to understand when and how life originated on our planet. One prominent class of models that we do not tackle here is the notion that life was transported from elsewhere to Earth, known as panspermia; this is tackled in Chapter 10.

2.1 WHEN DID LIFE ORIGINATE ON EARTH?

First we will outline the earliest evidence for life on Earth and concomitantly present a brief chronology of when the Earth became habitable. This topic has been studied in detail by many authors and constitutes a fast-evolving field. For further details, the reader may consult Zahnle et al. (2007, 2010), Arndt and Nisbet (2012), Pearce et al. (2018), and T. M. Harrison (2020).

2.1.1 The habitability timeline for the Earth

Before discussing the different lines of evidence for the earliest biosignatures (signs of life) on Earth, it is instructive to sketch the timeline of pivotal events that led to abiogenesis (Sleep 2018).

The age of the Solar system is approximately 4.568 Gyr (Bouvier & Wadhwa 2010), which is based on the isotopic dating of the oldest components of meteorites called calcium-aluminum-rich inclusions (CAIs).[1] The Earth is slightly younger because it formed after the coalescence of planetesimals from the protoplanetary disk. Subsequently, it underwent differentiation—that is, due to the temperature profile, denser liquids sink under gravity while low-density fluids rise upward. It has been estimated

1. Note that 1 Gyr is 10^9 yr and 1 Myr is 10^6 yr, whereas 1 Ga and 1 Ma refer to events that took place 1 Gyr and 1 Myr in the past, respectively.

that the differentiation process took a few tens of Myr. On the basis of isotopic evidence, the age of the Earth is currently held to be 4.54 ± 0.05 Gyr. However, we cannot treat this age as the outer limit for habitability because there were several major events that happened afterward.

The first among them was the formation of the Moon. The Moon is believed to have formed from the impact of a large planetary embryo (Theia) and the Earth. This theory is based on the fact that the isotopic composition of the Moon and the Earth are virtually identical, thereby indicating comprehensive mixing and exchanging of materials between these two bodies. The date of the Moon-forming impact is quite tightly constrained by the use of multiple dating techniques including rubidium-strontium (Rb-Sr), hafnium-tungsten (Hf-W), and uranium-lead (U-Pb). Collectively, these suggest that the Moon had formed around 4.42–4.52 Ga. However, as a result of this giant impact, the Earth's primitive mantle melted and gave rise to a short-lived magma ocean that subsequently cooled and solidified. The duration of this process is not yet well understood, as it is controlled by several variables—including the temperature of the magma ocean, the extent of tidal heating, and the thickness of the atmosphere—which remain unknown. The total time for cooling and the formation of the Earth's crust and oceans is presumably between 0.02 and 100 Myr (Pearce et al. 2018). Thus, we may assign an outer boundary of 4.32–4.52 Ga for the earliest time that life could have sprung into being on Earth.

Another line of geological evidence supports the above chronology—namely, the discovery and analysis of zircons ($ZrSiO_4$), arguably one of the great breakthroughs in the past few decades. Zircons, which crystallize from melts, are the oldest terrestrial minerals that have been preserved on Earth. The oldest zircons have been dated to 4.4 Ga (Wilde et al. 2001) and were found in Jack Hills, Yilgarn Craton, Western Australia. Zircons helped usher in a major shift with regard to our understanding of the Hadean eon (4.0–4.6 Ga). Previously, the Hadean environment was believed to have been inhospitable to life owing to the alleged existence of very high temperatures and lava oceans.[2] Studies of zircons have, instead, furnished evidence for the presence of continental crust and oceans during the Hadean, potentially dating as far back as 4.4 Ga.

---

2. The use of *Hadean* to describe the Earth stems from Hades, the Greek god of the underworld.

The initial formation of continental crust "freezes" the isotope ratio of Hf, and the subsequent melting and recycling of the crust gives rise to differing $^{176}Hf / ^{177}Hf$ ratios that can be used to date the ages of different zircons. Zircons also yield information about the availability of liquid water during the Hadean epoch via the oxygen isotope ratio $^{18}O / ^{16}O$. When low-temperature water-rock interactions figure prominently, they lead to higher values of $^{18}O / ^{16}O$ in the zircons owing to the preferential incorporation of ^{18}O in minerals because of the formation of stronger chemical bonds. Thus, higher concentrations of ^{18}O than the norm have been interpreted as evidence for liquid water. The collective evidence from zircons suggests that liquid surface water and continental crust were present throughout most of the Hadean—that is, from 4.4 Ga onward.

The next major event that *may* have occurred in Earth's history is the Late Heavy Bombardment (LHB). The LHB, sometimes also called the lunar cataclysm, is believed to have occurred between 4.0–4.2 Ga and ~ 3.5 Ga. The LHB is believed to have been characterized by a sharp spike in the bombardment of the inner Solar system bodies by asteroids. The main line of evidence is based on the radiometric dating of lunar basins by argon isotopes, which suggests that a cluster of large asteroids (~ 100 km) impacted the Moon around 3.9 Ga. However, this interpretation is not universally accepted, since several authors have argued that the basins formed during the course of a sustained, but declining, bombardment. In light of the uncertainties involved, and the rapid advancements in simulations and observations, it is dangerous to make a definitive prediction at this stage. Nonetheless, it seems plausible that there were at least two stages involved (Bottke & Norman 2017):

1. A relatively short and early period of bombardment caused by leftover planetesimals in the Solar system that ended at ~ 4.4 Ga.
2. A second period of bombardment initiated at 4.0–4.2 Ga due to the migration of giant planets triggered by gravitational interactions with planetesimals. The second wave might have been characterized by a long bombardment tail, with fairly significant impacts continuing until as recently as 2–2.5 Ga.

The potential existence of the LHB is very important, since it plays a major role in defining the point at which the Earth became habitable.

Most of the studies in the twentieth century (and the early twenty-first century) concerning the origin of life have operated under the assumption that life was impossible during the LHB. In other words, these analyses either suppose that life originated for the first time after the LHB or that life was present prior to the LHB but was eradicated during the bombardment period and had to originate again. Yet, it seems to us that there are several compelling reasons for supposing that microbial life (if present earlier than the LHB) could have survived the bombardment, thereby making the Earth continuously habitable since 4.32–4.52 Ga.

- Numerical modeling indicates that only ~ 1 percent of the Earth's crust (by volume) would have been in a molten state during the LHB at any given time (Abramov et al. 2013). While the existence of intermittent melting is consistent with the zircon record, the evidence also indicates that some portion of the crust did not undergo subsequent melting. Collectively, there are grounds for supposing that at least some part of the (sub)surface was continuously habitable during the Hadean eon.
- Moreover, the Earth is home to many organisms with the capacity to survive at high temperatures (known as hyperthermophiles), even those exceeding 373 K (100 °C).[3] Hence, it seems plausible that such microbes would not have died out altogether during the LHB if they had evolved by that time.
- Another possibility worth highlighting here is that inhabited debris ejected during the LHB could have fallen back to the Earth over short ($\sim 10^3$ yr) timescales (Wells et al. 2003). If these life-bearing ejecta survived reentry, they may have been capable of reseeding the Earth even if it had been been temporarily sterilized. This mechanism could have ensured that life was almost continuously prevalent in the Hadean (4.0–4.6 Ga) and Archean (2.5–4.0 Ga) eons over the period that the LHB occurred.

All of these claims must, however, be balanced against the extent of the bombardment, which remains poorly constrained. While some studies indicate

3. Unless stated otherwise, we will hereafter assume that the temperature on the Kelvin and Celsius scales is separated by an interval of 273 as opposed to the actual 273.15.

the Earth's oceans were subjected to repeated boiling until ~ 4 Ga, other models contend that the impact rates were sufficiently low to enable the onset of habitability around 4.4 Ga. It is therefore imperative to combine accurate empirical constraints on the time-dependent impactor flux derived from the inner Solar system objects with theoretical models that predict the temporal evolution of the surface temperature for a given atmospheric and oceanic composition.

One other matter concerning the habitability of the Earth during the Hadean epoch goes by the name of the "faint young Sun problem" (Sagan & Mullen 1972). The basic premise underlying this issue stems from the fact that the solar luminosity was only ~ 70 percent of its present-day value. As noted earlier, the evidence from zircons suggests the presence of surficial liquid water during this period. Hence, a variety of mechanisms have been proposed to explain how the Earth's oceans may have avoided being completely frozen over. Most of the proposed solutions rely on higher concentrations of greenhouse gases compared to the present era (Charnay et al. 2020), although the effects of albedo (due to cloud cover), radiogenic, and tidal heating cannot be discounted.

Thus, to summarize our discussion so far, we have seen that there exist two possibilities for the earliest time life could have originated on Earth. If the LHB did not sterilize the entirety of the Earth's surface, the earliest point in time at which our planet was ostensibly habitable is approximately 4.3–4.5 Ga; in this scenario, we will henceforth use 4.5 Ga because it represents a stringent bound on the conditions for habitability. On the other hand, if the LHB was indeed catastrophic to life, the habitability window opens at ~ 3.9 Ga. As both these scenarios lie within, or just after, the Hadean eon, we will employ it as a shorthand notation to describe the environment in which prebiotic chemistry and life arose. This does not, however, imply that we are ruling out the legitimate hypothesis that life originated during the Archean epoch, i.e., around $\lesssim 4.0$ Ga.

2.1.2 The first signatures of life on Earth

Before embarking on a discussion of the fossil and phylogenetic evidence for the earliest lifeforms on our planet, we will consider a toy mathematical model (Lingam & Loeb 2017d). It is important to recognize that this model is merely a conduit for guiding our subsequent discussion rather than an accurate representation of reality.

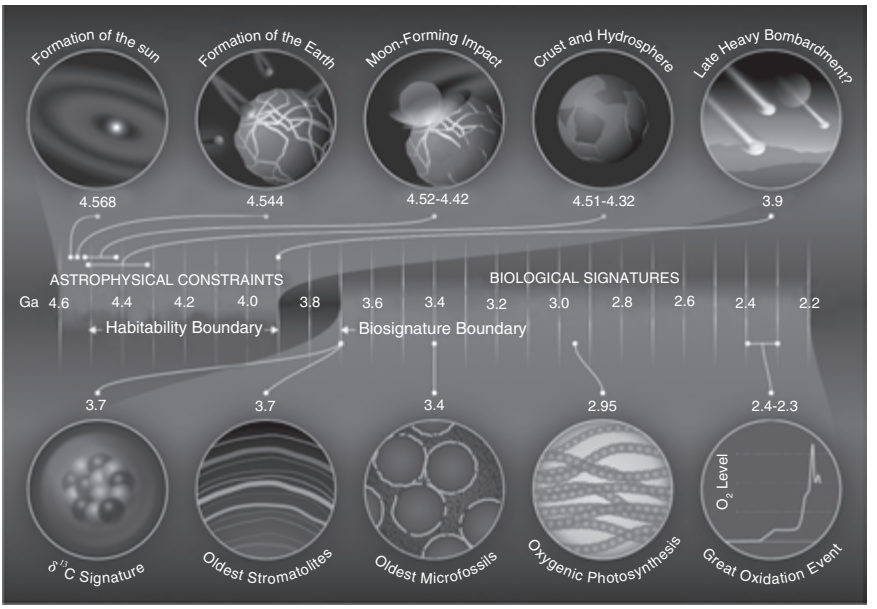


Figure 2.1 Timeline of the major events in Earth's astrophysical, geological, and biological history, with a focus on the Hadean and Archean eons—that is, the duration between 4.6 Ga and 2.5 Ga (© Mary Ann Liebert, Inc. *Source:* Ben K. D. Pearce, Andrew S. Tupper, Ralph E. Pudritz, and Paul G. Higgs [2018], Constraining the time interval for the origin of life on Earth, *Astrobiology* 18[3]: 343–364, fig. 1).

We will suppose that the total number of species on our planet (species richness) grows exponentially with time until a certain stage, after which it saturates and eventually declines when the Earth becomes increasingly uninhabitable.[4] During the exponential phase, we have

$$N_\star(\Delta t) = \exp\left(\frac{\Delta t}{\tau_c}\right) - 1, \tag{2.1}$$

where $N_\star$ is the species richness, Δt is the time that has elapsed since the Earth first became habitable, and τ_c is the characteristic *e*-folding timescale. We can also include a prefactor in front of the right-hand side, but it does not change our results since it is of order unity (Russell 1983). Note that

4. We reiterate that this model is only semi-accurate, since a more realistic model requires the superposition of three logistic curves (Purvis & Hector 2000).

(2.1) vanishes at $\Delta t = 0$ since evolution has not been initiated at this stage. The timescale for abiogenesis, denoted by $t_{0,\oplus}$, is found from $N\left(t_{0,\oplus}\right) = 1$, which yields $t_{0,\oplus} = \tau_c\,(\ln 2)$. Hence, a knowledge of $N_\star$ for present-day Earth enables us to estimate both τ_c and $t_{0,\oplus}$. The total number of microbial species is not properly constrained for modern Earth and is thus subject to intense controversy, but we will utilize the optimistic value of $N_\star \sim 10^{12}$ obtained in Locey and Lennon (2016). Fortunately, our result is quite insensitive to the choice of $N_\star$, even if we revise it downward by five orders of magnitude (Louca et al. 2019), due to the logarithmic dependence on this quantity:

$$\tau_c \approx \frac{\Delta t}{\ln N_\star}. \tag{2.2}$$

We need to consider two cases here. The first corresponds to the situation wherein the habitable epoch of the Earth was initiated shortly after the formation of the Earth—that is, around 4.5 Ga. In this scenario, we specify $\Delta t = 4.5$ Gyr, which consequently yields $\tau_c \approx 163$ Myr and $t_{0,\oplus} \approx 113$ Myr. Thus, one would expect the emergence of life to have transpired ~ 4.39 Ga. Rather intriguingly, this estimate is consistent with a recent proposal that RNA formation and the advent of proto-life occurred $\sim 4.36 \pm 0.1$ Ga (Benner et al. 2020).

The next case requires us to start the evolutionary clock at ~ 3.9 Ga after the LHB had subsided. When we choose $\Delta t = 3.9$ Ga and use (2.1), we end up with $\tau_c \approx 141$ Myr and $t_{0,\oplus} \approx 98$ Myr. Hence, in the first case we predict that life may have originated at ~ 4.4 Ga, whereas in the post-LHB scenario the first living organisms could have arisen at ~ 3.8 Ga. At the risk of putting the proverbial cart before the horse, our subsequent discussion will make the case that the earliest noncontroversial signatures of life date from ~ 3.7 Ga. However, the apparent absence of smoking-gun signatures for life before 3.7 Ga ought not be construed as evidence that life had not originated earlier. Hence, it can be seen that this timing is consistent with the values obtained from (2.1). In both cases, our model predicts that life had originated before the first unequivocal biosignatures appear in Earth's geological record.

Now, we shall commence the review of the multifarious lines of evidence commonly invoked for constraining the timing of the origin of life (Javaux 2019). As this represents an active area of research, these findings should not viewed as being the last word on the subject.

2.1.2.1 Microfossils

Owing to the recycling of the Earth's crust and high-grade metamorphism (changes in structure wrought by high temperature and pressure), the fossil record is very scanty in the Hadean and Archean eons (2.5–4.6 Ga). This issue is further compounded by the fact that the fossilized organisms are microbes (microfossils) and several abiotic processes can give rise to similar structures (Brasier et al. 2015). Hence, the identification of authentic microfossils remains a challenging endeavor. However, certain characteristics are expressly associated with microfossils and therefore perceived as genuine biomarkers. The relevant criteria include (1) the existence of distinctive three-dimensional walled compartments (cell lumina) that are usually devoid of any internal material (i.e., mostly hollow), (2) the carbonaceous composition of these wall components (specifically composed of insoluble organic matter known as kerogen), enriched in nitrogen, and (3) the existence of multiple specimens at the same site in differing stages of preservation.

With the advent of sophisticated tools such as transmission electron microscopy (TEM) and laser Raman spectroscopy, it has become possible to analyze whether potential microfossils fulfill (1) and (2). If we impose these stringent constraints, many of the oldest candidates fail to satisfy these requirements. It is, however, worth pointing out that failure to clearly establish these two criteria does not, by itself, rule out these candidates. The earliest microfossils that seem to fulfill the requisite conditions are from the Strelley Pool Formation and the Apex chert in Western Australia. The former have been dated to 3.43 Ga and appear to have been derived from sulfur-metabolizing bacteria (Wacey et al. 2011). The latter have been assigned the slightly older age of 3.465 Gyr and are believed to comprise primitive photosynthesizers and methane producers and consumers, based on the analysis of carbon isotope ratios (Schopf et al. 2018).

A recent discovery that has attracted much attention concerns putative microfossils from the Nuvvuagittuq supracrustal belt located in Quebec, Canada (Dodd et al. 2017) that have been dated to 3.77–4.28 Ga. The tubular and filamentary structures were discovered in sedimentary rocks that have been interpreted as evidence of ancient hydrothermal vent precipitates, and they possess morphologies that are akin to those found in modern hydrothermal precipitates. Another notable point is that these structures are coexistent with carbonaceous material in the rocks. Although their great age and location—to wit, near potential hydrothermal vents, which are

regarded as hospitable environments for abiogenesis—makes them appealing candidates, we caution that definitive evidence for either (1) or (2) is lacking. Hence, the question of whether these structures are indeed evidence of early life remains unresolved at this juncture.

2.1.2.2 Stromatolites

Stromatolites ("layered rock" in Greek) are layered structures that often take the form of sheets, columns, or mounds and typically entail the interaction of microorganisms with flowing water and sediments (Bosak et al. 2013). In one scenario, they form due to the action of photosynthesizing microbes that migrate upward—namely, toward the light (an example of phototaxis)—to avoid being eventually buried in sedimentary grains and silt, as this event would reduce their photosynthetic yield. As a consequence of this movement, a new microbial mat layer is formed. The process is repeated multiple times, until it leads to the formation of stromatolites—that is, several layers of sedimentary rock and carbonates.

Several stromatolite-like structures have been discovered in Western Australia, with ages ranging from 3.43 to 3.48 Ga. The Dresser Formation of the Pilbara Craton, in particular, has yielded very well-preserved stromatolites dating back to 3.48 Ga (Baumgartner et al. 2019). Some of the oldest evidence for stromatolites, which has been contested by Allwood et al. (2018), originates from the Isua Supracrustal Belt (ISB) in Greenland by Nutman et al. (2016) and has been dated to 3.7 Ga. A noteworthy aspect of the stromatolite-like structures discovered in the ISB is that they share certain morphological and chemical similarities with younger stromatolites (of varying ages) discovered at multiple locations in Western Australia (Nutman et al. 2019). Figure 2.2 depicts the ISB stromatolites along with a comparison against more recent stromatolites from Western Australia.

2.1.2.3 Carbon isotope ratios

One of the most commonly used isotopes for detecting the existence of biogenic activity is ^{13}C; on Earth, about 1.1 percent of natural carbon is present in this form. The chief advantage of using the ^{13}C / ^{12}C isotope ratio stems from the fact that biological activity inherently discriminates between ^{13}C and ^{12}C. In general, as a result of the lower mass, lighter isotopes tend to be more mobile and diffuse faster relative to heavier isotopes. Hence, studies

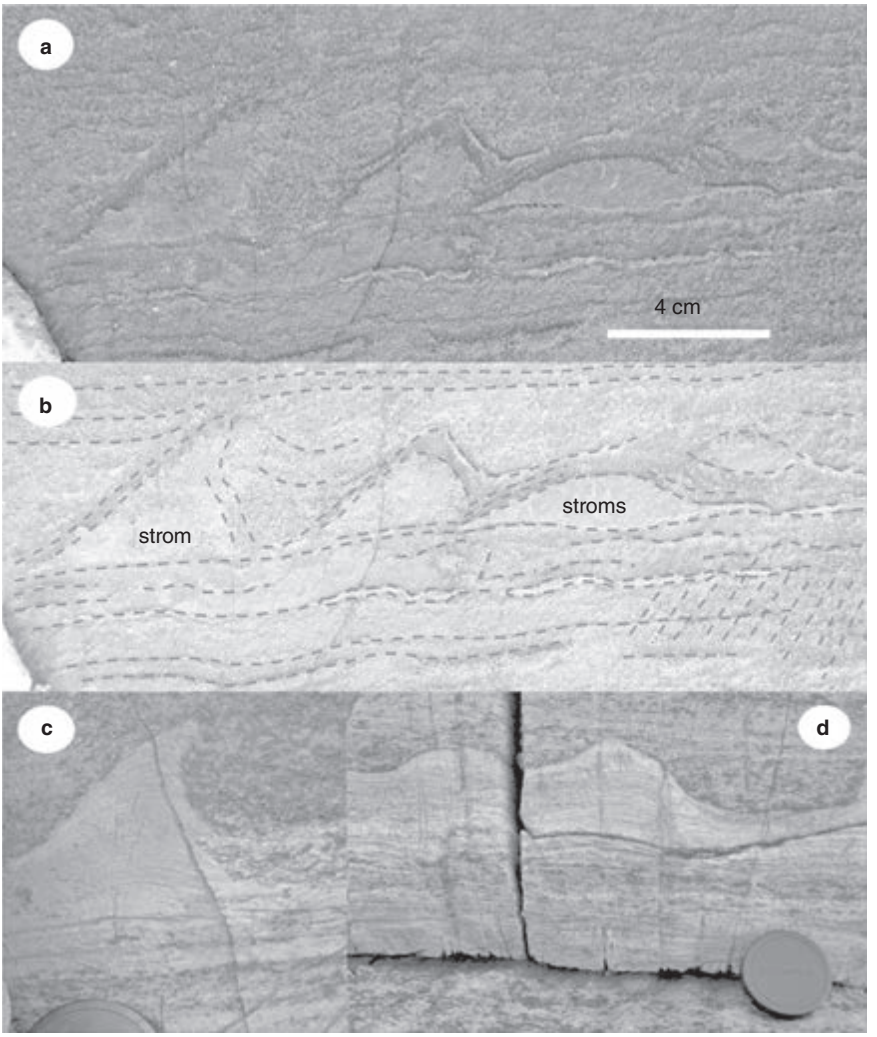


Figure 2.2 Panel (*a*) displays the 3.7 Ga stromatolites from the ISB located in Greenland. Panel (*b*) provides an interpretation for (*a*) with *strom* and *stroms* designating an isolated stromatolite and a collection of stromatolites, respectively. Panels (*c*) and (*d*) depict interlinked dome-shaped stromatolites from the more recent (~ 2 Ga) Wooly Dolomite carbonate platform situated in Western Australia. (© Macmillan Publishers Limited. *Source:* Allen P. Nutman, Vickie C. Bennett, Clark R. L. Friend, Martin J. Van Kranendonk, and Allan R. Chivas [2016], Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures, *Nature* 537[7621]: 535–538, fig. 1.)

have established that photosynthetic pathways and many key enzymes tend to favor ^{12}C on account of its lighter mass. In other words, as a consequence of enzymatic activity, biotic material displays a tendency toward preferential depletion of ^{13}C. This depletion is measured by the ratio $\delta^{13}C$ defined as follows:

$$\delta^{13}C = \left(\frac{\left(\frac{^{13}C}{^{12}C}\right)_{\text{sample}}}{\left(\frac{^{13}C}{^{12}C}\right)_{\text{standard}}} - 1 \right) \times 1000\,‰, \tag{2.3}$$

where the numerator and denominator on the right-hand side represent the isotope ratios for the given sample and an established reference standard, respectively. Highly negative values of $\delta^{13}C$ measured for organic matter within rocks are indicative of biological activity.

An important point to bear in mind here is that the rocks in which the organic matter is ensconced should preferably be sedimentary in origin. This additional requirement is rendered essential because there exist abiotic mechanisms, such as the removal of carbon dioxide during metamorphism, that can give rise to negative values of $\delta^{13}C$ in nonsedimentary rocks. Moreover, the presence of sedimentary rocks furnishes indirect evidence for habitable conditions (e.g., liquid water and moderate temperatures). Some of the oldest sedimentary rocks (as established by TEM) exhibiting suitable depletion of ^{13}C—to wit, featuring $\delta^{13}C = -19\,‰$—are derived from the aforementioned ISB in Greenland and purportedly date back to ~ 3.7 Gyr (Ohtomo et al. 2014).

Several other samples are distinguished by significantly negative values of $\delta^{13}C$ and are distinctly older than the ISB rocks described above. However, in most cases, the rocks are either not sedimentary or may have undergone contamination by high-grade metamorphic activity. In particular, preserved graphite dating from 4.1 Ga and extracted from the Jack Hills zircons, whose importance was discussed in Section 2.1.1, evinced $\delta^{13}C = -24 \pm 5\,‰$ (Bell et al. 2015). While it has been established that this striking negative value is probably not because of contamination by subsequent geological activity, an abiogenic origin for this sample could not be definitively ruled out.

2.1.2.4 Summary

We have seen that multiple lines of evidence seemingly converge to yield a timing of 3.7 Ga for the earliest unambiguous signatures of biological activity. There are some grounds to suppose that life was existent at even

earlier stages (4.1–4.3 Ga), but further research is imperative before the biosignatures that yielded these dates become widely accepted.

In closing, we note that another line of evidence (albeit indirect) for deducing the timing of early life is molecular phylogenetics. It involves the reconstruction of evolutionary history via genetic sequencing to map out the relationships among different biological groups. For reconstructing the early evolutionary timeline of the Earth using phylogenetic methods, molecular clocks must be brought into play. The basic principle underlying these clocks is that mutations occur at some specified (typically constant) rate, implying that the time at which two organisms diverged can be calculated by determining their gene sequences and computing the cumulative number of mutations that separate the two sequences. The clock is calibrated against the dates at which different clades (groups of organisms with a common ancestor) evolved, based on the available geological evidence, thereby enabling the reconstruction of the origination times for other nodes in the tree of life.

A major issue with using molecular clocks, especially as one approaches the Hadean and Archean eons, is that evolutionary rates vary significantly among organisms and even between genes within the same organism. Hence, the timing from molecular clocks is subject to large margins of error and variability, thus leading to very different estimates by different research groups for the same evolutionary event. With these caveats in mind, we note that a recent molecular clock analysis suggests that the Last Universal Common Ancestor (LUCA) originated much earlier than 3.9 Ga, i.e., predating the cessation of the putative LHB, although the exact date remains uncertain (Betts et al. 2018).[5]

2.2 THE BASIC BUILDING BLOCKS OF LIFE

Let us embark on a brief survey of the basic building blocks necessary for constructing the biomolecules that are essential ingredients of the first cells (protocells). The subsystems that comprise a minimal protocell were

5. While LUCA was the most recent common ancestor of all life on our planet, it should be concomitantly recognized that other organisms could have predated it—in other words, there is no reason to presume that LUCA was also the very first living organism on Earth.

cogently articulated by Tibor Gánti in 1971. Gánti's unit of life, the *chemoton,* is endowed with metabolism, information storage and transmission, and a membrane (Gánti 2003). These functions are carried out by proteins, nucleic acids, and lipids in all living organisms on Earth.

2.2.1 Amino acids

Amino acids are organic compounds characterized by the presence of two functional groups: amine (-NH_2) and carboxyl (-COOH). These two functional groups in biology are attached to a carbon atom that goes by the name of α-carbon. The importance of amino acids stems from the fact that they can undergo polymerization to form proteins. This process occurs via the formation of peptide bonds (-CO-NH-) via the removal of O and H from the carboxyl group and the removal of H from the amine group. Thus, peptides are formed via a condensation reaction from two or more amino acids, entailing the formation of water (H_2O) as a product. It is imperative, however, to recognize that peptides may be synthesized through other avenues, such as compounds known as aminonitriles (Canavelli et al. 2019).

The importance of proteins in the context of carrying out myriad cellular functions has been thoroughly documented. The diversity of proteins stems from their ability to selectively (and tightly) bind to target molecules. The best-known role of proteins is their capacity as enzymes, which serve as highly efficient catalysts for particular chemical reactions. They achieve this objective by lowering the activation energy (E_a) because the reaction rate k is given by

$$k = A \exp\left(-\frac{E_a}{k_B T}\right), \tag{2.4}$$

where A is a constant pre-exponential factor, while T denotes the temperature. Owing to the exponential dependence of the reaction rate on the activation energy, lowering the energy barrier via enzymatic action increases the rate by many orders of magnitude. The vast majority of all metabolic pathways in the cell require the assistance of enzymes to take place at adequately high rates. Moreover, enzymes are necessary for carrying out the reactions involved in the replication of DNA. Proteins also play a vital role in the transport of molecules and ions across the cell membrane—for example, against the concentration gradient (transport from low to high concentrations).

The genetic code is responsible for translating the information encoded in nucleic acids (DNA / RNA) to synthesize proteins. Life-as-we-know-it uses only twenty-two amino acids in the production of proteins, of which only twenty of them are encoded in the standard genetic code. In other words, the genetic code specifies the instructions for the synthesis of proteins from amino acids. We shall not delve into the specifics of how this is undertaken, but it suffices to say that the process entails the use of codons (nucleotide triplets) to encode amino acids. Of the twenty-two proteinogenic amino acids, only twenty-one of them are used in eukaryotes (organisms whose cells have well-defined nuclei that contain genetic material), with pyrrolysine (Pyl) being used only by some archaea and bacteria. Among these twenty-one amino acids, the odd one out is selenocysteine (Sec) because it is not encoded in the standard genetic code, and it is not present in all lineages of life. Figure 2.3 depicts the basic characteristics of the twenty-one proteinogenic amino acids that occur in eukaryotes.

However, this point brings up an immediate question: Why does life-as-we-know-it only use twenty amino acids for the most part, despite the fact that over 500 exist in nature? The answer acquires considerable relevance when it comes to searching for signs of life via remote-sensing or in situ life-detection missions. Two diametrically opposite explanations instantly spring to mind. The first is that the selection of these twenty amino acids was purely a matter of chance. The second is that these twenty amino acids were selected because they collectively possess certain properties that make their assimilation into biology quasi-inevitable. This dichotomy between chance and necessity, or chance and inevitability, is a fundamental one that will recur throughout this book.

However, the truth is plausibly situated somewhere in the midst of these two extremes—to wit, the use of these twenty amino acids may have stemmed from a combination of chance and inevitability. A number of factors have potentially influenced the consolidation of the twenty standard amino acids in biology and the accompanying issue of the (co)evolution of the genetic code (Knight et al. 2001; Koonin & Novozhilov 2017). A classic summary of this subject can be found in Weber and Miller (1981); the reader should also consult Higgs and Pudritz (2009), Philip and Freeland (2011) and Doig (2017) for later analyses. We will discuss some of the salient factors below. Before embarking on this exposition, we remark that enzymes comprised of less than twenty amino acids have proven

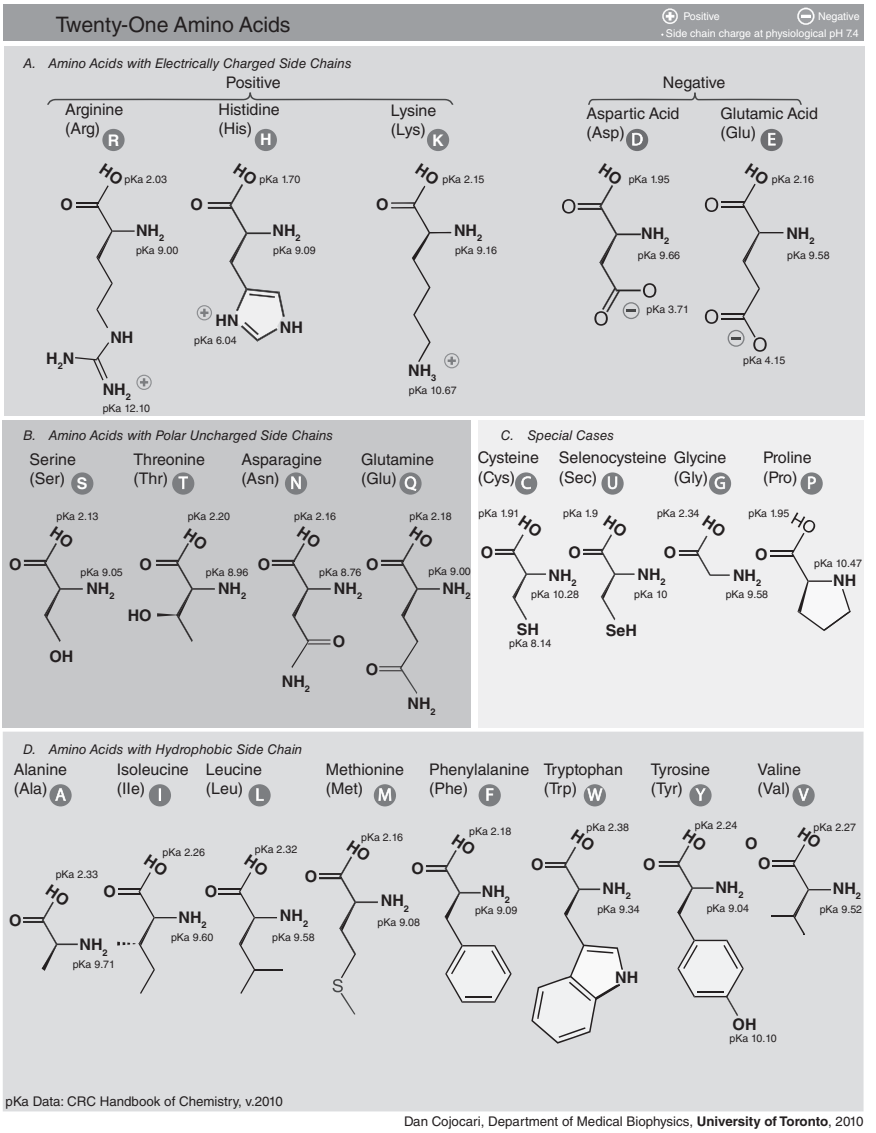


Figure 2.3 The chemical structures and nomenclatures of the twenty-one proteinogenic amino acids found in eukaryotes (cells with nuclei). Of these twenty-one molecules, only selenocysteine (Sec) is not encoded in the genetic code. (Dan Cojocari / Wikimedia Commons / CC BY-SA 3.0 / GNU Free Documentation License.)

effective at carrying out their engineered tasks: as few as nine were adequate to construct a serviceable enzyme (Walter et al. 2005).

If a particular amino acid belonging to a given category is formed more easily by prebiotic chemistry, it seems plausible that it could be availed in protein synthesis with greater ease, provided that all other factors are held equal. The analysis of the abundance of amino acids from a variety of sources (e.g., meteorites) has confirmed the intuitive notion that the simplest amino acids have the highest abundances. This feature is manifested since simpler amino acids typically possess a lower Gibbs free energy of formation. The concentration C_{rel} of a particular proteinogenic amino acid relative to glycine (the simplest amino acid) is

$$C_{\mathrm{rel}} \approx 15.8 \exp\left(-\frac{\Delta G_s}{31.3\,\mathrm{kJ/mol}}\right), \tag{2.5}$$

where ΔG_s is the Gibbs free energy of formation for the amino acid at the surface of an ocean at 291 K (18 °C) and a pressure of approximately 1 bar.

The amino acids utilized must be capable of facilitating the diverse roles played by proteins, implying that the chosen amino acids should enable proteins to acquire their characteristic structures and functions. The latter two features depend on basic chemical properties such as the size, acidity, charge, and hydrophobicity of amino acids.[6] Hence, if the twenty amino acids were optimally selected, they ought to be characterized by higher coverage—to wit, jointly composed of greater breadth (viz., the difference between maximum and minimum allowed values) and evenness (i.e., lower variance)—insofar as the selected properties are concerned, relative to the scenario where they had been chosen completely at random. Although further research is called for, there are strong indications that the set of proteinogenic amino acids fulfills this criterion (Ilardo et al. 2019).

One other notable issue regarding the selection of amino acids pertains to the stability of both these molecules and their polymers (peptides) in water. If amino acids are subject to rapid decomposition, they are not likely to accumulate in sufficiently high concentrations in aquatic environments. In turn, this can pose difficulties vis-à-vis their polymerization to yield peptides and proteins. The factors discussed herein, in tandem with

6. Hydrophobicity is a thermodynamic property of certain nonpolar molecules that leads to their clumping and segregation from water molecules.

other considerations, led Weber and Miller (1981) to argue that fifteen of the twenty amino acids would constitute the building blocks of extraterrestrial life as well. In closing, we observe that our discussion has dealt with amino acids in isolation, but this ignores the possibility of their coevolution with the genetic code (Koonin & Novozhilov 2017).

2.2.2 Nucleobases

Nucleobases, sometimes referred to as nitrogenous bases, are one of the primary building blocks of nucleic acids. The five canonical nucleobases—adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U)—form the basic units of the genetic code. Of these five nucleobases, T is present in deoxyribonucleic acid (DNA), whereas U occurs in ribonucleic acid (RNA); the other three nucleobases are common to both RNA and DNA. The canonical nucleobases are depicted in Figure 2.4, from which it is evident that the only difference between T and U is the presence of an extra methyl ($-CH_3$) group in the former.

In a similar vein, this figure establishes that A and G share a common structure. The underlying reason is that A and G are both derived from an organic compound by the name of purine, owing to which they are widely known as purine bases. A subtle point worth highlighting is that adenine possesses the chemical formula $C_5H_5N_5$, which represents the pentamer (i.e., polymer with five monomer units) of hydrogen cyanide (HCN). Hence, based on this knowledge, we may conjecture that the synthesis of adenine from HCN or its derivatives ought to be feasible. It is not surprising that, in a landmark experiment, Joan Oró (1960) was able to synthesize and validate adenine for the first time by heating aqueous solutions of ammonium cyanide at temperatures below 373 K.

Figure 2.4 reveals that C and T (as well as U) have a different structure compared to A and G. These nucleobases are derived from pyrimidine and were therefore christened the pyrimidine bases. In DNA, the four canonical nucleobases form base pairs by means of hydrogen bonding. Thus, the canonical nucleobases are a vital component of DNA, the latter of which is responsible for the storage of biological information in all known living organisms on Earth.[7] Hence, DNA plays a vital role in the reproduction

7. This statement implicitly assumes that viruses are not "living organisms" per se, but the validity of such a claim remains a highly contentious point. Irrespective of their biological

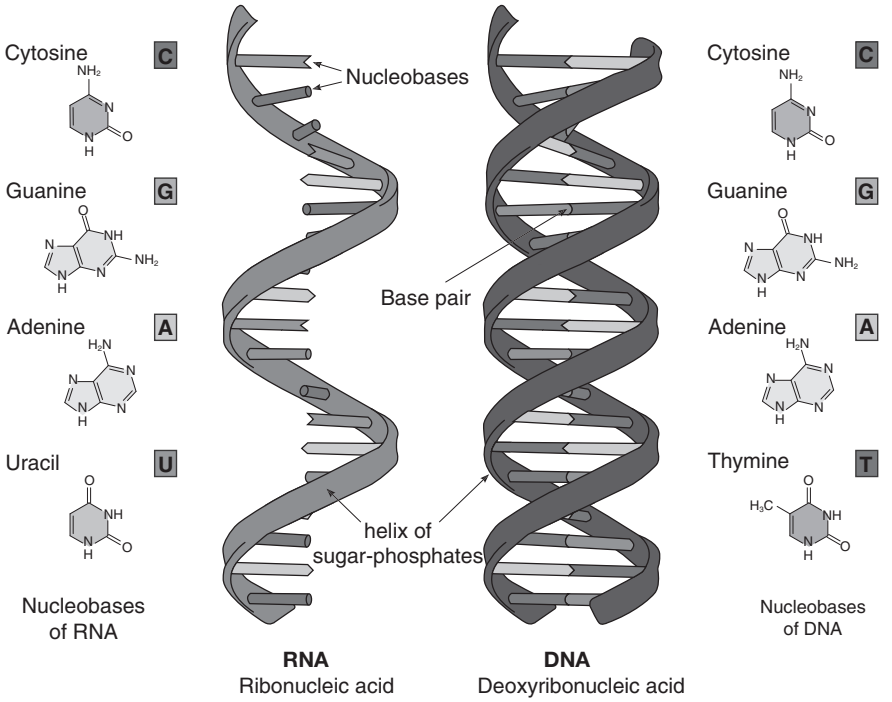


Figure 2.4 The five canonical nucleobases on Earth are depicted along with the structure of their respective nucleic acids. The so-called Watson-Crick base pairs (A-T and C-G) comprise the rungs of DNA by virtue of hydrogen bonding. (Sponk / Wikimedia Commons / CC BY-SA 3.0 / GNU Free Documentation License).

and functioning of all life on our planet. An in-depth discussion of RNA is deferred to Section 2.5, which deals with the paradigm of the RNA world.

Owing to this importance of nucleobases, we are confronted with the same question that we encountered in the context of amino acids: Were the canonical nucleobases the only available compounds? If not, what was the rationale behind their selection? While there is no definitive answer at this stage, a couple of points regarding the former question are worth highlighting. First, we observe that certain unusual nucleobase analogs have been

---

classification, viruses might have played an important role in shaping the origin and evolution of life on Earth through an evolutionary "arms race" and the (horizontal) transfer of genes with cellular organisms. The reader can find an overview of the status and significance of viruses in Berliner et al. (2018).

detected in carbonaceous meteorites, which do not occur on Earth, thereby suggesting an extraterrestrial origin. Second, many alternative pyrimidine and purine bases have been synthesized in laboratory experiments—for example, by subjecting urea solutions to freeze-thaw cycles. Lastly, artificial RNA and DNA with modified nucleobases are known to exist. The synthesis of the so-called *Hachimoji* DNA and RNA indicates that nucleic acids comprising eight nucleobases are technically viable (Hoshika et al. 2019); this word roughly translates to "eight-letter" in Japanese. On the whole, there is little doubt that other candidate nucleobases were prevalent in the Hadean environment.

Answering the second question (i.e., the reasons behind the selection of the canonical nucleobases) is much harder. One important breakthrough in recent times concerns the photostability of the canonical nucleobases. Laboratory experiments have established that the canonical nucleobases are relatively stable when subjected to ultraviolet (UV) radiation compared to most (but not all) of the other candidates (Rios & Tor 2013).[8] This stability is measured by evaluating the lifetime of the excited states induced by radiative excitation. It has been found that the canonical nucleobases have a lifetime of picoseconds ($\sim 10^{-12}$ s) and that they return to their ground states via effective nonradiative pathways (Beckstead et al. 2016). However, the issue of photostability does not, by itself, explain the selection of the canonical nucleobases, especially if they were incorporated into biological functions in a dark (e.g., underwater) environment. In the same vein, the modest stability of nucleobases in water—given that they are subject to fairly rapid hydrolysis (chemical breakdown in water)—does not yield any clear-cut information about why the canonical nucleobases were chosen; if anything, it raises the opposite question, as we shall chronicle in more detail later.

Other factors that may have played a role in the selection of the canonical nucleobases encompass the effective storage and transmission of genetic information by nucleic acids and the stability of base pairs in water at varying temperature, pH, and ionic composition. A major issue with the canonical nucleobases, which we shall elaborate on in Section 2.5, is that the formation of chemical bonds with sugars (i.e., N-glycosidic bonds) to eventually produce nucleotides (i.e., monomers of nucleic acids) has proven to be quite

8. When the nucleobases are assumed to take part in stacking or base pairing, their photostability is further increased.

challenging in laboratory settings. As a result, it has been proposed that alternative nucleobases were involved in the synthesis of early genetic polymers. Melamine ($C_3H_6N_6$) and barbituric acid ($C_4H_4N_2O_3$) seem to be promising candidates in this regard as high yields of nucleotides and base pairs were spontaneously obtained in water after the addition of these molecules (Cafferty et al. 2016).

Hitherto, we took it for granted that exactly four nucleobases (two base pairs) were utilized in the first genetic polymer, but there are no reasons to suppose that this was truly the case. Szathmáry (2003) has, however, outlined theoretical grounds for contending that four nucleobases might have represented the optimal number at the putative stage during which nucleic acids were responsible for both replication and catalysis: the considerations taken into account included replicability, stabilizing selection, and attaining maximal evolvability. Although some of these criteria are attractive, perhaps even plausible, they are also compatible with the premise that the deployment of the four letters of the genetic alphabet was an important (but nonessential) feature that was selected at a later period.

One potential scenario is that only two or three nucleobases were the sine qua non of the genetic alphabet at its inception. At first sight, this conjecture gains some credibility because cytosine has a very short half-life in water, as pointed out in Section 2.3.7. In this event, the G-C base pair may not have featured in the earliest genetic polymers. The other possibility, which has been explored extensively by numerous scientists, is that the first nucleic acids could have featured an expanded genetic alphabet (i.e., with more than four nucleobases); a notable example is the Hachimoji DNA and RNA mentioned earlier. In this context, we point out that codons (nucleotide triplets) can encode a maximum of $4^3 = 64$ amino acids, provided that there are a total of four nucleotides. Hence, in order to encode for twenty amino acids, assuming that codons are triplets and nucleotides exist as base pairs (i.e., the total number of nucleotides is even), at least four nucleotides are necessary; of course, as remarked previously, it is not imperative for extraterrestrial life to comprise twenty amino acids.

All things considered, it seems plausible that the building blocks of extraterrestrial life will not be universally composed of the canonical nucleobases. If this conjecture is correct, it would also imply that the storage and transmission of information may very well rely on alternative biopolymers and therefore not ineluctably on RNA and DNA (Cleaves et al. 2019).

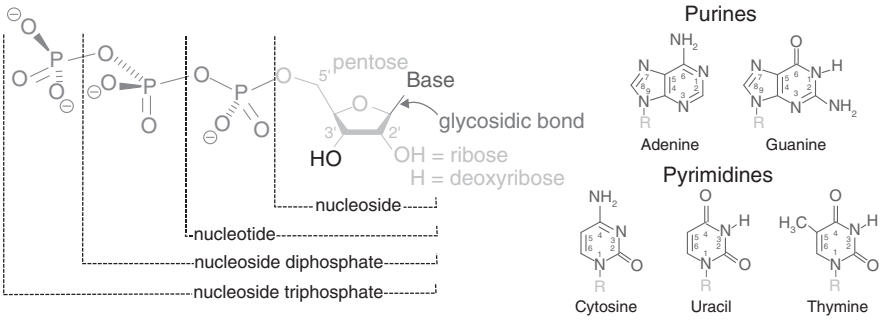


Figure 2.5 The components of nucleosides, nucleotides, and nucleoside triphosphates. The differences between the nucleotides of RNA and DNA are that ribose and uracil should be substituted by deoxyribose and thymine, respectively. (Boris [PNG], Sjef [SVG] / Wikimedia Commons.)

2.2.3 Carbohydrates (Sugars)

Carbohydrates are another important class of organic molecules comprising carbon, hydrogen, and oxygen. They are usually, but not always, characterized by the general formula $C_m(H_2O)_n$, where m and n are not necessarily equal. An exception to this norm is the monosaccharide deoxyribose that has the chemical formula $C_5H_{10}O_4$ since the ratio of hydrogen to oxygen atoms does not equal 2:1; in contrast, ribose ($C_5H_{10}O_5$) obeys the above rule with $m = n = 5$. Carbohydrates play a variety of roles in biology such as providing structural support (e.g., cellulose in plants) and storing energy.

What we are more interested in, however, is their role insofar as the origin of life is concerned. Two important functions stand out immediately. The formation of N-glycosidic bonds between sugars (ribose in RNA and deoxyribose in DNA) and nucleobases leads to nucleosides. The building blocks of RNA and DNA (nucleotides) consist of the phosphate group conjoined with the appropriate nucleosides. Likewise, when the triphosphate group is attached to nucleosides, the resulting compounds are known as nucleoside triphosphates (NTPs). Of the NTPs, the best known is adenosine triphosphate (ATP), which is often called the *energy currency* of the cell. Figure 2.5 illustrates the chemical structures of nucleotides and NTPs. We will not discuss ATP in further detail at this stage, but we have already seen that sugars, especially ribose and deoxyribose, play an important role in both nucleic acids and metabolism.

Of the two, deoxyribose is derived from ribose by the removal of an oxygen atom, and it must be emphasized that the latter also occurs in the analog of ATP. Thus, in light of the centrality of ribose in prebiotic chemistry, we must ask ourselves whether alternatives to ribose exist. On the one hand, ribose appears to be well suited for maintaining the structure of RNA. On the other hand, several difficulties are associated with synthesizing nucleotides and nucleic acids from ribose, which will be explicated in Section 2.5. Furthermore, laboratory experiments have demonstrated that ribose can potentially be replaced with threose ($C_4H_8O_4$) or glycerol ($C_3H_8O_3$), the latter of which is not a cyclic sugar.

Hence, on account of these reasons, we cannot and ought not dismiss the possibility that ribose was absent from the first genetic polymers—to wit, the precursors of RNA and DNA. If the role of ribose could be taken over by alternative (in)organic compounds, it automatically follows that extraterrestrial biology need not encompass ribose or deoxyribose scaffolding.

2.2.4 Lipids

> Traube, a Berlin physiologist, has succeeded in making artificial cells. Needless to say, they are not completely natural cells, being without a nucleus. If a colloidal solution, e.g. of gelatine, is combined with copper sulphate, etc., this produces globules surrounded by a membrane that can be made to grow by intussusception. Here, then, membrane formation and cell growth have left the realm of hypothesis!
>
> —Karl Marx to Pyotr Lavrov, June 18, 1875,
> *Collected Works*: *Vol. 45. 1874–1879*

Lipids refers to organic molecules that are insoluble in water and typically consist of fatty acids and their derivatives. Fatty acids represent a class of carboxylic acids—they are characterized by a long chain of carbon atoms with a carboxyl group (-COOH) at the end. Other notable examples of lipids include triglycerides (fats) that are derived from three fatty acids and glycerol ($C_3H_8O_3$).

As lipids include a very diverse array of compounds with varying functional groups, they are capable of fulfilling many distinct functions. One feature unique to many lipids is that they have a hydrophobic end and a hydrophilic (water-loving) end that is charged; lipids endowed with these properties are termed *amphiphilic*. One of the most important types

of lipids are phospholipids, whose structure consists of fatty acid "tails" (hydrophobic) and a phosphate "head" (hydrophilic). The importance of phospholipids stems from the datum that they are core components of modern cell membranes. It must also be recognized that fatty acids are capable of combining with other prebiotic molecules to yield peptolipids and nucleolipids. The former are theoretically capable of facilitating transport across protomembranes, whereas the latter have interesting chemical properties (e.g., hydrogen bonding) that might have led to their participation in the replication of the first genetic polymers.

The importance of membranes is evident when we survey known extant and extinct life on Earth since all of them are based on the existence of cells and the exchange of energy and matter between the external world and the internal components of cells. The basic functions of membranes include preventing metabolic products from diffusing out of the cell, excluding toxic materials from reaching the cell, and assisting in the transport and concentration of nutrients into the cell. Owing to the manifold advantages accorded by membranes, the formation of cell-like compartments (vesicles) with semipermeable closed boundaries that are capable of growth and reproduction has been posited by some as comprising an essential prerequisite for the origin of protocells and life (Zhu & Szostak 2009; Sarkar et al. 2020). This perspective has a long and distinguished history, with roots dating back to the nineteenth century, as indicated by the above quotation from Karl Marx.[9] Clearly, going by the lipid world scenario, the prebiotic synthesis of lipids is an essential ingredient for the emergence of vesicles and primitive protocells.

2.2.5 Phosphates

Unlike all of the previous molecules that we have discussed, the phosphate (PO_4^{3-}) anion is clearly not an organic compound. Yet, as we have seen, phosphates occur in discussions of nucleic acids (the sugar-phosphate backbone), ATP (energy transfer), and phospholipids (component of cell membranes). This fact raises the intriguing question: Why phosphates? This question has a long history, and it was thoroughly analyzed in the reviews

9. The authors are grateful to Michael Russell for bringing this document to their attention in connection with the origin of life.

by Gulick (1955), Westheimer (1987), and Kamerlin et al. (2013). The gist of these arguments is as follows.

First, phosphate links together two of the sugar molecules in nucleic acids—that is, it acts as the glue for joining the nucleotides. Even if extraterrestrial life used nucleic acids other than RNA / DNA, it would still be necessary to link the monomers together. In turn, this would imply that the chemical species has to be at least divalent. Second, the linking unit should have a net charge in order to prevent the molecule from being ejected out of the cell membrane; the ionization ensures that the compound is insoluble in the lipids that comprise the membrane. This requirement implies that at least three ionizable groups exist. Third, in order to maintain the stability of the molecule against hydrolysis, a negative charge is required to repel the hydroxide ion, which is nucleophilic in nature (an electron donor). Last, while the phosphate ester bonds that link the nucleotides are difficult to cleave, they are also very versatile and sufficiently reactive under the right circumstances owing to their capacity to regulate electrostatic forces.

The long and short of the above discussion is that no real alternatives to phosphates have been identified thus far in extant organisms. In 2011, it was reported that arsenate was capable of taking over the role of phosphate in the extremophile bacterium GFAJ-1 (Wolfe-Simon et al. 2011), but this purported discovery attracted substantive criticism, and the validity of the experimental results have been questioned (Reaves et al. 2012). Even if we do adopt the premise that phosphates are *currently* an inescapable part of life, it is nonetheless quite possible that phosphate was incorporated into biological functions at a later juncture. For example, computational analysis of metabolic networks has indicated that phosphate-free metabolic pathways based on thioesters (involving the functional group R-S-CO-R′), and mediated by inorganic enzymes, might have facilitated the origin of life (Goldford et al. 2017). Alternatively, the phosphate group in RNA could be replaced by glyoxylate ($C_2HO_3^-$), although the respective yields of the nucleotide analogs are comparatively low.

One of the main reasons why scientists are interested in identifying viable alternatives to phosphate is because of its unwelcome combination of low solubility in water (at neutral or alkaline pH) and low reactivity. To be precise, most of the phosphate in Earth's crust is locked up in the form of apatite minerals that are relatively insoluble in water. In addition, the formation of organophosphates entails dehydration reactions (to wit, loss of water), thereby making it difficult for them to transpire in aqueous

solutions as the reverse reaction (i.e., hydrolysis) can also efficiently occur. Collectively, these difficulties have come to be dubbed the *phosphate problem* in academic parlance.

Because of the difficulties associated with the availability of phosphates, several studies have chosen to focus instead on the synthesis of reduced phosphorus (P). Two of the most commonly produced forms of reduced P are phosphites and phosphides, with oxidation states of +3 and 0, respectively, as compared to phosphates, in which the oxidation state of P is +5. Reduced forms of P are characterized by their higher solubility and reactivity, and they can duly undergo subsequent oxidation to result in the formation of phosphodiester bonds that link the nucleotides. We shall outline a few avenues by which reduced P is synthesized and provide rough estimates of the annual yields (Pasek et al. 2017).

The first pathway entails the production of phosphites by means of lightning (electrical discharges), a high-energy source, acting directly on minerals containing trace amounts of P. The production rate of reduced P (in mol/yr) is

$$\mathcal{S}_P \sim 10^5 \,\text{mol/yr} \left(\frac{R}{R_\oplus}\right)^2, \tag{2.6}$$

where $\mathcal{S}_P$ denotes the production rate of dissolved P and R is the radius of the planet. In most cases where the scaling is proportional to R^2, it stems from the surface area, assuming that the rate per unit area is held constant. Although this production rate is much lower than the two alternatives discussed below, this value could be higher when lightning acts on volcanic eruptions comprising reduced gases and phosphate-containing minerals. In fact, the production rate per unit area (mol m^{-2} yr^{-1}) might approach $\sim 10^{-5}$ mol m^{-2} yr^{-1} in this environment—about five orders of magnitude higher than the global estimate.

The next source that merits study is the exogenous delivery of reduced P by meteorites in the form of the phosphide mineral schreibersite, whose chemical formula is $(\text{Fe,Ni})_3\text{P}$. Although the exact amount of reduced P delivered by meteoritic sources is subject to much variability, by two orders of magnitude or even more, taking the geometric mean yields a makeshift production rate of

$$\mathcal{S}_P \sim 3 \times 10^8 \,\text{mol/yr} \left(\frac{R}{R_\oplus}\right)^2, \tag{2.7}$$

provided that the mass flux of meteorites impacting the planet is comparable to the amount received by the Earth during the Hadean eon; we reiterate that the mass flux of impactors is subject to a high degree of variability, owing to which this estimate should be used with due caution. Nevertheless, the above mean value is not inconsistent with the theoretical predictions laid out in Ritson et al. (2020). The reduced P delivered by meteorites, after reactions with water and H_2S in the presence of UV radiation, is capable of undergoing full oxidation to yield phosphate as the final product (Ritson et al. 2020).

Although we have discussed reduced sources of P, it must be appreciated that apatite minerals are not invariably highly insoluble in water. On present-day Earth, most of the dissolved P inventory in the oceans is a consequence of continental weathering and transport by rivers. While the evidence from zircons (see Section 2.1.1) indicates that continental crust was present during the Hadean era, theoretical models and empirical data suggest that the volume of continental crust was lower than today. In this event, most of the weathering of apatite minerals would occur on the ocean floor. The corresponding production rate of dissolved P depends on the pH of the oceans at that time. By utilizing a heuristic approach that we elucidate later (see Section 7.6.2.1), we obtain

$$\mathcal{S}_P \sim 1.3 \times 10^9 \,\mathrm{mol/yr} \left(\frac{R}{R_\oplus} \right)^2 , \tag{2.8}$$

where we have assumed that the pH of the oceans in the twilight of the Hadean period (~ 4 Ga) was approximately equal to the early Archean value of ~ 6.6 (Halevy & Bachan 2017; Krissansen-Totton et al. 2018a). It should, however, be recognized that the average pH of Hadean oceans may have been higher due to the weathering of impact-generated ejecta by carbonic acid; in particular, a pH of ~ 7.9 (which is close to the modern value) at 4.3 Ga has been predicted by numerical models (Kadoya, Krissansen-Totton et al. 2020).

Among the three routes considered up to this stage, we see that (2.8) is dominant, albeit only by a factor of order unity compared to (2.7). Using the methodology developed in Section 7.6.2.1 that accounts for the abiotic sources and sinks of P, we find that the steady-state concentration (ϕ_P) of dissolved P in the ocean is

$$\phi_P \sim 0.2\,\mu\mathrm{M}\left(\frac{R}{R_\oplus}\right)^{-1.3}. \tag{2.9}$$

A concentration of 1 μM indicates that 10^{-6} moles of the dissolved solute is present in 1 liter of the solvent (water in our case). Apart from these pathways, one other route for synthesizing dissolved phosphate is worth highlighting. After a mixture of basalt and calcium phosphate (to mimic apatite minerals) was heated to temperatures $\gtrsim 1000\,^\circ$C to simulate volcanic magmas, water-soluble polyphosphates were produced from the hydrolysis of P_4O_{10} (phosphorus pentoxide). When the condensates of volcanic gases from Mount Usu were analyzed, the concentration of phosphate was shown to be ~ 1 μM. Hence, in the vicinity of local environments with prominent volcanic activity, phosphate might conceivably approach concentrations that are necessary for the prebiotic synthesis of organic molecules.

A significant fraction of laboratory experiments hitherto undertaken have relied on unusually high concentrations (~ 1 M) of phosphate that were, by and large, probably unrealistic in the Hadean environment. It has been proposed that the minimum concentration of P, in the form of phosphate, should have exceeded 10^{-3} M in order to successfully initiate biochemical reactions leading to the origin of life (Pasek et al. 2017). From the preceding discussion, it appears likely that the global concentrations of phosphate were much lower. However, this outcome does not preclude the existence of myriad microenvironments (i.e., localized patches) endowed with phosphate concentrations that were sufficiently high to facilitate the requisite prebiotic chemical reactions.

Carbonate-rich lakes seemingly represent a promising candidate in this respect because they evince the capability to sequester calcium through the formation of carbonate minerals, consequently suppressing the depletion of phosphate that would otherwise occur via the formation of apatites (Toner & Catling 2020); in these lakes, steady-state phosphate concentrations of ~ 0.1 M may be realizable. Along similar lines, ponds that were rich in cyanide and its derivatives (e.g., urea) in tandem with access to metals could have transformed insoluble phosphate minerals into more soluble and reactive species, thereby surmounting the phosphate problem (Burcar et al. 2019).

2.3 SYNTHESIS OF THE BASIC BUILDING BLOCKS OF LIFE

We will restrict our focus to the synthesis of organic compounds described in Section 2.2 by considering the various energy sources that are capable of playing a significant role in prebiotic chemistry (Chyba & Sagan 1992; Deamer & Weber 2010). We emphasize that this represents a field that is not only incredibly vast but also continuously evolving. Hence, we will restrict ourselves to a few select examples; for instance, we will not discuss prebiotic pathways mediated by the energy derived from volcanism. The interested reader may consult the reviews by McCollom (2013), Luisi (2016), and Meadows et al. (2020) for more details.

2.3.1 Ultraviolet radiation

Among the various energy sources available, UV radiation was predominant on early Earth insofar as the magnitude of energy flux is concerned (Rapf & Vaida 2016). We are primarily interested in the flux of UV photons with wavelengths ranging between 200 nm and 400 nm. The upper bound is set by visible light, whereas the lower bound arises because molecules such as CO_2 and H_2O are efficient absorbers of UV radiation when $\lambda \lesssim 200$ nm. We will, however, restrict ourselves to the range 200–300 nm to maintain consistency with most laboratory experiments. For this range, the UV energy flux (Φ_{UV}) is

$$\Phi_{UV} \sim 2.7 \times 10^7 \,\mathrm{J\,m^{-2}\,yr^{-1}} \left(\frac{M_\star}{M_\odot}\right)^{\nu} \left(\frac{a}{1\,\mathrm{AU}}\right)^{-2}, \tag{2.10}$$

where $M_\star$ is the mass of the host star and a represents the radius of the planet's circular orbit. Note that $\nu \approx 6$ for $M_\star \lesssim M_\odot$ and $\nu \approx 4$ for $M_\star \gtrsim M_\odot$ and the fluxes were expressed in units of yr^{-1} as per the standard convention in this field. It is evident from inspecting this formula that Φ_{UV} declines sharply when $M_\star$ is lowered. In fact, planets orbiting stars with $M_\star \approx 0.1\,M_\odot$ at a fiducial distance of $a \approx 0.02$ AU receive UV fluxes that are about three orders of magnitude lower relative to the Earth.

A wide range of prebiotic compounds have been synthesized using UV radiation. They were formed when gaseous mixtures consisting of hydrocarbons like methane (CH_4), ammonia (NH_3), and hydrogen sulfide (H_2S) were subjected to irradiation by UV light with wavelengths of ~ 200 to 300 nm. The most notable among them include at least fourteen of the

twenty standard amino acids: Ala, Gly, Ser, Glu, Asp, Cys, Asn, Val, Gln, Thr, Leu, Ile, Arg, and Pro (Zaia et al. 2008). The rate of amino acids produced depends on the specific pathway, which in turn depends on many factors, such as the composition of the gases. Although we will return to this matter in Section 2.3.2, to briefly illustrate this point, let us consider the following example.

A photochemical model was used by Tian et al. (2011) to conclude that the deposition rate of HCN on the planetary surface (denoted by $\mathcal{S}_{HCN}$), originally formed in the Earth's atmosphere due to the action of extreme UV radiation (10 to 120 nm), was $\mathcal{S}_{HCN} \sim 5 \times 10^{-6}$ mol m^{-2} yr^{-1}. If we assume that the yield of amino acids synthesized from HCN is on the order of 10 percent, the rate of amino acids produced via this pathway will be $\sim 5 \times 10^{-7}$ mol m^{-2} yr^{-1}. In contrast, if we consider the classic experiment conducted by Carl Sagan and Bishun Khare (1971), in which amino acids were observed after UV irradiation, the production rate ($\mathcal{M}_{UV}$) is about four orders of magnitude higher than the previous value: we end up with $\mathcal{M}_{UV} \sim 2.7 \times 10^{-3}$ mol m^{-2} yr^{-1}.

Another promising area of research that has garnered steady attention, predicated on UV light acting as the energy source, involves the synthesis of key prebiotic compounds wherein formamide is the solvent in lieu of water. We will discuss the relevance, advantages, and production of formamide in Section 2.3.4, owing to which we shall not delve further into the subject here. Suffice to say that, when subjected to UV irradiation and heating at 403 K (130 °C), the synthesis of the canonical nucleobases guanine and adenine, as well as hypoxanthine (regarded as a potential nucleobase candidate), in formamide solutions has been documented.

The last prebiotic pathway involving UV radiation that we shall describe is a relatively recent and genuinely promising breakthrough epitomized by the experiments reported in Patel et al. (2015) and reviewed in Sutherland (2016). As pointed out earlier, sugars, nucleic acids, amino acids, and lipids are vital constituents of protocells, and therefore it would be very advantageous to identify a common pathway by which the precursors of these molecules could be synthesized. Patel et al. demonstrated via laboratory experiments that the aforementioned biomolecular building blocks could be synthesized by starting with simple one-carbon molecules. The reaction network was based on hydrogen cyanide constituting the essential "feedstock" molecule, with H_2S playing the role of reducing agent, as seen in Figure 2.6. Other ingredients utilized in this network included cyanamide

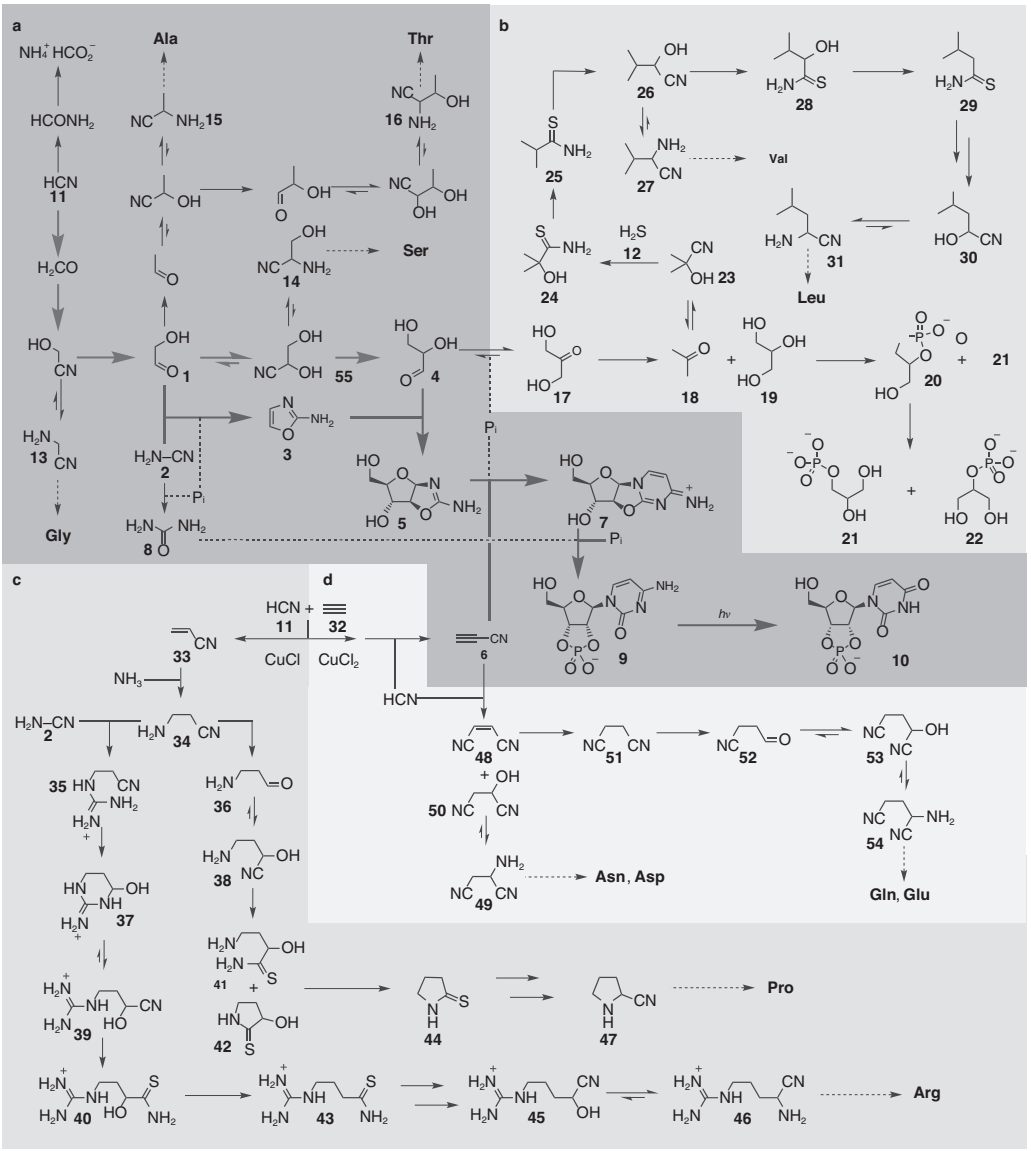


Figure 2.6 Reaction network leading to the precursors of RNA, lipids, and proteins. Panel (*a*): HCN (11) gives rise to the 2- and 3-carbon sugar glycolaldehyde (1) and glyceraldehyde (4). These sugars take part in subsequent reactions to give rise to the RNA nucleotides as depicted by the arrows leading from (1) to (10). The precursors for the amino acids Gly, Ala, Ser, and Thr are also produced. Panel (*b*): Two major products are acetone (18) and glycerol (19). Acetone yields the precursors of amino acids (Val and Leu), whereas glycerol gives rise to glycerol-1-phosphate (21), a lipid precursor. Panel (*c*): Acrylonitrile (33) results in the formation of the precursors of Pro and Arg. Panel (*d*): Cyanoacetylene (6) undergoes subsequent reactions to yield the precursors of Asn, Asp, Gln, and Glu. P_i refers to the inorganic phosphate group. (© Macmillan Publishers Limited. *Source:* Bhavesh H. Patel, Claudia Percivalle, Dougal J. Ritson, Colm D. Duffy, and John D. Sutherland [2015], Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism, *Nature Chemistry* 7[4]: 301–307, fig. 1.)

(CH_2N_2), cyanoacetylene (C_3HN), and inorganic phosphate, with copper (in oxidation states of +1 and +2) serving as the catalyst.

The reliance of the reaction network on HCN and H_2S has led to the coinage of the term "cyanosulfidic protometabolism" (Wu & Sutherland 2019). Two difficulties with this reaction network were addressed in a subsequent publication by Xu et al. (2018). First, the low solubility of H_2S in water implies that achieving sufficiently high concentrations is problematic. This was addressed by replacing H_2S with the sulfite (SO_3^{2-}) anion derived from the dissolution of atmospheric sulfur dioxide (SO_2). Second, the limited availability of copper on Earth is not ideal in light of its role as a putative catalyst. Hence, copper cyanide was replaced by ferrocyanide $[Fe(CN)_6]^{4-}$ because of the fact that the ferrous (Fe^{2+}) ion was relatively abundant in the Hadean-Archean environment. Another notable development worth highlighting is the discovery that iron-sulfur clusters (2Fe-2S and 4Fe-4S)—ancient structures known to mediate vital functions in metabolism—could be synthesized through a UV-mediated pathway (Bonfio et al. 2017).

Despite the undoubted promise and potential significance of this reaction network, certain questions remain unanswered. We begin by noting that the synthesis of these precursors has been posited to occur in a network of streams and lakes, implying that the final products depend on the order in which various compounds are delivered. Hence, when the pathways are order dependent, the absence of the right order could pose difficulties with regard to prebiotic synthesis. More importantly, the concentrations of reactants, such as HCN and phosphate compounds, utilized in these pathways were undoubtedly on the higher side. Further work is necessary to ascertain whether these concentrations were attainable in the geochemical habitats existent on the early Earth. Insofar as achieving high concentrations of HCN is concerned, we will delineate a couple of potential solutions below. Lastly, the fact that none of the lifeforms documented thus far on Earth, either alive or extinct, appears to use the cyanosulfidic metabolic network is regarded as a drawback by some scientists (S. A. Harrison & Lane 2018), but the possibility of the earliest lifeforms using similar biochemical pathways cannot be ruled out altogether.

Our discussion has highlighted why hydrogen cyanide (HCN) was probably an important prebiotic molecule.[10] Hence, it is instructive to

10. This conjecture has a long and rich history (Eschenmoser 2007), with one of its first proponents being Alexander I. Oparin, whom we shall encounter later in the chapter.

determine the steady-state concentration of HCN in water. We will adopt a simple source-sink model of the form

$$\frac{d\mathcal{C}_{HCN}}{dt} = \mathcal{A}_w \mathcal{S}_{HCN} - \frac{\mathcal{C}_{HCN}}{\tau_h}, \tag{2.11}$$

where $\mathcal{C}_{HCN}$ (units of mol) denotes the total concentration of HCN in the water body and τ_h is approximately the half-life timescale for the depletion of HCN. $\mathcal{A}_w$ represents the area of the water body, and the first term on the right-hand side is the source while the second term is the sink; we have introduced $\mathcal{A}_w$ to ensure that the correct units for the source term (mol/yr) are obtained. This simple model assumes that all of the HCN deposited at the surface of the water body is completely soluble.

Solving for the steady-state, we obtain $\mathcal{C}_{HCN} = \tau_h \mathcal{A}_w \mathcal{S}_{HCN}$. However, we are interested in the steady-state concentration $\phi_{HCN} = \mathcal{C}_{HCN}/(\rho_w \mathcal{A}_w \mathcal{H}_w)$, where ρ_w and $\mathcal{H}_w$ are the density and scale height of the water body, respectively. It is evident that the denominator quantifies the volume of the water body. Putting this information together, we find

$$\phi_{HCN} = 10^{-10}\,\mathrm{M} \left(\frac{\tau_h}{100\,\mathrm{yr}}\right) \left(\frac{\mathcal{S}_{HCN}}{10^{-6}\,\mathrm{mol\,m^{-2}\,yr^{-1}}}\right) \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}}\right)^{-1}. \tag{2.12}$$

Determining the depletion timescale of HCN (τ_h) is a difficult endeavor because of its dependence on several factors, such as evaporation and the hydrolysis of HCN. Even if we restrict ourselves solely to the latter phenomenon, τ_h depends on both the temperature (T_w) and pH of the water body under consideration (note that T_w is measured in K). In fact, as per the theoretical model proposed by Stribling and Miller (1987), this timescale depends exponentially on both the variables as follows:

$$\tau_h \approx \exp\left(\frac{20519}{T_w}\right) \left[\exp\left(-2.3\,\mathrm{pH} + 0.04\,T_w - 58.18\right) + \exp\left(-37.8 - \frac{8754}{T_w}\right) \right] \tag{2.13}$$

This formula is relatively accurate provided that the pH lies between six and nine. The hydrolysis rates for a wider range of physical parameters have been cataloged in Miyakawa et al. (2002).

Let us choose the fiducial values of pH = 6.6 and $T_w = 298$ K (25 °C), which yields $\tau_h \approx 1.3 \times 10^3$ yrs. We have noted earlier that $\mathcal{S}_{HCN} \sim 5 \times 10^{-6}$ mol m^{-2} yr^{-1} for the production of HCN via UV radiation. By substituting these quantities into (2.12), we end up with

$$\phi_{HCN} = 6.5 \times 10^{-9}\,\mathrm{M} \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}} \right)^{-1}. \tag{2.14}$$

An important point is that the temperature of the Earth around 4 Ga remains unknown. Several studies have investigated the possibility that much of the Earth, and therefore its oceans, may have been close to freezing. In this scenario, if we adopt $T_w = 273$ K and a pH of 6.6, we obtain $\tau_h \approx 2.7 \times 10^5$ yr. The corresponding value of ϕ_{HCN} is given by

$$\phi_{HCN} = 1.3 \times 10^{-6}\,\mathrm{M} \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}} \right)^{-1}. \tag{2.15}$$

Many laboratory experiments indicate that the critical threshold for HCN to undergo polymerization and form other organic molecules is at least 0.01 M. At the same time, we point out that several other experiments have used even higher concentrations, in the vicinity of 1 M. Thus, it seems reasonable to work with the intermediate value of $\phi_c \sim 0.1$ M for the critical HCN concentration. An inspection of (2.14) and (2.15) reveals that the condition $\phi_{HCN} = \phi_c$ cannot be fulfilled for realistic water bodies with depths of at least a few meters. Before proceeding further, we emphasize that the preceding discussion was based on the assumption that Earth's primordial atmosphere possessed minimal quantities of CH_4. If the concentration of methane was a few parts per million by volume (ppmv) and the abundance of CO_2 was 3 percent, $\mathcal{S}_{HCN}$ could increase by about two orders of magnitude (Tian et al. 2011). In this event, (2.15) will be duly enhanced to yield

$$\phi_{HCN} = 2.6 \times 10^{-4}\,\mathrm{M} \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}} \right)^{-1}, \tag{2.16}$$

provided that all other parameters are held constant. This is particularly interesting in light of the fact that $\mathcal{H}_w \sim 2.6$ m would lead to $\phi_{HCN} = \phi_c$. Thus, UV radiation may have generated concentrations higher than ϕ_c in shallow water bodies, thereby enabling more complex prebiotic compounds

to be synthesized. This optimism should, however, be balanced against the fact that it hinges on a number of favorable criteria being fulfilled.

2.3.2 Lightning (Electrical discharge)

The electrical discharge experiments (to simulate lightning) conducted by Stanley Miller in 1952 with the support of his doctoral advisor, Harold Urey, the 1934 Nobel laureate in chemistry, represent a genuine landmark in the field of prebiotic chemistry (Miller 1953). The Miller-Urey experiments are important for a couple of reasons. First, they sought to mimic the planetary conditions believed to have been prevalent on the young Earth, and they yielded a number of amino acids in this setting. Second, viewed from a historical perspective, Miller's paper appeared in the same year when the first amino-acid sequence of a protein (insulin) was documented and, more importantly, the double-helix structure of DNA was propounded by James Watson and Francis Crick, motivated (to a significant degree) by the X-ray crystallography studies undertaken by Rosalind Franklin, Maurice Wilkins, and their respective collaborators.[11]

The Miller-Urey experiments sparked (to use an unapologetic pun) much excitement since they bolstered the notion that the building blocks of more complex biomolecules could have formed in conditions akin to Hadean Earth. However, a problematic issue with the original Miller-Urey experiments as well as analogous subsequent studies—irrespective of whether they rely on UV radiation, electrical discharge, or shock-wave heating as the energy sources—is that the atmospheric composition of Hadean Earth is not tightly constrained. Miller's original experiments used a combination of H_2, CH_4, NH_3, and H_2O,[12] but subsequent geochemical

---

11. Although Crick, Watson, and Wilkins won the Nobel Prize in 1962 for this achievement, Franklin's substantive contribution remained deplorably overlooked or underappreciated for several decades. Thankfully, the tides of history have begun to reverse their course with regard to this matter. Other noteworthy scientists who paved the way for the discovery of DNA's structure include Oswald Avery, Erwin Chargaff, Raymond Gosling, Alec Stokes, and Herbert Wilson; by no means should this list be viewed as definitive.

12. The pathway by which the amino acids in the Miller-Urey experiments were produced has been described as a variant of the Strecker reaction, named after its discoverer, Adolph Strecker, who synthesized alanine (Ala) in 1850.

modeling suggests that his choice of gaseous mixture was too reducing and not reflective of the conditions on early Earth. While there exists a tentative consensus that the Earth's atmosphere was mildly reducing (i.e., composed mostly of CO_2 and N_2), there are several unknowns as well.

To offer one specific example, some numerical atmospheric models indicate that the escape of H_2 (owing to its low molecular weight) from the Hadean atmosphere was slower by two orders of magnitude with respect to earlier predictions (Tian et al. 2005). If this prediction is correct, it is then feasible for the primordial Earth to have retained a hydrogen-rich atmosphere over a fairly appreciable timescale, thus resembling the classic proposals of the 1950s in certain respects (Urey 1952). Furthermore, when it comes to methane and carbon monoxide (CO), both of which are important volatiles from the standpoint of prebiotic chemistry, comparatively high concentrations could have persisted as long-lived transients. Lastly, gases such as H_2S outgassed due to volcanic activity were probably not globally abundant but were conceivably enriched in select local environments and may have facilitated prebiotic synthesis therein.

With these caveats out of the way, the energy flux available via electrical discharges is

$$\Phi_{UV} \sim 2.9 \times 10^3 \,\mathrm{J\,m^{-2}\,yr^{-1}}. \tag{2.17}$$

Although (2.17) is about four orders of magnitude lower than the UV flux given by (2.10), the efficiency of prebiotic synthesis via lightning is generally much higher, thereby preserving the importance of this energy source. The next aspect that we shall tackle is the production of HCN through this channel. Bearing in mind the fact that the deposition rate of HCN depends on the atmospheric composition, the yield of HCN spans at least two orders of magnitude (Stribling & Miller 1987). In the presence of high concentrations of CH_4, the production rate reaches $\mathcal{S}_{HCN} \sim 2.9 \times 10^{-5}$ mol m^{-2} yr^{-1}. On the other hand, in CO_2 (or CO) atmospheres, the production rate can become $\mathcal{S}_{HCN} \sim 2.9 \times 10^{-7}$ mol m^{-2} yr^{-1}. Even under relatively optimal conditions, using $\tau_h \approx 2.7 \times 10^5$ yr from Section 2.3.1 and the upper bound for $\mathcal{S}_{HCN}$, we end up with

$$\phi_{HCN} = 7.8 \times 10^{-6}\,\mathrm{M} \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}} \right)^{-1}, \tag{2.18}$$

implying that HCN concentrations as high as ϕ_c are probably not achievable if the depth of the water body is at least a few meters.

We will briefly chronicle the organic molecules that have been synthesized by way of spark-discharge experiments. The best known, and the most widely studied, among them are the amino acids. Miller's original experiment in 1952 led to the detection of five amino acids, of which three of them (Gly, Ala, and Asp) are the components of proteins. However, the subsequent analysis of Miller's original samples after his death in 2007 revealed that over twenty amino acids had been produced, and possibly even more. The Miller-Urey experiments were expanded on in many ways (Bada 2013); some of the notable accomplishments include the following:

- It was originally held that a weakly reducing atmosphere (consisting mostly of CO_2 and N_2) gave rise to very low yields of amino acids. The addition of oxidation inhibitors such as Fe^{2+}, which were prevalent on early Earth, has demonstrated that the yields approach those obtained in strongly reducing atmospheres with gases like CH_4 and H_2.
- Miller had carried out several ingenious experiments to mimic how prebiotic synthesis would have taken place in volcanic environments during the Hadean epoch. A jet of steam was injected into the electric spark to mimic steam-rich volcanic eruptions. The reducing agent hydrogen sulfide, encountered in Section 2.3.1, was added to gaseous mixtures motivated by the fact that volcanoes spew out reduced gases (including H_2S).
- All in all, Miller's experiments from the 1950s have yielded thirty amino acids (and probably more) in varying yields. Among them, at least ten of the twenty standard amino acids have been detected.

Apart from amino acids, all of the four RNA nucleobases have been detected at concentrations of a few ppmv when a gaseous mixture of NH_3, CO, and H_2O (at a total pressure of ~ 0.01 atm) was subjected to spark-discharge (Ferus et al. 2017).

2.3.3 Shock waves from impacts

In Section 2.1.1, we discussed the evidence for and against the Late Heavy Bombardment (LHB). Regardless of the actual existence of the LHB, there

are strong theoretical and experimental grounds to believe that the rate of impactors striking the Earth was a few orders of magnitude higher than today. We have already seen that the putative LHB might have led to detrimental effects on habitability, but there are several compelling reasons why the high rate of impact events was potentially beneficial for the origin and evolution of early life.

- Impactors delivered water, bioessential elements such as phosphorus, and a diverse array of organic compounds, which we shall discuss in Section 2.3.6.
- Rocky planets with primordial hydrogen-helium (H / He) atmospheres are widely considered to have a low probability of being habitable. Numerical simulations suggest that such atmospheres can be destroyed and subsequently rebuilt (with gases such as CO_2) via large impactors.
- As per some models, meteorite impacts on Earth may have contributed to the (initially transient) emergence of plate tectonics and the magnetic field, both of which are often listed among the basic criteria for habitability.
- If the Earth was in a frozen state for the most part owing to the Sun's lower luminosity, impacts would have ostensibly supplied enough energy to enable the episodic melting of oceans.

Apart from the above reasons, let us turn our attention to another benefit of impact events that is of particular interest from the standpoint of prebiotic chemistry. After the collision of a large impactor, a significant amount of heat is dispelled and shock waves are produced. The shock heating of atmospheric gases could lead to the production of organic compounds. Alternatively, the vaporized combination of rock or oceans (with dissolved carbonates) and impactor material undergoes post impact recombination reactions to yield prebiotic molecules of interest.

The surface-averaged energy flux (Φ_{IE}) as a result of shock heating due to impact events is given by

$$\Phi_{IE} \sim 2 \times 10^5 \,\mathrm{J\,m^{-2}\,yr^{-1}} \left(\frac{V}{15\,\mathrm{km/s}} \right)^3, \tag{2.19}$$

with V denoting the characteristic speed of the impactors. To obtain this formula, we have used $\Phi_{IE} \propto \dot{M}_{acc} V^2/R^2$ (power per unit area) with R and $\dot{M}_{acc}$ representing the radius and accretion rate of the planet. It was assumed that the latter quantity is approximated by the geometric accretion rate, but a more sophisticated prescription based on the Bondi-Hoyle-Lyttleton accretion rate can be found in Section 7.4.1. Lastly, we utilize $\dot{M}_{acc} \propto R^2 V$ for geometric accretion from a homogeneous medium with identical particles. The constant of proportionality corresponds to Φ_{IE} for the Earth. To simplify our analysis, we have neglected the escape velocity (V_{esc}) of the planet, as otherwise the right-hand side of (2.19) would have to be replaced by $V\left(V^2 + V_{esc}^2\right)$.

The shock heating of atmospheric gases, as remarked earlier, produces precursor molecules such as HCN under certain circumstances. However, the yields of these molecules are very sensitive to the composition of the primordial atmosphere. In the best-case scenario (i.e., in methane-rich atmospheres), about 8.3×10^{-7} moles of HCN are produced per Joule (McKay & Borucki 1997). We will drop the V dependence for now and use (2.19). Thus, we end up with an HCN production rate of $\mathcal{S}_{HCN} \sim 0.17$ mol m^{-2} yr^{-1}. In contrast, for a neutral atmosphere with CO_2 and N_2, the production rate drops by nearly eight orders of magnitude compared to the reducing case (Chyba & Sagan 1992). Since a neutral atmosphere is currently held to be more realistic, we conclude that HCN would not have been produced in appreciable concentrations via this pathway.

On the other hand, the atmosphere of early Earth may have been enriched with CH_4 transiently, as explained later. To err on the side of caution, we will choose a production rate that is an order of magnitude lower than the maximum value listed in the preceding paragraph. We invoke (2.12) and make use of the fiducial estimate of $\tau_h \approx 2.7 \times 10^5$ yr introduced previously. Thus, the steady-state concentration of HCN is given by

$$\phi_{HCN} = 4.5 \times 10^{-2}\,\mathrm{M}\left(\frac{\mathcal{H}_w}{1\,\mathrm{km}}\right)^{-1}, \tag{2.20}$$

from which it is evident that $\phi_{HCN} = \phi_c = 0.1$ M is readily attained for any water body with depths up to 0.5 km. Of course, this analysis hinges on the presence of a methane-rich atmosphere during the Hadean era. Nonetheless, sophisticated models endowed with realistic impactor rates, post-shock

ejecta physics, and atmospheric chemistry yield similar results. In particular, the probability that the LHB resulted in at least one impact event that yielded an HCN concentration of 0.01 M (for a mixing depth of 100 m) is 0.24 to 0.56 (Parkos et al. 2018).

We may approach the issue of HCN concentration through a different channel. Recall that organic material from the impactors "destroyed" after the collision may be subject to post impact chemical reactions and give rise to prebiotic compounds. Let us consider an impactor of radius 300 m; most of the smaller impactors (with radii $\lesssim 100$ m) were probably burnt during their passage through the atmosphere, and larger impactors are relatively infrequent. A carbonaceous body of this size has $\sim 8 \times 10^{11}$ mol of carbon, and laboratory experiments indicate that at least $\sim 8 \times 10^{8}$ mol of HCN is produced from the vaporized carbon (Kurosawa et al. 2013). An accurate determination of the deposition area for HCN is difficult because of the multiple post impact physical processes involved. Instead, for a heuristic estimate, suppose that the produced HCN is distributed over the area of the debris cloud that would be formed by the impactor. The radius of this cloud R_d is computed by using the expression from Hills and Goda (1993):

$$R_d \approx R_i + 3.74 \left(\frac{\rho_a}{\rho_i}\right)^{\frac{1}{2}} H_a, \tag{2.21}$$

where R_i and ρ_i are the radius and density of the impactor, while H_a and ρ_a signify the scale height and density of the atmosphere. For Earth, we specify $H \approx 8 \times 10^3$ m, $\rho_i \approx 3 \times 10^3$ kg/m^3, and $\rho_i \approx 1.2$ kg/m^3. Thus, for an impactor with $R_i = 300$ m, we conclude that $A_d = \pi R_d^2 = 2.6 \times 10^6$ m^2 is the area of the debris cloud. Using this area in conjunction with the amount of HCN produced, we see that the density of HCN deposited is $\Sigma_{HCN} \sim 310$ mol/m^2. The concentration of HCN can be estimated from the above data via

$$\phi_{HCN} = \frac{\Sigma_{HCN} \mathcal{A}_w}{\rho_w \mathcal{A}_w \mathcal{H}_w} = 3.1 \times 10^{-4}\,\mathrm{M} \left(\frac{\mathcal{H}_w}{1\,\mathrm{km}}\right)^{-1}. \tag{2.22}$$

This formula clearly reveals that the likelihood of attaining $\phi_{HCN} = \phi_c$ is quite high if there are shallow water bodies, whose depths are a few meters, within the area spanned by the debris cloud. Of course, it is necessary to

recognize that our estimate was for a *single* impact event and does not represent a global average. The latter is predicted to be very low in magnitude and transient in nature, although local effects are potentially significant at <1 Myr (Todd and Öberg 2020).

The prebiotic synthesis of biomolecular building blocks as a consequence of impact events was, and continues to be, the subject of extensive laboratory investigations. Many of the twenty standard amino acids were produced in these studies: noteworthy examples are Gly, Ala, Ser, Asp, Glu, Val, Leu, Ile, and Pro. Over the past decade, the four nucleobases of RNA have also been detected at concentrations of a few ppmv or higher. In some of the experiments, however, the vapor mixture comprised certain reduced gases whose long-term presence in the primordial atmosphere of our planet remains indeterminate, although these gases could (and probably did) exist for a transient period, as described below. The ramifications of impacts for the origin of life are theorized to have extended far beyond prebiotic chemistry, e.g., triggering the formation of transient habitats for abiogenesis and delivering catalytic compounds like clays (Osinski et al. 2020).

In closing this section, we wish to highlight another crucial phenomenon that may have been actuated by giant impactors—that is, whose sizes were $\sim$100 km or higher. There is some isotopic evidence that these putative impactors contained substantial inventories of highly reduced materials such as metallic iron. The delivery of these compounds could have led to the reduction of water, CO_2, and N_2, thereby yielding transient atmospheres endowed with reducing gases such as NH_3, CH_4, and H_2, as seen in Figure 2.7; over time, these gases are jointly depleted via oxidation, photolysis, and atmospheric escape. The reducing gases described here, as we have seen earlier, readily facilitate the synthesis of prebiotic building blocks and nitrogenous organic compounds (Benner et al. 2019, 2020). Detailed numerical calculations by Zahnle et al. (2020) indicate that transient reducing conditions might have persisted for $\mathcal{O}(10)$ Myr, with the ensuing cumulative production of organics—encompassing amino acids among other things (Takeuchi et al. 2020)—covering Earth's surface up to an average depth of 0.5 km.

2.3.4 Radioactivity

The Earth contains a number of long-lived radioactive elements such as uranium-238 (^{238}U) that produce heat upon decaying. The total amount of heat produced per unit time is conventionally modeled as being proportional

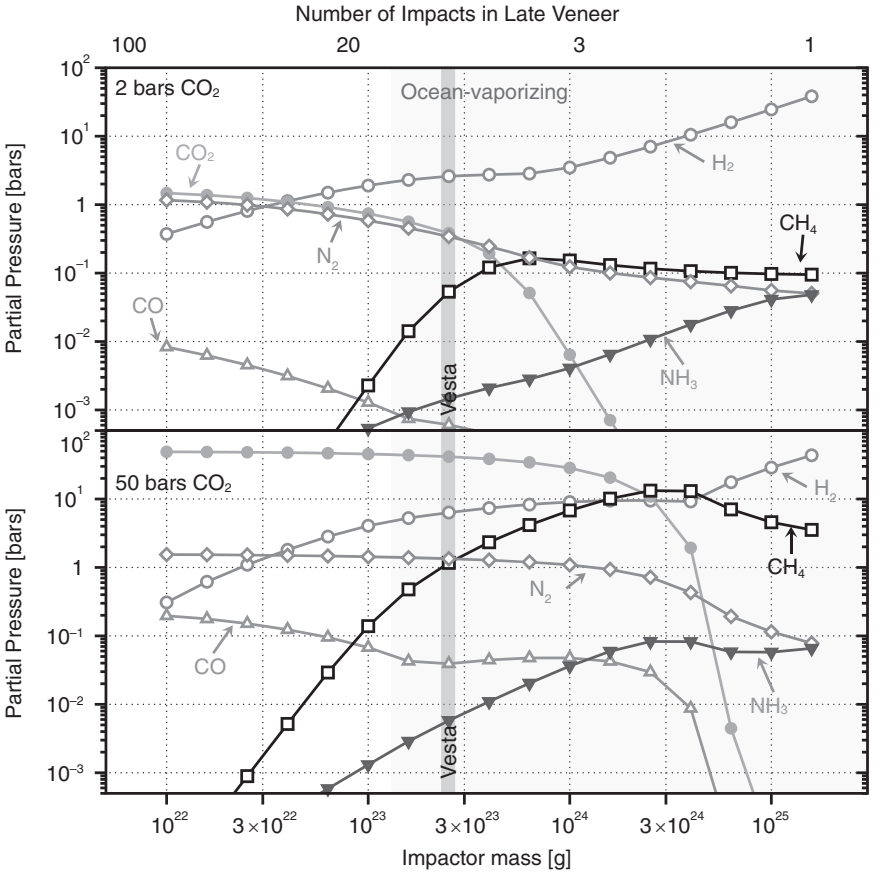


Figure 2.7 Transient atmospheres generated post impact as a function of the impactor mass for two choices of pre-impact atmospheric CO_2 inventories; both cases are assumed to possess 1 bar of N_2 and 1.85 oceans (5 km) of water. The shaded bar embodies a Vesta-sized impact (525 km in diameter), whereas the region broadly to the right of the bar demarcates impactors capable of vaporizing the oceans. At the top, a rough estimate of the number of impacts of a given size in the Hadean is depicted. The production of reducing gases occurs via the conversion of Fe delivered by the impactor into FeO. (CC-BY 4.0. *Source:* Kevin Zahnle, Roxana Lupu, David Catling, and Nick Wogan [2020], Creation and evolution of impact-generated reduced atmospheres of early Earth, *Planetary Science Journal* 1[1]: 11, fig. 3.)

to the mass of the planet. Hence, the energy flux available via radioactivity (Φ_R) on the surface of early Earth has been estimated to be

$$\Phi_R \sim 1.2 \times 10^5 \,\mathrm{J\,m^{-2}\,yr^{-1}} \left(\frac{R}{R_\oplus}\right)^{\beta-2}, \tag{2.23}$$

where $\beta \approx 3.3$ for $R \lesssim R_\oplus$ and $\beta \approx 3.7$ for $R \gtrsim R_\oplus$. The overall radiogenic heat flux for the Earth is about 50 percent of the total internal heat flux today, because the remaining half is contributed by primordial heat leftover from the formation of the planet via gravitational collapse. Although the energy flux from radioactivity is comparable to, or even exceeds, the other energy sources (barring UV radiation), comparatively little research has been undertaken in this field.

One of the underrated aspects about radioactivity is that the distribution of radioactive elements is inexorably subject to spatial heterogeneity. Hence, localized environments endowed with high concentrations of radioactive elements, occurring in the form of minerals, might serve as plausible sites of abiogenesis. We shall not elaborate on this aspect here, because we shall tackle it in Section 2.7.3. For now, though, it suffices to remark that nature ingeniously "invented" a nuclear fission reactor long before *Homo sapiens* on Earth. The natural fission reactor at the Oklo deposit in Gabon, discovered in 1972, was presumably functional by ~ 1.8 Ga (Meshik 2005). Apart from natural reactors, radioactive beaches may have existed at the time of abiogenesis on Earth by virtue of the accumulation of radioactive minerals in shallow sediments.

Due to the localized action of intensive prebiotic synthesis by radioactive minerals, it has been shown that formamide ($HCONH_2$) may be produced near placer deposits (radioactive beaches and rivers) at rates of $\mathcal{S}_F \sim \mathcal{O}\left(10^{-7}\right)$ mol m^{-2} yr^{-1} (Adam et al. 2018). If we consider natural fission zones akin to the Oklo reactor instead, the rates are amplified by many orders of magnitude and attain a maximum of $\mathcal{S}_F \sim 1$ mol m^{-2} yr^{-1}. Unlike other prebiotic environments previously proposed for synthesizing formamide, this setting is relatively amenable to attaining high concentrations because the production of formamide occurs at a much higher rate than its degradation. Why is formamide so important though?

The core reason behind the appeal of formamide stems from its dual capacity to function as the solvent in which prebiotic chemistry takes place and as the substrate for synthesizing more complex compounds. We shall delineate some of the chief advantages attributed to formamide below, but we invite the reader to consult Saladino et al. (2012, 2019) for lucid expositions of this issue.

- One of the most prominent difficulties encountered when water functions as the solvent is that many of the essential biochemical

bonds undergo rapid degradation. The RNA phosphodiester bond that links two ribonucleotides has a short half-life of less than 1 yr in water at neutral pH and 303 K (30 °C), while the peptide bond joining amino acids displays a half-life of ~ 100 yr at pH 7 and 298 K (25 °C). This difficulty is bypassed by formamide solutions that are known to facilitate the formation of biopolymers such as nucleic acids and peptides owing to their electrophilic (electron acceptance) reactivity.

- All of the five canonical nucleobases for RNA and DNA have been synthesized by starting from formamide under a wide range of conditions using different candidates as catalysts. The formation of nucleosides and the subsequent phosphorylation to yield nucleotides (the monomers of nucleic acids) was also documented.
- Apart from the precursors of nucleic acids, formamide has also led to the formation of amino acids and carboxylic acids (lipid precursors) through different pathways. Collectively, it is evident that a wide range of biologically relevant molecules can be produced by starting from formamide.
- Formamide also remains a liquid over a wider range as compared to water under standard conditions. Its melting point is around 275 K (2 °C), almost equal to that of water, but its boiling point is 483 K (210 °C). Thus, compared to certain solvents that have been studied as potential replacements for water, formamide is less volatile.

Yet, it must be recognized that there are notable downsides to the hypothesis that formamide comprised a viable prebiotic solvent. One of the chief reasons concerns its availability—to wit, access to adequately high concentrations across a large number of "sites"—but it seems plausible that radioactivity could mitigate this deficiency to an extent. A point of direct significance in this context is that formamide is often synthesized from the hydrolysis of HCN. Thus, our preceding discussion concerning the availability of HCN acquires another degree of relevance.

Recent experiments have subjected HCN in conjunction with several other basic feedstock molecules like phosphate, sodium chloride (NaCl), and ammonium chloride (NH_4Cl) to repeated radiolysis by gamma rays and drying (Yi et al. 2020). This continuous reaction network gave rise

to a number of crucial RNA building blocks such as glycolaldehyde and cyanamide (both of which were mentioned in Section 2.3.1) without the necessity of having to regulate the timing or order of the addition of reagents. Other empirical results have established the emergence of amino acids, carboxylic acids, and ribonucleotide precursors via radiolysis as well as the unique topological properties of these prebiotic pathways.

2.3.5 Solar energetic particles and cosmic rays

Stars, including our Sun, sporadically give rise to eruptions of high-energy radiation that have been christened stellar (or solar) flares. The most powerful flares documented on the Sun possess energies on the order of 10^{25} J, but other stars have been shown to produce gigantic flares (superflares) that are as much as five orders of magnitude greater. The most powerful flares are accompanied by the release of large masses of plasma and magnetic fields called Coronal Mass Ejections (CMEs). High-energy particles arise directly or indirectly (via CMEs) as a result of flares. They are known as solar/stellar energetic particles (SEPs) and reach maximum energies of ~ 10 GeV (giga-electron-volt), although the majority evince lower energies of keV to MeV.

Another source of high-energy particles in the Earth's atmosphere are cosmic rays (CRs). CRs typically have much higher energies: their energy range is ~ 1 GeV to $> 10^{15}$ eV; particles with energies equal to the upper bound are naturally very rare. CRs are commonly produced during the death of high-mass stars (supernovae), but several other sources and mechanisms also exist. Both SEPs and CRs are primarily comprised of high-energy protons (and electrons). Let us begin by considering the energy flux contributed by CRs, which equals

$$\Phi_{\mathrm{CR}} \sim 4.6 \times 10^2 \,\mathrm{J\,m^{-2}\,yr^{-1}}, \tag{2.24}$$

with Φ_{CR} denoting the energy flux of CRs received at the *surface* of Earth. Now we turn our attention to SEPs. Estimating the energy flux is a much more complex endeavor since two different factors must be taken into account. First, the Sun is believed to have been active (i.e., producing large flares very frequently) during the Hadean era. Hence, we must not only determine the number of large flares that occurred but also assess the number of SEPs released per flaring event. Second, not all of the SEPs make their way to the surface because of their lower energies (unlike CRs). Thus,

numerical simulations are necessary to determine the energy deposited by SEPs at the surface during a single flare. These issues were studied in depth by Lingam et al. (2018), wherein the SEP energy flux received at the planet's surface (Φ_{SEP}) was given by

$$\Phi_{\mathrm{CR}} \sim 50\,\mathrm{J\,m^{-2}\,yr^{-1}} \left(\frac{a}{1\,\mathrm{AU}}\right)^{-2}, \tag{2.25}$$

which is smaller than the CR flux by about an order of magnitude for solar-type stars. Yet, there is a crucial distinction that lies hidden: while the supply of CRs is continuous, that of SEPs is discontinuous since it occurs only whenever a flare erupts. Hence, it is more instructive to view (2.25) as a time-averaged value. During the "on" phase, when the SEPs are hitting the planet, the energy flux will be around an order of magnitude higher than (2.25). Another point worth mentioning is that a planet with the same effective temperature as the Earth orbiting a star roughly 10 percent of the mass of Sun will probably receive a SEP flux that is ~ 1000 times higher with respect to early Earth. In our subsequent discussion, we will only deal with SEPs since the corresponding results for CRs would be approximately one order of magnitude higher for our planet during the Hadean period.

A comparison of the CR and SEP energy fluxes against the other sources clearly reveals that they are much lower, sometimes even by several orders of magnitude. This basic observation immediately brings up the question: Are these energy sources of any importance? Despite the uncertainties (e.g., atmospheric composition) involved, there are grounds for believing that the answer is in the affirmative. If we compare a UV photon of wavelength 300 nm with a GeV proton, the energy of the latter dominates over the former by eight orders of magnitude. The energy carried by a particle plays a significant role in its ability to cleave chemical bonds and assist in the formation of new ones. The number of essential prebiotic molecules produced per Joule of energy has been shown to be among the highest in laboratory experiments for energetic particles of a few MeV.

Next, let us turn our attention to the range of biomolecular building blocks that emerge when high-energy particles serve as the energy source.[13]

---

13. This subject has a long history, going back to the 1950s, when formic acid was produced from carbon dioxide after the latter was bombarded by high-energy particles (Garrison et al. 1951).

The irradiation of gaseous mixtures consisting of N_2, CO_2, CO, and H_2O has demonstrated that four of the five canonical nucleobases are formed, with thymine being the exception. Furthermore, the proteinogenic amino acids Gly, Ala, Asp, Ser, Pro, and Thr were also synthesized under the same conditions. By invoking the restrictive assumption that the efficiency of prebiotic synthesis does not depend on the energy of individual particles, Lingam et al. (2018) concluded that the maximum production rates on the Earth were $\sim 10^8$ mol/yr for amino acids and $\sim 10^5$ mol/yr for nucleobases through this pathway. It was proposed that the yield of amino acids generated through this avenue was comparable to more widely studied routes entailing UV radiation or electrical discharges.

All five of the canonical nucleobases and a host of other prebiotic compounds have also been documented via another channel involving high-energy particles—namely, when powdered meteorites in liquid formamide were subjected to irradiation by energetic protons. Recent numerical simulations also reveal that sizable quantities of HCN are produced at low altitudes for neutral (or mildly reducing) atmospheric compositions. The use of Henry's Law, which states that the concentration of dissolved gas is proportional to its partial pressure, implies that HCN concentration might approach $\sim 10^{-3}$ M. Despite this value being smaller than ϕ_c by two orders of magnitude, this pathway ought not be altogether discounted for HCN synthesis, because the concentration thus obtained is not extremely low.

2.3.6 Exogenous delivery of organic compounds

There are three different routes by which the delivery of organic compounds from extraterrestrial objects take place: (1) interplanetary dust particles (IDPs), (2) cometary and asteroid impacts, and (3) meteorites. In fact, the distinction between (2) and (3) is mostly one of size. Of the three channels, the accretion of IDPs yields the highest quantity of intact organics but is also less understood because of the small size of the IDPs (meteoroids).

The mass flux of organic carbon via IDPs ($\mathcal{M}_{\rm DP}$) is estimated by using the prescription elaborated on in Section 7.4.1, which the reader should consult for the derivation. The mass flux is expressible as

$$\mathcal{M}_{\rm DP} \sim 2 \times 10^{-7}\,{\rm kg\,m^{-2}\,yr^{-1}} \left(\frac{\sigma}{26\,{\rm km/s}}\right)\left(\frac{\rho_d}{2 \times 10^{-22}\,{\rm kg/m^3}}\right), \tag{2.26}$$

when $R > R_B = GM/\sigma^2$ with R_B denoting the Bondi radius. In this formula, M is the mass of the planet; $\sigma = \sqrt{V_d^2 + c_s^2}$, where V_d is the relative velocity between the dust particles and the planet and c_s is the sound speed; and ρ_d is the density of dust particles. All quantities have been normalized in terms of the values observed in the vicinity of Earth, while the prefactor of $2 \times 10^{-7}\,\mathrm{kg\,m^{-2}\,yr^{-1}}$ corresponds to the delivery rate of intact organics via IDPs on Earth at ~ 4 Ga. In contrast, in the regime where the Bondi radius exceeds the planetary radius (i.e., $R_B > R$), the geometric accretion rate must be replaced by the Bondi-Hoyle-Lyttleton accretion rate. This leads us to

$$\mathcal{M}_{\mathrm{DP}} \sim 2 \times 10^{-7}\,\mathrm{kg\,m^{-2}\,yr^{-1}} \left(\frac{R}{R_\oplus}\right)^{2\beta-2} \times \left(\frac{\sigma}{26\,\mathrm{km/s}}\right)^{-3} \left(\frac{\rho_d}{2 \times 10^{-22}\,\mathrm{kg/m^3}}\right), \tag{2.27}$$

where we have introduced the definition of β immediately below (2.23).

Next, let us turn our attention to asteroid and cometary impacts. Chyba and Sagan (1992) impose a cutoff of $\mathcal{O}(100)$ m for the radii of impactors in their discussion of asteroid and cometary impacts. The delivery rate of intact organics varies by a few orders of magnitude between 4.5 Ga and 3.0 Ga, but we shall adopt the value at 4.0 Ga, as the possibility of life originating in the Hadean eon cannot be ruled out. Thus, the mass flux via cometary impacts ($\mathcal{M}_C$) is

$$\mathcal{M}_C \sim 4 \times 10^{-10}\,\mathrm{kg\,m^{-2}\,yr^{-1}}, \tag{2.28}$$

where we have used the value predicted for the Earth at 4 Ga. We have not provided any dependence on planetary or stellar parameters, since the mass flux of impactors is likely to depend on the dynamics of a particular planetary system. The last channel concerns the delivery of organic carbon via meteorites and does not differ greatly from the second route apart from the size of the impactors. The corresponding mass flux via meteorites ($\mathcal{M}_{ME}$) is

$$\mathcal{M}_{ME} \sim 4 \times 10^{-12}\,\mathrm{kg\,m^{-2}\,yr^{-1}}, \tag{2.29}$$

with the mass flux being calculated at ~ 4 Ga.

Over the past few years, missions such as *Stardust* and *Rosetta,* which studied the comets Wild 2 and 67P/Churyumov-Gerasimenko (67P),

respectively, have shed new light on the organic molecules within comets. One of the most notable discoveries was that comets harbor glycine, the simplest proteinogenic amino acid, as well as the precursors of this important prebiotic molecule. Another important breakthrough was the discovery of phosphine (PH_3) in 67P, especially since its abundance was about three times higher than the typical Solar system value. Lastly, SO_2, H_2S, and HCN were detected in 67P, and the significance of these molecules stems from their centrality to the cyanosulfidic metabolism pathways discussed in Section 2.3.1. These discoveries lend credence to the classic hypothesis that comets may have delivered vital prebiotic compounds to Earth, either as intact material or subjected to post impact recombination (Oró 1961).

Most of our understanding of organic compounds that existed in the early Solar system has been derived from the study of carbonaceous chondrites (CCs)—a class of primitive meteorites that contain ~ 0.3 to 3 percent (by weight) C, most of which is manifested in a dazzling array of organic compounds. Among the CCs, the best known and most widely studied is the Murchison meteorite (with a mass > 100 kg). The analysis of the Murchison meteorite has led to the identification of $> 10^4$ chemical compounds (with millions of distinct structures), including several proteinogenic amino acids, three of the four canonical nucleobases of RNA (with cytosine being the exception), and carboxylic acids (precursors of lipids). Collectively, the total inventory of organic molecules relevant for prebiotic chemistry derived from CCs includes ≥ 75 amino acids (of which at least thirteen are proteinogenic), canonical and noncanonical nucleobases, (di)carboxylic acids, sugar acids and sugar alcohols (both of which are related to sugars), ribose, and other bioessential sugars (Furukawa et al. 2019). Apart from these organic compounds, a wide range of molecules with nitrogen (e.g., NH_3) and sulfur have been detected. Interestingly, the suite of molecules with phosphorus appears to be much more limited. A recent overview of this subject can be found in Pizzarello and Shock (2017).

2.3.7 Hydrothermal conditions

By *hydrothermal conditions,* we refer to alkaline (pH of 9 to 11) aqueous environments endowed with high temperatures ($\lesssim 373$ K) and pressures (~ 300 atm) that mimic alkaline "low-temperature" hydrothermal vents discovered at the bottom of the oceans. We shall revisit this environment in Section 2.7.

A wide array of laboratory experiments have sought to simulate the conditions of hydrothermal vents, although a substantial fraction of them did not explicitly incorporate the effects of high pressure. Most of the experiments heated a mixture of formaldehyde (HCHO) and cyanide (CN^-) compounds in conjunction with other carbon (e.g., CO_2) and nitrogen sources (Huber & Wächtershäuser 2006); examples of the latter are ammonium (NH_4^+) compounds or ammonia (NH_3). The amino acids detected include at least seventeen of the twenty standard amino acids, with Gln, Trp, and Tyr being the possible exceptions. Barge et al. (2019) demonstrated the synthesis of Ala starting from a simpler set of feedstock molecules and under plausible hydrothermal vent conditions. Sugars were obtained from formaldehyde after heating by means of the classic formose reaction, which was discovered in 1861 by Aleksandr Butlerov.

When aqueous solutions containing cyanides and / or ammonium are heated, it has been found that the purine nucleobases A and G are produced, albeit at relatively low yields. The synthesis of the pyrimidine nucleobases C, T, and U is more challenging but has been achieved by heating compounds such as malic acid ($C_4H_6O_5$), cyanoacetylene, and potassium cyanate (KCNO). Alternatively, the pyrimidine bases have been produced by heating formamide solutions with the addition of catalysts like titanium dioxide (TiO_2). Nucleobases can also be synthesized by means of Fischer-Tropsch-type (FTT) reactions when NH_3 is included in the mix. The Fischer-Tropsch process was discovered in the 1920s by Franz Fischer and Hans Tropsch and entails the conversion of CO into organic compounds by sequential reduction (typically involving H_2) and polymerization at high temperatures ($\sim$ 373 to 573 K) in the presence of catalysts.

FTT reactions have also been used to facilitate the formation of fatty acids and alcohols (lipid precursors) with up to twenty-two carbons in the laboratory; by the same token, FTT synthesis has been employed to synthesize vesicles from mixed amphiphiles under hydrothermal conditions resembling those found at the seafloor (Jordan et al. 2019). Apart from the experiments described here, detailed thermodynamical calculations have been undertaken to assess the likelihood of synthesizing amino acids, nucleobases, sugars, and fatty acids under hydrothermal conditions. On theoretical grounds, it has been demonstrated that the production of these compounds, for the most part, is thermodynamically favorable under such circumstances (Amend et al. 2013)—that is, their Gibbs energy of

formation is negative ($\Delta G < 0$), implying that the reactions are exergonic (energy yielding) in nature.

Although these experiments in conjunction with theoretical calculations establish that biomolecular precursors could be synthesized under hydrothermal conditions, several issues remain unresolved. For starters, most of the experiments rely on the availability of highly reactive compounds at concentrations that appear to be implausible in hydrothermal environments at first glimpse; the reality is more complex, as explained in Section 2.4.4. When the same experiments were repeated with either less reactive compounds or lower concentrations, the yields of these molecules were minimal (perhaps even zero). Another point that must be noted in this regard is that many of the compounds produced have incredibly short half-lives at high temperatures in aqueous solutions. One of the most striking examples in this context is the decomposition of cytosine (via hydrolysis) with a half-life of only nineteen days at 373 K; the corresponding values for the other RNA nucleobases are also quite limited at the same conditions ($\sim$ 1 to 12 yr).

2.4 THE POLYMERIZATION OF MONOMERS

We will assume that the biomonomers of interest (i.e., nucleotides, amino acids, and lipids) have already been synthesized through the appropriate pathways. This postulate allows us to direct our attention toward the myriad avenues through which polymerization can take place. The astute reader will notice that we have used *nucleotides* in place of *nucleobases,* although we have not yet elucidated how nucleotides are assembled from their components—namely, nucleobases, sugars (ribose/deoxyribose), and the phosphate group.

Two prominent questions immediately arise. Is the spontaneous formation of biopolymers in aqueous solutions at high concentrations feasible? And if not, what is the reason? The first question has a negative answer, and the explanation for the second is based on the datum that polymerization is thermodynamically disfavored—to wit, the reaction is endergonic ($\Delta G > 0$)—under "standard" conditions of neutral pH, pressure of 1 atm, and temperature of 298 K.[14] For instance, the dimerization of Gly in water,

14. We caution, however, that contra textbook treatments, the standard Gibbs energy of reaction (ΔG^0) is *not* always evaluated at the aforementioned parameters (Amend & LaRowe 2019).

$$\mathrm{Gly} + \mathrm{Gly} \rightarrow \mathrm{Gly} - \mathrm{Gly} + \mathrm{H_2O}, \tag{2.30}$$

has $\Delta G^0 \approx 15$ kJ/mol. Similarly, for glycine, the assembly of an $(n+1)$-mer from an n-mer and Gly typically requires $\Delta G^0 \approx 9.6$ to 10.5 kJ/mol. Note that the equilibrium constant K_{eq} can be computed from ΔG^0 and vice versa by invoking

$$\Delta G^0 = -RT \ln K_{eq}, \tag{2.31}$$

where $R = 8.314$ J mol^{-1} K^{-1} is the molar gas constant, and T denotes the temperature. Thus, if we consider the dimerization of Gly in water at 298 K, we find $K_{eq} \approx 2.3 \times 10^{-3}$ M. In other words, even if we start with a high concentration of Gly (~ 1 M), the concentration of Gly-Gly is only $\sim 2.3 \times 10^{-3}$ M. From the above data, selecting $\Delta G^0 \approx 10$ kJ/mol leads us to $K_{eq} \sim 1.8 \times 10^{-2}$ M for the production of an $(n+1)$-mer from an n-mer. Thus, it becomes evident that forming long polymers in water is a highly difficult endeavor. For instance, if we wish to synthesize an 18-mer of Gly starting with 1 M concentration of this monomer, we find that the resultant equilibrium concentration of the 18-mer would be $\sim 2.3 \times 10^{-3} \left(1.8 \times 10^{-2}\right)^{16} = 2.8 \times 10^{-31}$ M, which is infinitesimally small.

The same set of arguments apply to the synthesis of oligonucleotides (short polymers of nucleotides), but the above issues are further exacerbated due to the higher free energies. For instance, it has been estimated the formation of the DNA phosphodiester bond (that links two nucleotides) in water necessitates a free energy of $\Delta G^0 \approx 22$ kJ/mol, whereas the corresponding value for the RNA phosphodiester bond is $\Delta G^0 \approx 14$ kJ/mol. Apart from thermodynamics, we must also consider kinetics, but the latter issue is arguably not as pressing as the former insofar as peptide bonds are concerned. This is because peptide bond formation (and decomposition via hydrolysis) has a characteristic half-life of $\mathcal{O}(100)$ yr at neutral pH at 298 K, which has been argued to be smaller than typical geological timescales and therefore relatively fast (Lambert 2008), but not so much as to pose stability issues. On the other hand, the hydrolysis of RNA has a much shorter half-life ($\lesssim 5$ yr); we shall revisit this in Section 2.5.

The pathways to achieving polymerization are numerous, owing to which we shall focus only on a select few examples, not all of which are mutually exclusive.

2.4.1 Wet-dry cycles

To understand why wetting-drying cycles are favorable on thermodynamic grounds, we will sketch the arguments delineated by Ross and Deamer (2016, 2019). Before we proceed, we emphasize that our primary focus is on the drying phase, which leads to the concentration of monomers and facilitates oligomerization (formation of short polymers). However, the wetting phase is also very important since it enables the redistribution of organic molecules, confinement of polymers within vesicles, (un)desired cleavage of chemical bonds, and, in geological environments, delivery of new organics.

We will focus on the formation of oligomers via the generic chemical reaction

$$\mathrm{N}_m + \mathrm{N}_n \leftrightarrow \mathrm{N}_{m+n} + \mathrm{H_2O}, \tag{2.32}$$

where the forward reaction indicates the production of the $(m+n)$-mer from an m-mer and n-mer. The Gibbs free energy of the reaction is given by

$$\Delta G = \Delta G^0 + RT \ln \left(\frac{\mathrm{a}_{m+n}}{\mathrm{a}_m \mathrm{a}_n} a_W \right), \tag{2.33}$$

where a denotes the thermodynamic activity of the species, and the subscript identifies the species (W is water). The expression inside the brackets on the right-hand side of (2.33), known as the reaction quotient, approaches unity for dilute solutions. The activity a_s is defined as

$$a_s = \exp \left(\frac{\mu_s - \mu_s^0}{RT} \right), \tag{2.34}$$

where μ_s is the chemical potential of the species and μ_s^0 is the chemical potential at some specified standard conditions. The activity is a measure of the "effective" concentration and is related to the actual concentration through a variable known as the activity coefficient.

The basic point here is that the alternation of wetting and drying cycles leads to the evaporation of water and the concentration of reactants, thereby increasing their concentrations. There is a corresponding decrease in the activity of macromolecules on account of crowding, and this effect results

in the second term on the right-hand side of (2.33) becoming negative,[15] counteracting the positive first term. Another source responsible for lowering the reaction quotient is a_W, whose value is given by the Kelvin equation

$$\ln a_W = -\frac{2\gamma_w V_w}{r_c RT}, \tag{2.35}$$

where γ_w and V_w are the surface tension and molar volume of water, respectively, whereas r_c quantifies the representative size of water cavities in the reaction mixture. As a result of the evaporative phases, the molecules are increasingly crowded and the r_c is lowered. In turn, this leads to a considerable decrease in a_W owing to the exponential dependence. The collective effects of evaporation have been estimated to lower the second term on the right-hand side of (2.33) by a factor of ~ 3 to $4\,\Delta G^0$, thereby rendering $\Delta G < 0$ and facilitating oligomerization.

An important point to note here, however, is that the process of forming oligomers cannot continue indefinitely, even if the requisite raw materials are abundant. The reason is believed to stem from the expectation that the selfsame crowding effects wrought by evaporation, which are beneficial from a thermodynamic standpoint, also bring about reduction of the reaction rate. As a result, when sufficiently high concentrations of oligomers are attained, reaction kinetics becomes the dominant limiting factor, and their further production is anticipated to cease at some point. We have commented on wet-dry cycles and crowding in the narrow backdrop of polymerization, but these facets are theorized to play cardinal roles in the origin of life (Spitzer et al. 2015). The crowding of macromolecules, to take the latter, modulates catalysis, self-assembly, diffusive transport, and molecular recognition, to name a few.

Even though our discussion hitherto was theoretical, we observe that an array of experiments conducted since the 1970s have demonstrated that peptides, oligonucleotides, and vesicles (made of lipids) are formed when their corresponding monomers are subjected to wetting and drying cycles (Deamer 2019). In addition, empirical evidence indicates that the synthesis of biomolecular building blocks like nucleobases (e.g., cytosine from urea and cyanoacetaldehyde) can take place in wet-dry conditions.

---

15. The situation is somewhat analogous to how children are capable of getting through nooks and crannies that adults cannot on account of their larger volume.

2.4.2 Ice and eutectic freezing

Another scenario, closely connected to wet-dry cycles, concerns the use of freeze-thaw cycles. This mechanism is particularly relevant in the context of our discussion of icy worlds with subsurface oceans in Section 7.4.2. The starting premise of this hypothesis is that Hadean Earth was close to freezing. As we have seen in Section 2.3.1, one of the advantages with positing cold temperatures is that valuable prebiotic molecules such as HCN are longer-lived at lower temperatures since they are less vulnerable to hydrolysis.

However, readers may be inclined to expostulate that lower temperatures should concomitantly lower the reaction rates for oligomerization, owing to the exponential dependence of the reaction rate on the temperature, consequently stymieing the likelihood of this process. In this respect, they would undoubtedly be correct, but it is equally essential to realize that the reaction rate scales with the concentration of the reactants. This trend is easy to understand intuitively because the chemical species undergo more "collisions" per unit time at higher concentrations and thereby result in faster reaction rates. The importance of freezing stems from the fact that it facilitates the concentration of molecules for reasons adumbrated below.

The basic reason stems from a basic property of two-component liquid mixtures, of which one of them is water (the solvent) and the other is the solute (the organic molecules of interest). When the temperature drops below the freezing point of water, ice crystals begin to form. The solute undergoes subsequent concentration in the fluid cavities located amid the ice matrix. If the temperature is continuously decreased, ultimately both components are subject to crystallization. We pointed out during our analysis of (2.33) that the activities are diminished when the reactants are more concentrated. The same principle applies here, albeit for different reasons, thus lowering the Gibbs free energy and facilitating polymerization. Another point to note is that the activity of water also diminishes with temperature, if all other factors are held fixed, as can be seen by inspecting (2.35). This approach of concentrating reactant solutions, and setting the stage for prebiotic synthesis, often goes by the name of eutectic freezing.

A discussion of the physical and chemical basis of this mechanism can be found in Menor-Salván and Marín-Yaseli (2012) and Feller (2017), along with a recent summary of the attendant experimental studies. To begin with, we note that all four of the RNA nucleobases have been synthesized by starting with appropriate feedstock molecules (such as HCN) and subjecting

them to freezing. In 2001, the formation of oligomers up to eleven bases in length was demonstrated within a few days, when the mixtures were frozen at 255 K in the presence of soluble catalytic ions. In a similar vein, the oligomerization of amino acids using suitable catalysts engendered peptides at high yields. Apart from these milestones, ice (and freeze-thaw cycles) has been show to provide an unusually clement, perhaps even essential, medium for stabilizing and driving the replication of RNA.

Despite the advantages associated with ice as a medium, a note of warning must be struck here in connection with eutectic freezing. The mechanism is predicated on the assumption that amino acids and other prebiotic molecules are incompatible with ice—that is, they are not incorporated into the ice during freezing. This fundamental premise has been thrown into question by a recent experimental study, which concluded that amino acids are apparently *compatible* with ice (Hao et al. 2018). However, while this hampers the prospects for eutectic freezing, it simultaneously increases the likelihood of exogenous delivery (see Section 2.3.6) by a few orders of magnitude, as the relevant organic compounds are better protected within the ice matrix.

2.4.3 Minerals

The role of minerals is not merely restricted to their capacity to facilitate or speed up chemical reactions. Among other properties, they may enhance the thermal stability of biomolecular building blocks, enable the selective binding of prebiotic compounds to sites on their surfaces, catalyze the preferential synthesis and accumulation of longer biopolymers (e.g., RNA), promote the restructuring and organization of molecules, and assist in the emergence of biological homochirality, the last of which is encountered toward the end of this chapter; the reader is directed to Cleaves et al. (2012) and Mizuuchi et al. (2019) for surveys of this field.

One of the avenues by which minerals may play a positive role in polymerization is via adsorption, i.e., the process by which molecules (in gaseous, liquid, or solute form) adhere to the surface of a solid. Let us turn our attention to (2.32), but assume that the polymerization takes place in the adsorbed state. It is conceivable that the Gibbs free energy of this reaction is negative, even though the corresponding value for the same reaction in an aqueous medium is positive. This can mathematically occur when the free

energy associated with adsorption is sufficiently negative; more precisely, its magnitude should exceed 20 kJ/mol (Lambert 2008). In some cases, the net Gibbs free energy is indeed (slightly) negative, but the results also depend on the mineral chosen. Among the common oxide minerals, Gly polymerization is predicted to occur with highest efficiency when rutile (which consists mostly of titanium dioxide) is used.

The majority of experiments that entail the use of minerals, and other pathways for that matter, have relied on coupling the endergonic reaction of oligomer formation to another, thermodynamically favored (exergonic) reaction, akin to how cells function in today's biology. This coupling is typically accomplished by means of compounds called *activating agents,* but most of the ones employed in laboratory experiments were probably not widespread on early Earth nor continuously supplied (cf. Liu et al. 2020). Hence, in identifying the most realistic prebiotic pathways, care must be taken to ensure that the activating agents were compatible with Hadean geochemistry. The numerous methods by which inert amino acids and nucleotides have been converted into their activated versions will not be tackled further at this juncture.

One mineral that merits special attention is montmorillonite—a phyllosilicate (with silicon-to-oxygen ratio of 2:5) clay. Montmorillonite, abundant on Earth today, is formed when volcanic ash undergoes weathering. When we account for the higher level of volcanic activity ostensibly present on early Earth, there are strong grounds to believe that montmorillonite was fairly common. Montmorillonite has been shown to enable the formation of oligomers containing as many as 40 to 50 monomers in the case of both activated nucleotides and amino acids, whereas only short oligomers ($\lesssim 10$ monomers in length) were synthesized in its absence. The reaction rate for montmorillonite-mediated oligomerization is much higher (by a factor of ~ 100 to 1000) than the rate of hydrolysis, implying that the issue posed by the latter is mitigated to some extent. Finally, montmorillonite offers the added benefit of regioselectivity—that is, the formation of phosphate ester bonds occurs along a preferred direction (in a manner consistent with biology). Overviews of this remarkable mineral and the experiments undertaken on this front can be found in Ferris (2006) and Maurel and Leclerc (2016).

Let us consider a simple model to illustrate how minerals (in their capacity as catalysts) expedite polymerization. In equilibrium, the forward and backward reaction rates of (2.32) should be identical, thus yielding

$$k_p \phi_m \phi_n = k_h \phi_{m+n}, \tag{2.36}$$

where k_p and k_h are the rates of polymerization and hydrolysis, respectively; we also introduce the notation $K_{eq} = k_p/k_h$ for the equilibrium constant. It must be cautioned that the concentration of water is absent herein, since it effectively remains constant throughout. Given that an n-mer has $(n-1)$ bonds, we assume that the formation of each bond has probability P. The chosen ansatz gives rise to the widely used expression $\phi_n = \phi_1 P^{n-1}$ for the concentration of the n-mer; the topic is tackled in Section 2.8.1. Upon substituting this formula into the above equation, we find $P = \phi_1 K_{eq}$. Next, let us suppose that the total number of monomers is fixed at a particular concentration (denoted by ϕ_T). The imposed constraint leads us to

$$\sum_{n=1}^{\infty} n\phi_n = \frac{\phi_1}{(1-P)^2} = \phi_T, \tag{2.37}$$

and solving for P, by using $P = \phi_1 K_{eq}$, yields

$$P = 1 + \frac{1}{2K_{eq}\phi_T} - \sqrt{\left(1 + \frac{1}{2K_{eq}\phi_T}\right)^2 - 1}. \tag{2.38}$$

The mean length of the polymer $\bar{n}$ is computed from

$$\bar{n} = \frac{\sum_{n=1}^{\infty} n\phi_n}{\sum_{n=1}^{\infty} \phi_n} = \frac{1}{1-P}, \tag{2.39}$$

and combining (2.38) with the above expression, we end up with $\bar{n} \propto \sqrt{K_{eq}}$ when $K_{eq} \to \infty$, implying that the mean polymer length diverges at this limit. It is easy to verify that $\bar{n}$ increases monotonically with K_{eq}. As the function of catalysts is to facilitate polymerization over hydrolysis, the desired behavior is effectuated by an enhancement of K_{eq}. The latter feature would result in the production of polymers of increasing length.

We round off the discussion with a couple of comments. First, the recognition that minerals (especially iron-rich clays) could play pivotal roles in the polymerization of biomolecular building blocks has a long and distinguished history dating back to J. D. Bernal's pioneering treatise at the

minimum (Bernal 1951). Second, apart from the multitude of benefits associated with minerals discussed herein, a radical hypothesis was tendered by A. G. Cairns-Smith in the 1960s. Cairns-Smith proposed that the first lifeforms arose from clay crystallites capable of (hereditary) replication and mutations. The central postulate was that lattice imperfections in crystals function as substitutes for genetic information, implying that variations in these defects serve as mutations transmitted from one generation to the next; a lucid exposition of the salient details can be found in Cairns-Smith (1982). While the original premise remains conceptually novel, it lacks sufficient experimental support at this stage.

2.4.4 Hydrothermal systems

Although the polymerization of amino acids and nucleotides is unfavorable in aqueous solutions at standard temperature and pressure, we turn our attention to hydrothermal systems. Let us begin with a consideration of peptide bond synthesis. Several experiments that mimic hydrothermal conditions have been conducted in this regard. Most of them used Gly as the amino acid monomer, and it has been shown that short oligomers—i.e., at least up to the heptamer $(Gly)_7$—could be produced. Another positive feature discovered via theoretical calculations is that the stability of peptide bonds increases with temperature and pH (up to a point) relative to hydrolysis reactions. Furthermore, thermodynamic studies of Gly have revealed that the production of Gly-Gly becomes significantly more exergonic at higher temperatures, thereby indicating that this reaction becomes relatively favorable.

An inescapable issue with most of the early experiments is that they did not satisfy at least one of the following criteria: (1) alkaline pH and relatively low temperature ($\lesssim 373$ K), (2) realistic concentrations of Gly, (3) long experimental times, (4) the exclusion of geologically implausible condensing agents. When these requirements are properly taken into account, the yields of glycine oligomers were either extremely low or below detectable levels (Bada 2013). The tide has, however, begun to reverse to some degree with the latest suite of experiments. At high pressures and in the presence of divalent metal ions and minerals, empirical results indicate that the polymerization of glycine is spontaneous and that the formation of linear peptides is favored (Pedreira-Segade et al. 2019).

Yet, at the same time, it is essential to recognize that the oligomers produced in most prior experiments were short. When this fact is viewed in

unison with the intrinsic difficulties arising from the synthesis of nucleotides from their individual components, one could be forgiven for concluding that the synthesis of oligonucleotides is well-nigh impossible. However, this premature conclusion ignores a series of exciting developments in recent times that incorporate characteristics of real-world hydrothermal systems. Most of them are centered on the inclusion of new physical processes and geological information that had mostly been neglected previously in origin-of-life studies.

First, in situ observations have revealed that hydrothermal vents exist in the vicinity of highly porous structures (mineral precipitates). Two physical mechanisms act in the pore: the first is thermal convection and the second is thermophoresis. The latter refers to the movement of molecules (or colloids) as a response to the application of thermal gradients. The collective action of these two processes leads to a tremendous accumulation of prebiotic molecules in these pores. Baaske et al. (2007) demonstrated, for instance, that the concentration of nucleotides could be amplified by a factor of $\lesssim 10^{10}$. This result, among several others, may act to render many, but probably not all, of the earlier objections regarding the low concentrations of prebiotic molecules invalid.

Let us denote the ratio of the concentration at the bottom of the pore relative to the top by Υ_A. It can be computed via

$$\Upsilon_A = \exp\left(0.42\, S_T \Delta T \varepsilon_p\right), \tag{2.40}$$

where S_T is the Soret coefficient (units of K^{-1}), ΔT denotes the temperature difference, and ε_p is the aspect ratio of the pore (treated geometrically as a rectangular cavity). It is the exponential dependence on ΔT and ε_p that is responsible for the major amplification in concentration. The Soret coefficient is not constant, as it depends on the number of monomers (n_m) that were assembled to form the molecule. For a single nucleotide, we have $S_T \sim 0.015\ \mathrm{K}^{-1}$, and using $\Delta T = 30$ K and $\varepsilon_p = 125$, we obtain the strikingly high value of $\Upsilon_A \sim 1.8 \times 10^{10}$. Another important parameter is the time (t_s) taken to attain a steady-state concentration,

$$t_s = \frac{\left(\varepsilon_p d_0\right)^2}{\pi^2 D_m}, \tag{2.41}$$

where D_m is the diffusion constant of the molecule (which depends on n_m) and d_0 is the optimal width of the pore; as the latter expression is rather

complicated, we shall not derive it here. The above equation can be heuristically understood as the time required for diffusion along the pore—that is, the effective area of cross section divided by the diffusion coefficient. In the case of a single nucleotide, it has been found that $t_s \sim 5 \times 10^4$ s (around 14 h), indicating that the process is very rapid.

If the concentration problem is perceived as having been resolved to some degree, the question of polymerization springs up. Thanks to recent theoretical and experimental work, the formation of oligonucleotides seems not only possible but also plausible, provided that certain issues such as kinetic stability are set aside. Despite this rosy picture, it is not obvious whether we can actually ignore this matter, given that many studies have emphasized the importance of dynamic kinetic stability (DKS) in recent times (Pross 2016). It has been proposed, for instance, that the maintenance of kinetic barriers to prevent metabolites from decomposing imposes constraints on the available free energy sources and may therefore assist in identifying what environments were more conducive to the origin of life (Pascal et al. 2013; Danger et al. 2020).

From a theoretical perspective, it has been shown that there exists a critical height of the aforementioned thermal trap (i.e., the pore) at which the total concentration exceeds the dissociation constant K_D associated with the hydrolysis of oligomers (Mast et al. 2013). Beyond this threshold, the production of long oligomers is dramatic as it displays hyperexponential growth. For a temperature gradient of 10 K, a feedstock (monomer) concentration of $\sim 10^{-9}$ M, and a dissociation constant $K_D \sim 10^{-5}$ M, it was found that the critical height is ~ 0.05 m. When the height exceeds this threshold, the production of oligomers up to 200 nucleotides in length at concentrations of $\sim 10^{-6}$ M becomes feasible.

Another problem has often been encountered in experiments involving genetic polymers. It is almost always the case that shorter polymers tend to replicate faster and therefore surpass their longer counterparts in competition with them, implying that the synthesis of long polymers ought to be difficult. In the 1960s, Sol Spiegelman and his collaborators demonstrated that a self-replicating RNA sequence was reduced to one-sixth of its original length after seventy-four generations and that its rate of synthesis became increasingly faster during the course of this process (D. R. Mills et al. 1967); subsequent experiments have been able to decrease the size of the replicator even further by an order of magnitude. This influential discovery was dubbed *Spiegelman's Monster,* and the observed tendency toward favoring shorter polymers has been christened Spiegelman's paradox.

Interestingly, the combination of influx, thermophoresis, and convection in the pore system has collectively enabled the circumvention of Spiegelman's paradox by ensuring that longer oligonucleotides survive while shorter ones are flushed out (Kreysing et al. 2015). The concentrations of long and short oligonucleotides (with seventy-five and thirty-six base pairs), denoted by ϕ_L and ϕ_S, is determined via

$$\frac{c_S}{c_L} = \frac{c_S^0}{c_L^0} \exp\left(-\Delta k t\right), \tag{2.42}$$

where c_S^0/c_L^0 is the initial concentration ratio and Δk represents the differential growth rate. Experiments have revealed that the latter quantity is positive, with $\Delta k \sim 1.5 \times 10^{-4}\ \mathrm{s}^{-1}$, implying that the long oligomers are more competitive with respect to the short oligomers, thus resulting in the ratio c_S/c_L declining monotonically over time. One caveat that must be recognized in connection with the aforementioned experimental results is that they have focused on using DNA nucleotides instead of RNA. When experiments were carried out involving RNA nucleotides in conditions resembling alkaline hydrothermal vents, only short oligomers (up to four units in length) were detected (Burcar et al. 2015). The nondetection of long RNA oligomers to date may imply that the synthesis of RNA is not easy in hydrothermal vents, although a great deal of work is necessary to (in)validate this claim.

The above mechanism for concentrating molecules applies not merely to RNA and DNA but also to other classes of biomolecules, such as lipids. It has been demonstrated in the case of oleic acid (viz., a fatty acid) that the combination of thermophoresis and convection yielded high concentrations that exceeded a critical threshold in some regions, consequently resulting in the local formation of large vesicles. Thus, it appears plausible that the first membranes and protocells could have been produced through such a process, perhaps in conjunction with evaporation (facilitating further concentration) in some geochemical environments. We refer the reader to Blain and Szostak (2014) and Toparlak and Mansy (2019) for summaries of breakthroughs in the laboratory related to the synthesis of artificial cells.

2.4.5 Other approaches to polymerization

Several other synthetic pathways have been studied in the laboratory, and a summary of these methods is found in Ruiz-Mirazo et al. (2014) and

Kitadai and Maruyama (2018). In what follows, only a few representative, but intriguing, examples are sketched for the sake of brevity.

When the volcanic gas carbonyl sulfide (COS) was passed through a solution containing the proteinogenic amino acid Phe at concentrations of 25 to 50 mM concentrations, the formation of dimers and trimers was observed within minutes to hours; what's more, the yield of dimers was quite high (up to 80 percent). Hence, this result implies that COS may have served as an effective condensing agent and mediated peptide synthesis on early Earth. The natural levels of COS reach a maximum mole fraction of $\sim 10^{-3}$, but the experiments made use of higher values. If our understanding of the Hadean eon is correct, the levels of volcanic activity were much higher than today, indicating that COS might have facilitated the synthesis of peptides, and perhaps other biomolecules, in local environments close to volcanoes.

We have seen in Section 2.3.3 that impact events can stimulate the production of biologically relevant monomers. It is natural to accordingly inquire whether shock-driven synthesis can also facilitate the formation of oligomers. This question was experimentally investigated in a study by Sugahara and Mimura (2014), who subjected a frozen mixture of alanine, water, and forsterite (a silicate mineral) at 77 K to extreme pressures between 4.8 and 25.8×10^9 Pa; the conditions were chosen so as to approximate cometary impacts. Linear dimers and trimers of Ala were synthesized during this process, and the equilibrium concentration of dipeptides in the early oceans was estimated to be $\lesssim 10^{-10}$ M.

In the prior section, we encountered the role of pores in promoting polymerization. It is known that gases emanating from magmas were in contact with water-filled pore networks (Arndt & Nisbet 2012). The heated gas circulating in these regions creates bubbles, whose significance in their capacity as water-gas interfaces was recently investigated by Morasch et al. (2019) via laboratory experiments. The authors found that many of the mechanisms presumed to be relevant or even necessary for prebiotic chemistry were operational in this setting: (1) dry-wet cycling that enabled phosphorylation, (2) accumulation of oligonucleotides and enhanced catalytic activity, (3) agglomeration of nucleic acids to yield gels, and (4) encapsulation of DNA into aggregated vesicles and subsequent fission.

2.5 THE RNA WORLD

Here we shall focus on a widely investigated hypothesis, or group of hypotheses to be strictly accurate, for the origin of life on Earth: the RNA

world. The central idea behind the RNA world is that RNA or its kin fulfilled the dual function of replication and catalysis. The RNA world has engendered a vast corpus of research and musings, consequently leading to the emergence of passionate defenders and dissenters. Myriad obstacles have arisen in the RNA world scenario—thereby giving birth to picturesque epithets such as the "Molecular Biologist's Dream" and the "Prebiotic Chemist's Nightmare" (Orgel 2004)—and a host of ingenious solutions have been proposed to overcome or ameliorate them.

2.5.1 A brief history of the RNA world

DNA currently serves as the repository of genetic information and is therefore essential for the reproduction and development of all known living organisms on Earth. Proteins play an equally vital role in biological functions, as they are presently responsible to the catalysis of innumerable reactions. If we suppose that one or the other must have emerged first, we immediately run into a conundrum. The replication of DNA is facilitated by enzymes (proteins), whereas the synthesis of proteins necessitates DNA as a starting point. At the risk of oversimplification, DNA requires proteins and proteins require DNA: we are thus confronted with a classic chicken-and-egg problem prima facie.

The easiest way out of this quandary is by positing the existence of a fictional molecule that is capable of carrying and transmitting genetic information on the one hand and serving as a catalyst on the other. Fortunately for us, biology on Earth already has an organic macromolecule that is known to undertake both tasks: RNA. The hypothesis that RNA, sometimes whimsically dubbed *biology's handmaiden,* was the first biopolymer endowed with these dual functions is not new. The history of the RNA world is both complex and long, making it difficult to trace who was the first scientist to definitively propose that RNA was responsible for both genetics and catalysis. The first explicit mention in the literature appears to have been by Alexander Rich in 1962. A few years later (in 1967–1968), three of the leading molecular biologists of the twentieth century—Carl Woese, Francis Crick, and Leslie Orgel—came up with variants of the central argument independently.[16] The near-concomitant blossoming of the RNA world

16. The fact that three groups converged on the same proposition in near simultaneity is uncannily reminiscent of the discovery in 1964 of the Higgs mechanism independently

paradigm constitutes an exemplar of how many scientific breakthroughs (serendipitous or otherwise) occur in *multiples* (Merton 1973).

While this idea was considered attractive, it was unclear in the 1960s whether RNA was genuinely capable of serving as a catalyst. The apparent absence of RNA-mediated catalysis was explained by the fact that protein catalysts were superior and had therefore replaced RNA in this capacity during the course of evolution. However, this picture was changed by Thomas Cech and Sidney Altman in the 1980s when they discovered that certain RNA molecules, called ribozymes (ribonucleic acid enzymes), were responsible for the catalysis of certain chemical reactions. This momentous development led to the award of the Nobel Prize to Cech and Altman in 1989. To date, ribozymes have been shown to play an important role in reactions pertaining to protein synthesis, RNA splicing, and the synthesis of transfer RNA (tRNA), to name a few. The notion of the RNA world was eloquently expressed by Walter Gilbert in 1986, and it seems apposite to quote his own words on this subject:

> And if there are activities among these RNA enzymes, or ribozymes, that can catalyse the synthesis of a new RNA molecule from precursors and an RNA template, then there is no need for protein enzymes at the beginning of evolution. One can contemplate an RNA world, containing only RNA molecules that serve to catalyse the synthesis of themselves. (p. 618)

Before we proceed, we wish to emphasize that the notion of the RNA world is loosely predicated on the idea that the same biopolymer was responsible for the storage and replication of genetic information and catalyzing chemical reactions.[17] However, the term *RNA world* is reasonably broad, encompassing a wide array of possibilities that are included under the same umbrella. One such example is the idea that some other primitive polymer with dual functionality predated RNA and might have "invented" it later. Detailed reviews of the RNA world paradigm can be found in Robertson

---

by (1) Robert Brout and François Englert, (2) Peter Higgs, and (3) Gerald Guralnik, C. R. Hagen, and Tom Kibble.

17. It must be recognized that the prospects for a DNA or hybrid DNA-RNA world, in which DNA evolved much earlier than canonically expected (perhaps even coevally with RNA), are not altogether negligible, especially given the advances in recent prebiotic pathways vis-à-vis the synthesis of DNA nucleosides (Teichert et al. 2019; Xu et al. 2019, 2020).

and Joyce (2012), Higgs and Lehman (2015), Wachowius et al. (2017), and Lazcano (2018), although the classic accounts by Joyce (2002) and Orgel (2004) are also highly recommended.

2.5.2 The difficulties and breakthroughs in the RNA world

Here we will categorize some of the most common criticisms that were leveled against the RNA world and the potential solutions that have been proffered in its defense; the reader should peruse Bernhardt (2012) for a cogent summary of the salient arguments *pro et contra* the RNA world.

2.5.2.1 Assembling ribonucleotides is tricky

We have already seen that RNA nucleotides (ribonucleotides) comprise three parts: nucleobases, ribose, and the phosphate group. Hence, as per the traditional line of thought, one would merely need to assemble together these building blocks, eventually culminating in the formation of RNA nucleotides. Yet, there are numerous difficulties linked with successfully implementing this approach—the reader may peruse Shapiro (2006) and Le Vay and Mutschler (2019) for lucid expositions—but several innovative solutions have nonetheless emerged in recent times (Islam & Powner 2017; Fialho et al. 2020; Yadav et al. 2020).

First, let us suppose that nucleobases are available in sufficient quantities, as there are a number of pathways by which they can be synthesized in situ—or externally delivered by means of comets and asteroids—as pointed out in Section 2.3, although the exact magnitude of their abundances remains an open question. Turning our attention to sugars, the most common pathway is the formose reaction, whereby a number of sugars are generated from formaldehyde, of which one of them is ribose. The major issue is that compounds with the carbonyl group (C=O) are very susceptible to the "asphalt problem"—that is, they react with other carbonyl-containing species and form complex organic goo. To counteract this tendency, it has been shown that the inclusion of borate minerals (and perhaps silicates) stabilizes 5-carbon sugars (Scorei 2012) and that the synthesis of ribose from formaldehyde and glycolaldehyde ($C_2H_4O_2$) is feasible in the presence of molybdate minerals (Benner et al. 2012). In view of the alleged importance of these two minerals, their preponderance on habitable worlds merits serious investigation in the future; the availability will, in turn, depend on the oxidation state of the atmosphere and the element inventory of the crust.

Next, we turn our attention to the formation of RNA nucleosides from nucleobases and ribose via N-glycosidic bonds. When dry mixtures containing nucleobases and ribose were subjected to heating in the presence of Mg^{2+} and inorganic polyphosphates, the four nucleosides were generated, although their yields were rather low except for adenosine. In contrast, the RNA nucleobases along with the corresponding nucleosides were synthesized at reasonable abundances by employing formamide in the presence of TiO_2. Lastly, the phosphorylation of nucleosides to produce nucleotides has been documented in formamide solutions. Even in aqueous conditions, there are a number of pathways that culminate in the production of canonical and noncanonical nucleosides. For instance, as outlined further in Section 2.7.2, the systematic action of wet-dry cycles permits the continual synthesis of the purine and pyrimidine nucleosides of RNA.

The pathway discovered by Powner et al. (2009) for synthesizing pyrimidine ribonucleotides is one of the crucial developments of the past two decades. The starting materials used in this pathway were fairly simple—cyanamide, cyanoacetylene, glycolaldehyde, glyceraldehyde ($C_3H_6O_3$), and inorganic phosphate—and the reactions were conducted in the presence of UV radiation. The chief breakthrough arose when, in lieu of the pyrimidine base being paired with ribose as orthodoxy would dictate, a half-nucleobase and half-sugar chimeric molecule was utilized (Sutherland 2010). Figure 2.8 presents a schematic overview of this pathway. While the publication remains a landmark achievement in synthetic chemistry, it has attracted some criticism as the reaction network is ostensibly reliant on the addition of reactive compounds, which are prone to the asphalt problem, at high concentrations in a particular order. Another potential drawback is that phosphate serves as catalyst and buffer in three different reactions. Hence, the success of this method is contingent on the availability of soluble inorganic phosphate.

After this pioneering work by Powner et al. (2009), a number of subsequent studies unveiled promising avenues for the synthesis of canonical and noncanonical nucleotides. It is instructive to contemplate a few examples, with the explicit understanding that this list is far from exhaustive. Kim and Benner (2017) combined nucleobases with phosphorylated carbohydrates (i.e., sugar phosphates) in a discontinuous fashion to obtain a variety of purine and pyrimidine nucleotides. Stairs et al. (2017) effectuated the synthesis of pyrimidine ribonucleotides by starting with a mixture of simple

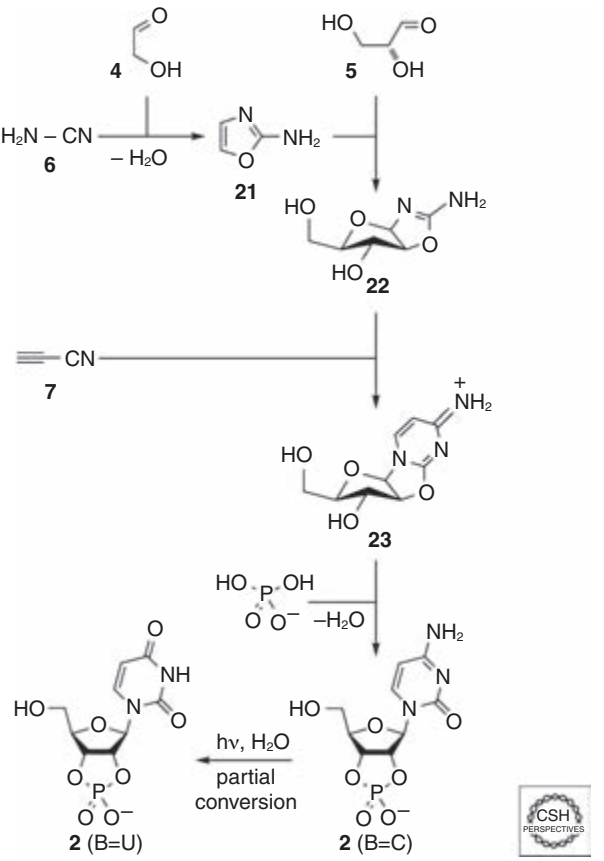


Figure 2.8 The pyrimidine nucleotides (2), where U is uracil and C is cytosine, are the final product, while 2-aminooxazole (21) roughly embodies a nucleobase-sugar hybrid. Note that (4)–(7) refer to glycolaldehyde, glyceraldehyde, cyanamide, and cyanoacetylene, respectively. The compounds (22) and (23) are intermediate compounds by the name of arabinose aminooxazoline and anhydronucleoside, respectively. (© 2010 Cold Spring Harbor Laboratory Press. *Source:* John D. Sutherland [2010], Ribonucleotides, *Cold Spring Harbor Perspectives in Biology* 2[4]: a005439, fig. 4.)

prebiotic compounds like glycolaldehyde, cyanamide, and thiocyanic acid. Islam et al. (2017) demonstrated the joint production of proteinogenic amino acids and ribonucleotides by starting with a complex mixture of 2-aminothiazole ($C_3H_4N_2S$), glycolaldehyde, and glyceraldehyde. A unified reaction network for prebiotic phosphorylation, polymerization, and the

self-assembly of major biomolecules—such as oligonucleotides and peptides from nucleosides and amino acids, respectively—was experimentally realized by Gibard et al. (2018), wherein the water-soluble diamidophosphate (DAP) ion $PO_2(NH_2)_2^-$ acted as the phosphorylating agent; DAP may have existed on Hadean Earth if schreibersite and other compounds had reacted with ammonia solutions.

Lastly, the phosphate ester bond that links together two nucleotides exhibits a uniform structure in RNA (or DNA)—that is, the 3′-end of one sugar ring is linked to the 5′-end of another. One could therefore suppose that replacing this specific linkage with another would cause a breakdown in the RNA functionality. However, laboratory experiments have demonstrated that this is not necessarily the case (e.g., when 2′-5′ linkages were substituted), implying that early biopolymers were possibly more robust than anticipated (Engelhart et al. 2013).

2.5.2.2 The replication of RNA

One of the basic premises underlying the RNA world is that RNA was capable of making a copy of itself (i.e., self-replication). There are a number of ways by which this mechanism could have operated, and the various candidates have been explored both experimentally and theoretically. The majority of these studies are based on template-directed replication, in which the replication process requires a preexisting template provided by a polymer strand of RNA; a mineral substrate (e.g., montmorillonite) may also assist in this regard.

The first approach is nonenzymatic replication—namely, an RNA strand synthesizes the complementary strand in the absence of any RNA / protein enzymes. While this route remains the most simple and direct one from a conceptual viewpoint (Sievers & von Kiedrowski 1994), empirical progress remains challenging owing to a number of issues—ranging from high error rates to the experimental reliance on unrealistically high concentrations of divalent metal ions, which induce the perils of hydrolysis—during the copying process. These issues, alongside the avenues whereby they were resolved to an extent, are summarized in Pressman et al. (2015) and Szostak (2012b, 2017). Despite the attendant difficulties, nonenzymatic synthesis of RNA oligomers $\gtrsim 100$ nucleotides in length has been actualized to date. As we will describe shortly hereafter, some of the issues associated with nonenzymatic replication are potentially bypassed at low temperatures

in ice. Experiments have demonstrated that oligomers enclosed in vesicles are stabilized to some degree against the effects of degradation, thus enabling them to efficiently participate in the act of nonenzymatic RNA synthesis (Adamala & Szostak 2013).

The second approach entails replication by means of an RNA polymerase ribozyme, i.e., an RNA molecule that catalyzes the polymerization (and replication) of RNA. In particular, if polymerase ribozymes are capable of copying RNAs of their own size (achieved recently, as elucidated below), they might pave the way toward synthesizing an RNA replicase, which has been christened the "Holy Grail" of the RNA world by some authors. This putative RNA molecule would function both as a template for storing and transmitting genetic information and as a polymerase that catalyzes its replication (Szostak et al. 2001). A careful scrutiny of this route reveals that there is a chicken-and-egg paradox since an RNA replicase cannot function as template and polymerase at the same time; in other words, at least two RNA molecules are necessary.

While a truly generic RNA replicase has not been achieved to date, there have been several breakthroughs in the field of RNA polymerase ribozymes. We shall not delve into this topic here, except for a few recent discoveries, as the reader may peruse the mostly up-to-date synopsis by Joyce and Szostak (2018). The first concerns the implementation of an RNA polymerase ribozyme by Attwater et al. (2013) that was capable of accurately synthesizing an RNA sequence longer than itself (up to 206 nucleotides in length), thereby possibly comprising the first steps in designing a bona fide RNA replicase. A notable feature of this discovery was that the polymerase under question attained its optimal functionality at temperatures as low as 254 K (−19 °C) in ices. The second concerns the fabrication of the 24–3 RNA polymerase that catalyzed the synthesis of a variety of complex RNA molecules, including transfer RNA, at fast rates, although the length of the templates was typically quite short (Horning & Joyce 2016). The last entails the execution of RNA-catalyzed RNA synthesis and replication of structured templates by deploying trinucleotide triphosphates (known as triplets) as substrates (Attwater et al. 2018).

The third category of interest is the concept of autocatalytic networks. In the strict sense of this phrase, all molecules in a given set are synthesized by reactions that are catalyzed by other molecules in that set. The important point to note here is that a given molecule is not autocatalytic, but the whole set is. A minimal representation of an autocatalytic set is

$$X_1 + Y_1 \xrightarrow{Z_2} Z_1; \qquad X_2 + Y_2 \xrightarrow{Z_1} Z_2, \tag{2.43}$$

with X_i and Y_i denoting the building blocks for Z_i, and $i \in (1, 2)$. A noteworthy example from this class was discovered only about a decade ago by Lincoln and Joyce (2009). Two ligases (enzymes that join two molecules) were shown to catalyze the synthesis of each other by starting from a total of four oligonucleotide precursors. The amplification of these molecules proceeded at a very fast rate, with a doubling time of about 1 h. Another impressive breakthrough entailed the identification of RNA fragments that self-assembled into a ribozyme (~ 200 nucleotides in length) by means of cooperative networks originated from simpler autocatalytic cycles (Vaidya et al. 2012). A cautionary point in connection with several laboratory studies of autocatalytic networks is that preexisting biological molecules (e.g., ribozymes) were fragmented and then investigated as to whether mutual catalysis and self-assembly could arise. In short, the selection of the underlying oligomers is not random, and they tend to be macromolecules in their own right, thereby raising the question of how they would have realistically arisen in prebiotic settings. In addition, when the number of molecules in the autocatalytic network is increased, one of the challenges from a practical standpoint is to ensure that each molecule can securely access its precursors.

Before proceeding, we observe that the three broad mechanisms for RNA replication delineated above are *not* exclusive. In fact, it seems plausible that they were capable of occurring in tandem to some degree.

2.5.2.3 The stability of RNA in water

One of the most commonly invoked arguments against the RNA world is that the RNA phosphodiester bonds linking ribonucleotides are quite susceptible to hydrolysis, as mentioned earlier. This issue is further compounded by the fact that RNA is very labile at high (or even moderate) temperatures.

If we accept the above statement at face value, a potential solution presents itself at once: namely, perhaps the RNA world arose in an icy setting. This topic has been the subject of much investigation in recent times and the results seem promising. For starters, it has been shown that ice increases the stability of RNA polymerase ribozymes and enables high-fidelity replication. Moreover, eutectic freezing in ice enables the concentration of molecules, as previously explicated. It is therefore noteworthy

that one of the longest RNA sequences synthesized by an RNA polymerase ribozyme was achieved within ice. On a related note, freeze-thaw cycles were successfully employed to synthesize more complex RNA polymerase ribozymes by starting from RNA oligomers with lengths ≤ 30 nucleotides by means of autocatalytic networks (Mutschler et al. 2015). In the absence of the temperature and concentration gradients induced by the freeze-thaw cycle, the catalytic activity of the network was diminished.

There are a couple of possibilities whereby long oligomers can be formed in liquid water without difficulty. First, instead of using the conventional nucleotides for assembly, one may opt to use their cyclic counterparts (Šponer et al. 2017). The latter, dubbed cyclic nucleoside monophosphates (cNMPs), are formed when the phosphate group forms bonds with two of the hydroxyl (-OH) groups of ribose. It has been shown that cNMPs yield oligomer chains of RNA $\gtrsim 120$ nucleotides in length with minimal assistance from organic or inorganic catalysts. Second, it is proven that the hydrolysis of RNA oligomers depends on the pH. Hence, as the phosphate ester bond is seemingly more stable at acidic pH under STP conditions, it is conceivable that the synthesis of RNA oligomers took place in such a medium. The third option, which we presented in Section 2.3.4, is to substitute formamide in place of water as the solvent.

2.5.2.4 Catalytic property of RNA molecules

One of the chief arguments against the RNA world is that only long RNA sequences possess the desired catalytic activity. If this point was indeed correct, clearly the synthesis of such long molecules in the first place would come across as highly improbable at first glimpse. For instance, let us suppose that we require an RNA molecule with ~ 200 nucleotides to serve as an efficient catalyst—that is, one that synthesizes RNA sequences as long as itself. The total number of combinations possible is $4^{200} \approx 2.6 \times 10^{120}$, clearly an unreasonably high number.

The most promising solution may stem from catalytic networks discussed in the preceding pages. In fact, we saw that oligomers ≤ 30 nucleotides in length gave rise to RNA polymerase ribozymes when subject to freeze-thaw cycles. Hence, if we choose a maximum cutoff length equal to this value, we find that the total number of combinations is $4^{30} \approx 10^{18}$. That is to say, given 10^{18} RNA oligomers, we may suppose that all of the desired molecules for initiating the catalytic network are present. While this

number may appear to be extremely large, it corresponds to approximately 2×10^{-6} mol. Consider a pool of radius ~ 10 m and depth ~ 1 m, and suppose that all these molecules are ensconced within it. By using this data, we deduce that the molar concentration ought to be $\sim 6.4 \times 10^{-12}$ M. Despite the fact that this abundance is probably not easily realizable, it is clear that the magnitude is, in itself, not very high.

Furthermore, if the 30-mers could, in turn, emerge from even smaller molecules via the same process, the desired concentrations of RNA oligomers will be further lowered and might end up being realistic. One problem with this proposed solution is that the catalytic activity of RNA oligomers usually decreases with length, implying that short oligomers are not likely to enhance the reaction rates significantly, which could therefore serve as a bottleneck for the emergence of longer oligomers with higher catalytic activity. While some innovative ideas and experiments have been advanced to bypass this conundrum, the pathways in question are not always founded on plausible feedstock molecules and environmental conditions.

In closing, it is imperative to recognize that we have highlighted only a fraction of the hurdles confronting the RNA world in the current era. Several others were omitted, of which the takeover of biological functions by protein enzymes from ribozymes is one of the most prominent (Bowman et al. 2015).

2.5.3 The possibility of pre-RNA worlds

As we have witnessed, the emergence of the RNA world is fraught with major obstacles, although one should also appreciate that an array of potential solutions have been propounded to overcome them. The existence of these conundrums has led many authors to seek out prior alternatives to RNA, often collectively known as pre-RNA worlds, which endeavor to bypass the accompanying RNA-related issues; the reader is referred to Hud et al. (2013), Cleaves et al. (2015), and Fialho et al. (2020) for cogent analyses of this burgeoning topic.

The first option is to expand or contract the genetic alphabet: not surprisingly, there are many examples of noncanonical base pairs. One such candidate, purine-2,6-dicarboxylate·copper(II)·pyridine, is mediated by the divalent copper anion (Cu^{2+}). This base pair was "recognized" by DNA polymerase in the presence of Cu^{2+}, implying that copper ions, and perhaps

other transition metals, might play a role in nucleotide synthesis. The second route entails the replacement of the phosphate group with a different functional group. However, there are several physicochemical advantages associated with phosphates as described in Section 2.2.5: it is nearly ubiquitous in the major components of cells, and its replacements have not been shown to be equally successful (as of yet). Hence, even if phosphate was not present in the first nucleic acid-like molecules—that is, xenonucleic acids (XNAs)—it seems to us that phosphate was an essential requirement for life on Earth in other respects, although further research is clearly needed to assess the veracity of this statement.

Setting aside the nucleobases, the next component worth addressing is the sugar-phosphate backbone of RNA formed via phosphodiester bonds. We have already commented on some alternatives in Section 2.2.3, and we recommend the reader peruse through it again. Two candidates that have been studied in this regard are threose nucleic acid (TNA) and glycerol nucleic acid (GNA), in which ribose is replaced by threose (a 4-carbon sugar) or glycerol (a 3-carbon non-sugar). These alternatives were chosen since they are less complex than ribose (a 5-carbon sugar), and this opens the possibility that they were more readily available than ribose on the young Earth. Apart from their capacity for base pairing, XNAs must also possess the ability to form duplexes with RNA or perhaps DNA, as they would serve as the initial template for the synthesis of nucleic acids, after which RNA and / or DNA would take over the genetic and catalytic functions. Recent experiments have revealed that both TNA and GNA fulfill these two basic criteria, while TNA also forms duplexes with its complementary strand. In addition, TNA has the rare ability to form three-dimensional structures capable of multiple functions, such as binding selectively to various proteins with high affinity.

Unlike the XNAs discussed hitherto, we may envision a candidate in which the sugar-phosphate backbone is dispensed with altogether. This idea led to the groundbreaking design and synthesis of peptide nucleic acid (PNA) by Peter E. Nielsen, Michael Egholm, Rolf H. Berg, and Ole Buchardt (Nielsen et al. 1991). The authors ingeniously substituted a polyamide backbone in place of the standard deoxyribose-phosphate backbone of DNA. The former was produced from the polymerization of N-(2-Aminoethyl)glycine. In place of the glycosidic bond that links sugars to nucleobases in RNA and DNA, the nucleobases are connected to the

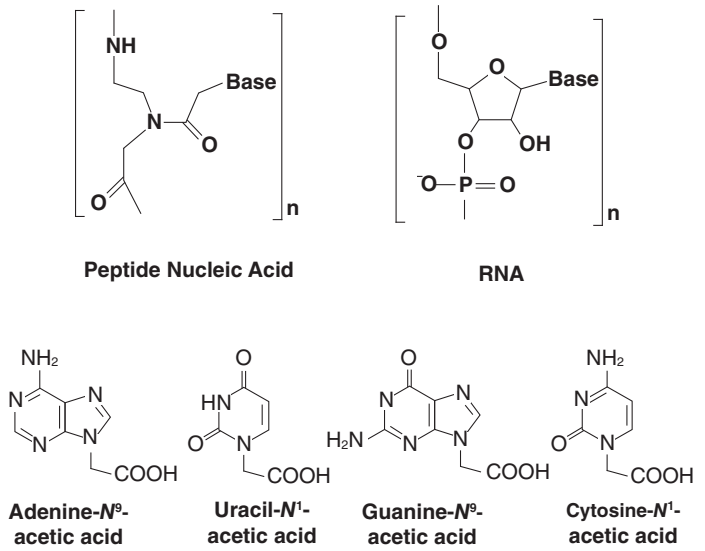


Figure 2.9 *Top row:* The monomer units of PNA are compared against their RNA counterparts. Note that the sugar-phosphate backbone is replaced with a polyamide structure. *Bottom row:* The four bases of PNA consist of the canonical RNA nucleobases bonded to an acetic acid linker. (© National Academy of Sciences. *Source:* Kevin E. Nelson, Matthew Levy, and Stanley L. Miller [2000], Peptide nucleic acids rather than RNA may have been the first genetic molecule, *PNAS* 97[8]: 3368–3871, fig. 1.)

polyamide structure via methylenecarbonyl bonds. The structure of PNA and its components is depicted in Figure 2.9. PNA possesses many desirable characteristics, some of which are outlined below.

- Unlike RNA and DNA, PNA is achiral in nature, namely, it has mirror symmetry. This feature is particularly relevant since explaining how and why biology uses chiral molecules—such as amino acids, sugars, and nucleic acids—remains a highly challenging endeavor, whereas this issue is not relevant for achiral molecules like PNA.
- PNA is an uncharged polymer, in contrast to RNA and DNA. The consequent relative lack of electrostatic repulsion plays an important role in enabling PNA to bind to complementary strands of RNA / DNA with high affinity. In addition, it was demonstrated that this binding is highly specific, for reasons that we shall not outline here.

- The building block of PNA, aminoethylglycine, was synthesized via Miller experiments when a gaseous mixture of NH_3, CH_4, N_2, and H_2O was subjected to electrical discharge (K. E. Nelson et al. 2000). Moreover, PNA-related molecules have been documented in carbonaceous chondrites.
- PNA oligomers are known to serve as effective templates and enable the transfer of information from PNA to RNA and DNA via the polymerization of their monomers.
- PNAs have proven to be successful when employed as catalysts (PNAzymes). In addition, PNAs were combined with other functional groups (e.g., thioesters) to produce molecules that exhibit traits of both nucleic acids and proteins.

Hence, PNAs do possess many desirable attributes prima facie; they are intriguing candidates insofar as pre-RNA worlds are concerned (Sharma & Awasthi 2017). On the other hand, however, further experiments simulating plausible geochemical conditions are necessary in order to properly gauge the feasibility of PNA worlds.

The relative merits of the sundry candidates for pre-RNA worlds, when measured against the RNA-first hypothesis (the strong RNA-world model), are far from being fully understood. When all is said and done, perhaps the most parsimonious explanation is that RNA was the first genetic polymer endowed with catalytic functions, instead of having to invoke one or more pre-RNA worlds that remain shrouded in mystery.

2.5.4 The RNA-peptide world

Several hypotheses contend that RNA and peptides (or proteins) coevolved together or that peptides and their variants wholly preceded RNA in prebiotic evolution (C. W. Carter 2015; Frenkel-Pinter et al. 2020). It must, however, be noted at the outset that this stance is not universally accepted; we refer the reader to Danger et al. (2012) and Wachowius et al. (2017) for detailed critiques of this topic. One noteworthy example that has been advanced in support of an RNA-peptide world is founded on putative evidence that the cores of large and small subunits of ribosomes (to wit, the sites of protein synthesis), existent in all known living organisms on Earth, contain a mixture of ribosomal RNAs and polypeptides. Phylogenetic studies of ribosomal RNA and ribosomal proteins have

also been adduced to bolster the claims of an RNA-peptide world, but these methods and conclusions have attracted some criticism (Bernhardt 2012).

If short peptides were indeed essential to the RNA world (or the lipid world, addressed in Section 2.5.5), there are multiple avenues through which they could have exerted their influence. To begin with, in a manner analogous to their protein successors, they may have enabled the catalysis of simple metabolic reactions. Looking beyond, they might have regulated the permeability of protocells, promoted the formation and stabilization of vesicles with the capability to enclose RNA (Cornell et al. 2019), and operated as cofactors (helper molecules necessary for enzymatic activity) for ribozymes. To round off this brief interlude, one of the primary roles of simple peptides possibly involved enhancing the thermal stability of three-dimensional RNA structures to a substantial degree.

When it comes to peptide formation, we already witnessed that many pathways are feasible for the synthesis of amino acids and that several mechanisms can enable the polymerization of amino acids to yield peptides. Hence, at the very least, it seems likely that the RNA-peptide world cannot be discounted solely based on geochemical considerations. It is also worth remarking that the formation of increasingly complex peptides might have taken place by virtue of autocatalytic networks (Lee et al. 1996; Zozulia et al. 2018)—although much research remains to be done on the experimental front—enabling peptides to conceivably fulfill cardinal and variegated roles in the origins of life.

2.5.5 The lipid world

We have already seen in Section 2.2.4 that lipids constitute an essential component of protocells. Reams of research have been undertaken, in terms of laboratory experiments and theoretical modeling, concerning the formation, growth, reproduction, and evolution of vesicles and other primitive compartmental structures from lipids. Two noteworthy aspects to bear in mind with regard to the formation of vesicles are (1) the reaction network is patently autocatalytic (Lancet et al. 2018), thereby ensuring that the newly formed vesicle accelerates the production of other vesicles, and (2) many types of surfaces catalyze the formation of vesicles, with minerals (e.g., montmorillonite) especially deserving of attention (Dalai & Sahai 2019).

For reviews of this subject, we direct the reader to Segré et al. (2001), Luisi (2016), Deamer (2017), and Joyce and Szostak (2018). Although there are myriad functions played by lipid vesicles, we shall primarily concern ourselves with their role in the RNA world paradigm. It is, however, imperative to remain cognizant of alternative hypotheses that posit the existence of pre-RNA lipid worlds. As per these scenarios, the immense diversity of organic amphiphiles predicted or observed in various astrobiological settings suffices to instigate credible pathways oriented toward the formation of micelles and subsequent emergence of protocells. For instance, mechanisms central to life such as self-assembly, replication, and selection are known to manifest in experiments involving lipid building blocks (Colomer et al. 2020). It has been suggested by Kahana et al. (2019), on the basis of the analysis of complex organic molecules detected in the plumes of Enceladus by the *Cassini* mission, that this moon is conducive to the origin of life by means of a pre-RNA lipid world.

Rajamani et al. (2008) investigated the fate of mononucleotides when they were subjected to wet-dry cycles in a lipid environment. At the end of the experiment, it was found that RNA-like polymers were assembled and thereafter encapsulated within vesicles. This discovery is pertinent because the first cell membranes supposedly made of fatty acids may have been leaky—to wit, it is plausible that they enabled the transport of organic molecules such as nucleotides from the exterior environment into the interior of the vesicle. At the same time, however, it is believed that these membranes were sufficiently robust to maintain structural stability over the temperature range where water is liquid at surface pressure and prevent biopolymers within the vesicle from diffusing outside.

Chen et al. (2004) demonstrated that the growth of vesicles and RNA (or DNA) replication are related to one another. When RNA situated within the vesicle undergoes replication, it leads to the vesicle swelling up, thereby exerting osmotic pressure on its membrane. The resultant intensification in membrane tension drives an increase in its area through the uptake of additional membrane components. Hence, it was shown that RNA replication is directly linked with protocell growth, implying that vesicles with efficient RNA replication (and faster growth) will outcompete other vesicles, consequently facilitating the potential emergence of Darwinian evolution at the level of protocells.

In a similar vein, it has been observed that the amplification of DNA (via replication) drove the division (reproduction) of vesicles, illustrating

once again that the efficiency of replication appears to be connected with the growth and division of vesicles. Other studies have also documented the two-way coupling between enclosed catalysts (e.g., ribozymes) and lipid membranes. For instance, vesicles endowed with an enzyme that promoted the synthesis of membrane components were concomitantly stabilized against the effects of Mg^{2+} (Adamala et al. 2016). This finding acquires significance because such vesicles would be preferentially selected, which is advantageous since Mg^{2+} is often recruited for ribozyme activity and oligomerization of RNA (cf. Motsch et al. 2020).

We bring this abbreviated discussion to a close by noting that lipid membranes can also serve as selective binding sites for the canonical nucleobases of RNA and ribose (Xue et al. 2020). As a result, they are capable of assisting in the formation of RNA nucleosides by ensuring that the apposite compounds are chosen from the plethora of prebiotic chemicals. In concurrence, the binding of ribose and RNA nucleobases to lipid membranes also confers a greater stability to the latter in aqueous media. Thus, it would seem reasonable to contend that lipids, RNA bases, and ribose are mutually beneficial to each other under certain circumstances.

2.6 DID METABOLISM ARISE FIRST?

> Life is the mode of existence of protein bodies, the essential element of which consists in *continual metabolic interchange with the natural environment outside them,* and which ceases with the cessation of this metabolism, bringing about the decomposition of the protein. Such metabolism can also occur in the case of inorganic bodies and in the long run it occurs everywhere, since chemical reactions take place, even if extremely slowly, everywhere. The difference, however, is that inorganic bodies are destroyed by this metabolism, while in organic bodies it is the necessary condition for their existence.
>
> —Friedrich Engels, *Dialectics of Nature*

Two of life's fundamental characteristics are replication and metabolism. The RNA world, which we have covered extensively, is based on the premise that replication arose first in the form of RNA or a pre-RNA molecule endowed with similar properties. However, as we have also seen, the RNA world has been criticized on several grounds. Moreover, even if we assume that the synthesis of RNA was somehow plausible, the leap from RNA to proteins and DNA remains unexplained by the RNA world.

It behooves us therefore to seriously contemplate the other possibility—namely that metabolism arose before replication. By *metabolism* here and elsewhere, we refer to the complex and coordinated set of chemical reactions that furnish all living entities with the requisite energy to sustain their existence. A strikingly modern exposition of the central role played by metabolism in life's functioning was presented by Friedrich Engels in his monograph *Dialectics of Nature* (1883), as illustrated by the above quotation. The majority of "metabolism-first" hypotheses are based on the premise that small organic molecules participate in a network of self-sustaining reactions (autocatalytic cycles), thereby producing energy. One of the chief reasons behind the attractiveness of metabolism-first is that the emergence of these small molecules is averred to be much more plausible than that of RNA, which is a very complex molecule. Yet, at the same time, it is necessary to recognize that not all "replication-first" hypotheses should be equated with those entailing RNA or its antecedents serving as the replicator.

A few aspects are worth emphasizing before proceeding further. First, the apparent dichotomy between these two basic paradigms is not as clear-cut as it may appear at first glimpse (Preiner et al. 2020a). For instance, Copley et al. (2007) proposed that mutual catalysis aided in the emergence of the RNA world by way of protometabolic networks, implying that genes and metabolism could have coevolved together: their reaction network was based on pyruvate (CH_3COCOO^-) as the starting point and culminated with the production of ribonucleotides. On a related note, a fair number of theoretical models predicated on the existence of metabolic replicators tend to draw on elements of both the metabolism- and replication-first hypotheses. Other authors, notably Fry (2011), have accordingly pointed out that the classification scheme of "replicants" and "metabolists" does not truly capture the latent subtleties.[18]

Alexander I. Oparin, and J. B. S. Haldane to a lesser extent, was the forerunner of modern metabolism-first theories in many respects, although his formulation cannot and ought not be viewed as a metabolism-first theory sensu stricto. Over a series of groundbreaking publications in the 1920s and 1930s, Oparin proposed that organic molecules endowed with catalytic properties gave rise to protometabolism within colloidal droplets (Oparin

18. The adoption of *replicants* in place of *geneticists* is recommended because the latter seems overly Earth-centric. In addition, the word *replicants* pays due homage to *Blade Runner,* a classic science-fiction movie adapted from the maverick novel by Philip K. Dick.

1938) and that the influx of new materials from the environment actuated processes such as growth, division (self-reproduction), and heredity (transmission of "organization"). Oparin's so-called coacervates—colloidal droplets infused with chemicals formed via liquid-liquid phase separation—retain some, although by no means all, of the central features prevalent in modern metabolism-first hypotheses (Lazcano 2010). While Oparin's work tended to emphasize metabolic functions, it was suggested by Vasas et al. (2012) that coacervate-like entities were theoretically capable of Darwinian evolution (by natural selection) via the inception of autocatalytic networks.

Another notable aspect in which Oparin's hypothesis differs from the majority of current metabolism-first models is the geological setting. Oparin posited that coacervates existed in a prebiotic soup on the surface, whereas the bulk of modern research in this realm has often tended to focus on hydrothermal vents at the bottom of the ocean. Notwithstanding these divergences, the coacervates paradigm has witnessed a partial resurgence on both theoretical and experimental fronts over the past decade. State-of-the-art analyses indicate that these droplets exhibit a diverse array of desirable characteristics including catalysis, spontaneous self-organization, selective sequestration and polymerization of prebiotic molecules, and cell-type growth and division by means of external energy sources (S. Koga et al. 2011; Strulson et al. 2012; Zwicker et al. 2017; Drobot et al. 2018; Poudyal et al. 2019; Donau et al. 2020). Aside from coacervate droplets, which typically comprise oppositely charged polymers (e.g., polypeptides and nucleotides), other membraneless droplets—such as those composed of α-hydroxy acids (Jia et al. 2019)—also evince the capacity for preferential self-assembly, sequestration, catalysis, and functionality of biomolecules.

The number of metabolism-first hypotheses are manifold and encompass multiple viewpoints and methodologies, owing to which we focus on a couple of specific metabolic pathways in addition to a brief description of collectively autocatalytic sets introduced by Stuart Kauffman and Freeman Dyson in the 1980s as a theoretical model for abiogenesis. A point to emphasize here is that the formation and reproduction of vesicles is autocatalytic and often classified under metabolism-first hypotheses. As we already touched on this topic in Section 2.5.5, we will not address it here.

2.6.1 Autocatalytic sets

The pioneering notion of collectively autocatalytic sets (CAS) was originated by Stuart Kauffman in 1971, although it was explicitly articulated in the context of peptide synthesis during the 1980s (Kauffman 1986). Freeman Dyson proposed a similar concept in 1982, which was further elaborated on in his elegant book a few years later (Dyson 1985). We will provide a short summary of their respective approaches here, although the reader is recommended to consult the original publications. Before embarking on our discussion, it is important to recognize that the basic axioms of autocatalytic sets are quite general and therefore not restricted only to metabolic networks. In fact, we already encountered a couple of autocatalytic systems when discussing the RNA world (replication-first) in Section 2.5.2.

As per Dyson's schema, the protocell is endowed with N active sites on its interior with the capacity to either adsorb or desorb monomers with equal probability. The monomers are divided into two classes, active and inactive, which are duly assigned the values of 1 and 0, respectively. An active monomer in this context refers to one that is capable of rendering other monomers active, i.e., it has catalytic properties—while the inactive ones induce zero catalysis. The monomers are assumed to be sampled from $(m+1)$ chemical species, of which only one species is active at a *given* site. It is assumed that the rate of desorption or adsorption for a single species (per unit time) is uniformly equal to r. Hence, the total rate of adsorption of inactive species equals mr because there exist m of them.

The total rate of desorption of monomers per filled site is given by qr, where q is an efficiency factor (not necessarily equal to unity) governed by the temperature and the energy required for dissociating the monomer from the binding site. The rate of adsorption of an active species monomer per site is given by $\psi(x)r$, and this factor of $\psi(x)$ (in place of unity) accounts for the catalytic effect of other monomers in promoting adsorption of the monomer under consideration; here, x is defined as the fraction of sites that are occupied by active monomers. If we specialize to a steady-state, the fraction of sites that have active monomers, have inactive monomers, and are empty will be proportional to ψ, m, and q respectively. Thus, by combining these postulates, we obtain

$$x = \frac{\psi(x)}{m+q+\psi(x)} = \left[1 + a_0\,(\psi(x))^{-1}\right]^{-1}, \tag{2.44}$$

where $a_0 \equiv m + q$. In general, the form of $\psi(x)$ is indeterminate, but we can invoke an ansatz along the lines proposed in Section 2.4.3. Consider the idealized scenario in which every site features an active monomer, and we suppose that the collective action of this perfect catalyst lowers the activation energy for the binding of an active monomer by ΔE. However, as per this model, only a fraction x of the sites are active, because of which the activation energy is lowered by only $x\Delta E$. Dyson proposed that the total enhancement $\psi(x)$ is encapsulated by the Boltzmann factor, i.e., we make use of

$$\psi(x) = \exp\left(\frac{x\Delta E}{k_B T}\right) = b_0^x, \tag{2.45}$$

where $b_0 \equiv \exp(\Delta E/k_B T)$. By substituting the above expression into (2.44), one can scan the parameter space for optimal choices of a_0 and b_0. Dyson concluded that, roughly speaking, the most likely values for these parameters are $a_0 \sim 8$ to 10 and $b_0 \sim 60$ to 100. Recall that a_0 serves as a proxy for the number of monomer species and b_0 for the gain in reaction rates via catalysis. Perhaps the most interesting point is the prediction that $a_0 = 4$ is disfavored, implying that the likes of RNA / DNA might not provide enough chemical diversity.

It is also possible to calculate the probability P_j of finding j active monomers in the steady-state, which we shall not tackle here. It suffices to state that the final expression takes the form

$$P_j = K_0 \exp\left[-N V_{\text{eff}}\left(\frac{j}{N}\right)\right], \tag{2.46}$$

where K_0 is a constant, and the effective potential V_{eff} is given by

$$V_{\text{eff}} = x \ln x + (1-x)\ln(1-x) + x \ln a_0 - \frac{x^2}{2} \ln b_0. \tag{2.47}$$

The existence of two minima and one maximum, in certain regions of parameter space, can be verified for V_{eff}. Of these, the lowest minimum S_1 represents the disordered state, the maximum S_2 presents a barrier that must be crossed, and the second minimum S_3 is the ordered state. Let us suppose that ΔU represents the normalized height of the barrier separating the maximum and minimum potential energies, i.e., S_2 and S_1, respectively. It is quite natural to model the time taken for the N-component cell to move

from an ordered to a disordered state via an exponential dependence on N and ΔU as follows:

$$t_f = \tau_a \exp(\Delta U N), \tag{2.48}$$

where t_f is the transition time and τ_a is the average time interval between adsorptions (or desorptions) at each site. Although this function is smooth, one can qualitatively identify a critical value of N above which the transition time becomes excessively long. It is, somewhat arbitrarily, assumed to be $N_c = 30/\Delta U$, as this would result in the very large value of $t_f/\tau_a \approx 10^{13}$. For the ranges of a_0 and b_0 discussed earlier, one ends up with $N_c \sim 2000$ to 20,000, indicating that the number of monomers in the protocell was probably smaller than (or comparable to) this value. Many other interesting points emerge from Dyson's analysis, one of which concerns the amount of information contained in the protocell. The probability of a binary message with information content $\mathcal{I}$ is given by $2^{-\mathcal{I}}$. If we equate this to the probability of moving from S_1 to S_3 and use (2.46), we can obtain the value of $\mathcal{I}$. In the case of large a_0 and large b_0, we find

$$\mathcal{I} \approx \frac{N}{2} \log_2 \left(\frac{b_0}{a_0^2} \right), \tag{2.49}$$

implying that the number of bits per site is dictated solely by b_0 and a_0. For instance, selecting $b_0 = 100$ and $a_0 = 8$ leads us to $\mathcal{I} \approx 0.32N$. Before closing our discussion of this model, we emphasize that its value is primarily conceptual as there are many idealized assumptions throughout.

Let us now turn our attention to Stuart Kauffman's corpus of research. We begin by considering a set of two distinct monomers and all combinations of polymers (formed from these two monomers) up to some specified length M. It can immediately be verified that the total number of possible configurations of polymers up to this length is $N_M = 2 + 2^2 + \cdots + 2^M = 2^{M+1} - 2$, and the second term can be neglected even for moderate values of M. Next, we shall calculate the number of reactions that are possible for all possible M-mers. It is evident that any M-mer is the result of $(M-1)$ condensation reactions, i.e., there are $(M-1)$ bonds formed. We must also take into account the fact that there exist 2^M combinations of M-mers in total. Thus, the total number of condensation reactions for M-mers is $2^M(M-1)$, as it represents the product of number of reactions per M-mer and the number of all possible M-mers. If we wish to determine the number

of condensation reactions in the set, denoted by R_M, we must carry out the same calculation for the lower order polymers as well, thus yielding

$$\begin{aligned} R_M &= 2^M\,(M-1) + 2^{M-1}\,(M-2) + \cdots 2^{M-(M-2)}\,(M-(M-1)) \\ &= 2^{M+1}(M-2) + 4, \end{aligned} \tag{2.50}$$

where the expression in the second line of the right-hand side follows from the fact that the first line of the right-hand side is a simple arithmetico-geometric series. For large M, it is therefore evident that $R_M/N_M \approx M-2$; in other words, the total number of permissible condensation reactions increases linearly with M relative to the total number of feasible molecules. Hence, while the number of potential molecular combinations may increase, this boost is superseded by the higher number of reactions, thereby increasing the likelihood of catalysis on the whole.

The next topic that we must address is the likelihood of the reaction network possessing an autocatalytic subset. Recall that, in an autocatalytic (sub)set, the formation of each entity must be catalyzed by at least one other entity. Kauffman (1986) proposed a sufficient condition for the existence of the autocatalytic subset: the longest polymer within this subset (which has length M) must have at least one reaction in the last step of its formation that is catalyzed by some member of the subset. Let us adopt the idealized situation in which each polymer catalyzes a reaction with the same probability P_0. We have seen previously that there exist $(M-1)$ condensation reactions to form a given M-mer from an $(M-1)$-mer, and the total number of other polymers is approximately 2^{M+1}. Thus, the probability $\bar{P}$ that a given polymer of this length is *not* catalyzed is given by

$$\bar{P} = (1-P_0)^{(M-1)2^{M+1}}. \tag{2.51}$$

If this probability is sufficiently low, it would imply that the existence of an autocatalytic set is at least plausible. Motivated by this desideratum, Kauffman suggested that a small-enough value for $\bar{P}$ should be on the order of 10^{-3}. By using this criterion, we solve for M as a function of P_0, thereby ending up with

$$M \sim 1.44\,\mathrm{W}\left(\frac{2.4}{P_0}\right), \tag{2.52}$$

where $\mathrm{W}(z)$ is a mathematical function known as the Lambert-W function; we have assumed $M \gg 1$ and $P_0 \ll 1$ to simplify matters. The total set of

polymers that is necessary for the existence of an autocatalytic set is $N_M \approx 2^{M+1}$. For instance, if we consider $P_0 \sim 10^{-5}$, we find $M \sim 14.5$ and $N_M \sim 4.8 \times 10^4$. We have only focused on condensation reactions, but cleavage and exchange reactions are also feasible. Kauffman showed that (2.52) must be replaced by approximately half its value. Thus, the length of the polymer M is not affected much, but this leads to a significant reduction in N_M. However, we are free to work with (2.52) as a reasonable upper bound.

The above result is predicated on the unspoken assumption that P_0 is independent of M. However, suppose that P_0 could be expressed as the product of two factors: the probability that a given polymer serves as a catalyst (P_0') and the probability that it catalyzes a *particular* reaction (P_0'')—that is, we have $P_0 = P_0' \cdot P_0''$. If we choose $P_0'' = 1$, clearly $P_0 = P_0'$, and (2.52) is duly obtained. On the other hand, recall that the total number of reactions feasible was determined in (2.50). If a polymer can catalyze only *one* reaction, it follows that $P_0'' = R_M^{-1} \approx (M-2)^{-1}\, 2^{-(M+1)}$. Using this value of P_0 in (2.51), it can be shown that $\bar{P}$ has an approximate lower bound of e^{-1} for large M. One of the resultant consequences is that the discussion following (2.51) would be rendered invalid. This analysis, attributable to Lifson (1997), serves to underscore the fact that the interpretation and choice of P_0 play a significant role in determining the likelihood of the emergence of autocatalytic sets.

For the sake of building intuition, let us set aside the above caveat, while acknowledging its importance. The basic qualitative picture of the reaction network is as follows. If we have a high probability of catalysis for each polymer, the likelihood of the autocatalytic set becomes almost unity. This can be seen from inverting (2.51) to solve for M and observing that the result diverges when $P_0 \to 1$. Thus, there is a critical transition of sorts from the case where only small and isolated subsets of reactions are catalytic to the scenario in which the whole reaction network becomes autocatalytic. In a subsequent study, Farmer et al. (1986) demonstrated via numerical simulations that the critical probability P_c at which the likelihood of an autocatalytic set becomes close to unity was given by

$$P_c \approx \mathcal{Q}^{-2M}, \tag{2.53}$$

where $\mathcal{Q}$ is the number of different monomers that exist. This result is closely connected to the theory of random graphs formulated by the mathematicians Paul Erdős and Alfréd Rényi in a seminal paper (1959), as rightly noted by Kauffman (1986).

Hence, we will take a brief detour into random graph theory to trace the origin of the above phenomenon. Let us suppose that there exist N different species in the reaction network. If we have $\mathcal{Q}$ different monomers and the maximum length of the polymer is M, then we have $N = \mathcal{Q} + \mathcal{Q}^2 + \ldots \mathcal{Q}^M \approx \mathcal{Q}^M$, where the last equality follows only in the limit $\mathcal{Q} \gg 1$. In the graph-theoretic picture, we shall envision each species as a node; the total number of nodes equals N_M (we have $N_M = N$ in our model). If one species catalyzes a reaction involving another species, this corresponds to joining the two nodes via an edge on the graph. Thus, the probability of two nodes being connected is given by P_0, which is just the probability of catalysis introduced previously. With this setup, the matter of interest to us concerns the appearance of subgraphs with q nodes and ℓ edges.

As we shall see later, some of the most popular hypotheses for metabolism-first models rely on potentially autocatalytic cycles. Thus, we shall focus primarily on cycles in this example. In graph theory, a cycle is a closed loop whose defining characteristic is that a common node is shared only by two consecutive edges. A cycle is precisely what one would imagine intuitively, i.e., a triangle is a cycle of order 3, a quadrilateral is a cycle of order 4, and so on. Thus, it can be verified that cycles are characterized by $q = \ell$. With this background, let us evaluate the number of subgraphs N_G with q nodes and ℓ edges that can be formed from the graph with N_M nodes.

As we can choose q nodes from N_M nodes, the total number of ways of doing this is given by the binomial coefficient C_q^N. Next, the probability of forming ℓ edges is given by P_0^ℓ. Finally, the q nodes selected can be permuted among themselves to give rise to different graphs, implying that the extra term $q!$ must be included. Mathematically, some of these graphs may be isomorphic to one another, and this is accounted for by dividing $q!$ with a constant factor. We do not include it in our analysis, as it does not affect our results. Thus, by including all these terms, we obtain

$$N_G = C_q^N \, q! \, P_0^\ell = \frac{N!}{(N-q)!} P_0^\ell \approx N^q P_0^\ell, \tag{2.54}$$

and the last equality on the right-hand side is formally valid provided that $q/N \ll 1$, which is reasonable when N is very large. Let us introduce the probability $P_c = N^{-q/\ell}$, as it serves an important purpose. If we consider

the case with $P_0 > P_c$, it can be shown that N_G diverges when N and ℓ are large. Similarly, if we have $P_0 < P_c$, it is found that N_G becomes infinitesimally small under the same conditions. Thus, we see that P_c functions as a critical probability that lies at the threshold of two distinct outcomes. When a rigorous derivation is carried out, an extra factor appears in the expression for P_c but does not change the scaling with respect to N (Albert & Barabási 2002). We use the relations $N \approx \mathcal{Q}^M$ and $q = \ell$ for cycles, leading us to

$$P_c \approx \mathcal{Q}^{-M}. \tag{2.55}$$

When this result is compared against (2.53), we see that both expressions fall off exponentially with M, although the exponential factors are different. In actuality, rigorous analyses of this subject have shown that (2.53), or equivalently (2.55), overestimate the likelihood of autocatalysis as a function of M. The emergence of autocatalytic sets is more likely to scale with M as a power law, not exponentially as predicted here. Nevertheless, despite the fact that these results are highly idealized, they lend some credence to the notion that large reaction networks are more susceptible to the emergence of autocatalytic subset(s). The observed critical transition is also explainable if we use other approaches in lieu of random graphs, such as percolation theory.

Autocatalytic theory has come a long way since the 1970s and 1980s, when it first emerged. From a mathematical and logical standpoint, it has been placed on a rigorous footing, and the number of attendant theoretical studies has also grown impressively; the ongoing developments in this area are summarized in Hordijk and Steel (2017) Kauffman (2019), and Adamski et al. (2020). To offer one such example, Xavier et al. (2020) investigated the autocatalytic networks conserved across microbial metabolic pathways and deduced that they mostly comprise small-molecule catalysts. From an experimental standpoint, we already encountered a few examples in Section 2.5.2. The formose reaction, which produces a variety of sugars using formaldehyde as the starting point, is known to be autocatalytic in nature, and so is the oligomerization of HCN to yield its tetramer and other products. Sousa et al. (2015) analyzed the metabolic reaction network employed by *Escherichia coli* and concluded that much of it was autocatalytic. Semenov et al. (2016) demonstrated that thiols and amides, potentially important players in prebiotic chemistry, could be synthesized by means of an autocatalytic network.

Many of the empirical studies involving autocatalytic networks have, however, either focused on advanced biomolecules or examples from (in)organic chemistry whose relevance to abiogenesis remains questionable. By and large, it may be said that the theory of autocatalytic sets has been placed on a sound footing but necessitates further experimental research and modeling (under plausible conditions) in order to fully establish its efficacy.

2.6.2 The iron-sulfur world

Most of the aforementioned pathways for the origin of life are implicitly heterotrophic in nature—that is, they operate under the premise that life originated through a "primordial soup" comprising organic compounds abiotically synthesized from a variety of sources sketched in Section 2.3. This hypothesis was independently originated by A. I. Oparin and J. B. S. Haldane in the 1920s (Tirard 2017) and received further impetus from the Miller experiments and their successors from the 1950s onward. In contrast, life on Earth is based on carbon fixation (autotrophy)—namely, the conversion of inorganic carbon compounds into organic molecules. All autotrophs on our planet rely only on six major avenues for carbon fixation. Although we will tackle two of them hereafter, the reader may consult Keller et al. (2014) for an example of how the other pathways could date back to the prebiotic world.

In 1988, Günter Wächtershäuser came up with a novel approach for resolving the origin of life conundrum by positing that the process was chemoautotrophic—that is, it required neither a plentiful supply of organic compounds nor the availability of solar radiation.[19] Wächtershäuser proposed that autocatalytic reaction networks, corresponding to protometabolic cycles, evolved in environments that were rich in reducing agents (electron donors) and iron-sulfur compounds. It was the latter feature that gave this class of models their name: the iron-sulfur world. In what follows, we shall discuss the initial formulation of the iron-sulfur world—an elegant summation was presented in Wächtershäuser (1988, 1990)—but it must be

19. A subtle point that is often overlooked is that the demarcation between autotrophic and heterotrophic modes is meaningful only in the context of living organisms. When dealing with prelife—a topic studied by origin-of-life models—the terminology is confusing since it runs the risk of mixing and mashing up different concepts.

recognized that some of the basic premises of the original model have been altered or discarded over the years.

A perceptive insight ascribable to Wächtershäuser is that a reaction that is disfavored on thermodynamic grounds (endergonic) can nevertheless function when it's coupled to another reaction that is thermodynamically favorable (exergonic). The latter reaction, in the iron-sulfur world, arises from the oxidation of iron sulfide (FeS) to form pyrite (FeS_2),

$$FeS + H_2S \rightarrow FeS_2 + H_2, \qquad \Delta G^0 = -38.4\,\text{kJ/mol}, \tag{2.56}$$

where the superscript '0' signifies that the reaction(s) were conducted at a pH of 0 and temperature of 298 K. Clearly, this pH is exceptionally acidic, but fortunately ΔG^0 does not change appreciably until a pH of ~ 8 is reached. As this exergonic reaction leads to the production of pyrite, Wächtershäuser coined the evocative phrase "pyrite-pulled reactions" to describe the other endergonic reactions. It is well-known that many of the early steps in carbon fixation are endergonic in nature, with one such example being the reduction of CO_2 to yield formic acid in water,

$$CO_2 + H_2 \rightarrow HCOOH, \qquad \Delta G^0 = +30.2\,\text{kJ/mol}, \tag{2.57}$$

and if we combine the above two reactions, we see that the net reaction is exergonic in nature:

$$FeS + H_2S + CO_2 \rightarrow FeS_2 + HCOOH, \qquad \Delta G^0 = -8.2\,\text{kJ/mol} \tag{2.58}$$

We point out that the Gibbs free energies associated with these reactions are subject to some variability, as noted in the review by Cody (2004). Apart from thermodynamics, it is evident that kinetics ought to also play an important role, but this issue remains less well understood. Experiments suggest that coupled microorganisms in Earth's subsurface environments have the capacity to mediate the transformation of FeS and H_2S into FeS_2 (Thiel et al. 2019).

The initial formulation by Wächtershäuser hinged on a version of the reverse tricarboxylic acid (rTCA) cycle—also variously known as the reverse Krebs cycle and the reductive citric acid cycle—an autocatalytic cycle that constitutes one of the six carbon fixation pathways on Earth; we will return to the rTCA cycle in Section 2.6.3. The reason it has been christened a

modified rTCA cycle is because the metabolic intermediates are thioanalogs of the compounds prevalent in the conventional rTCA cycle. A thioanalog essentially refers to compounds in which one of the carbon-oxygen bonds is replaced by a carbon-sulfur bond. The following is one of the pyrite-pulled reactions in the thioanalog rTCA cycle.

$$R - COSH + CO_2 + FeS \rightarrow R - (C = O) - COOH + FeS_2 \quad (2.59)$$

In other words, by starting from a combination of thioacetic acid (i.e., a sulfur analog of acetic acid) and carbon dioxide, pyruvic acid has been synthesized, with FeS serving as the electron donor.

The reason why the iron-sulfur world was thus named goes beyond simple metabolism, because Wächtershäuser argued that it plays a key role in other biochemical reactions as well. Equally important, it provides the surface onto which the molecules are adsorbed. On account of this postulate, it is evident that pyrite operates as a selective force, because some molecules will experience preferential adsorption compared to others. It was also theorized to function as a catalyst for enhancing the rates of polymerization, akin to montmorillonite in Section 2.4.3. Unlike most conventional models of abiogenesis, either heterotrophic or autotrophic, the iron-sulfur world does not hinge on the formation of a lipid membrane.

We can now ask the question, What level of experimental evidence favors or disfavors the canonical iron-sulfur world? The formation of simple organic molecules such as methane thiol (CH_3SH) is indisputable, but the situation with respect to more complex molecules is less promising on the whole. A variety of organic compounds including acetic acid and certain amino acids have been synthesized in the appropriate settings, and it was shown that amino acids are capable of undergoing oligomerization to yield short peptides. Despite these successes, most of the reactions were based on substituting the more reactive carbon monoxide (CO) in place of CO_2, with the concentrations unrealistically high in some cases.

As there are multiple reactions involved, it does not seem likely that pyrite will serve as an effective catalyst (or "puller") for all endergonic reactions. Moreover, there are some generic difficulties with all metabolic cycles. First, as the number of steps increases, the final yield will be very low even if the yield per step is moderate; for instance, the thioanalog rTCA cycle has sixteen steps in total. Second, as the reaction is cyclic, the (low) yield of the final step will serve as the initial condition for the first step of the

next cycle, and so on. We must note that some of the original premises of the iron-sulfur world (e.g., the thioanalog rTCA cycle) were revised in the 2000s (Wächtershäuser 2007), but the difficulties raised here are potentially still applicable mutatis mutandis. Further critiques of the classical formulation of the iron-sulfur world on kinetic, thermodynamic, and philosophical grounds can be found in de Duve and Miller (1991).

In the iron-sulfur world, it is evident that sulfur plays an important role. Recent experiments suggest that sulfides of other metals (e.g., iron, copper, and lead) could complement FeS in terms of promoting carbon fixation; in particular, experiments by Kitadai et al. (2019) show that these metal sulfides could collectively elevate the yields of intermediaries in the reverse tricarboxylic acid cycle, a crucial metabolic network that is described shortly hereafter. At this juncture, it is important to appreciate the fact that iron-sulfur clusters [Fe-S] play a vital role in biology in the form of iron-sulfur proteins. The fact that these compounds play a critical role in metabolic pathways, perhaps most notably as cofactors (facilitating enzymatic activity), is well-known. As we remarked in Section 2.2.5, there are some tentative grounds for supposing that iron-sulfur clusters may have constituted one of the pillars of protometabolic networks that were not reliant on phosphates.

Apart from the iron-sulfur world, several other metabolic models likewise underscore the centrality of sulfur, some of which we shall encounter later. One protometabolic pathway that will not be addressed here concerns the "thioester world" proposed by Christian de Duve, who received the Nobel Prize in 1974. The thioester bond is formed via the dehydration condensation of a carboxylic acid (R−COOH) and a thiol ($R' - SH$), as depicted in the following reaction.

$$R - COOH + R' - SH \rightarrow R - S - CO - R' + H_2O \qquad (2.60)$$

This reaction is endergonic and therefore requires energy, whereas the opposite reaction (hydrolysis) is exergonic ($\Delta G^0 \approx -31$ kJ/mol). In this respect, it resembles ATP, and the amount of energy released or required is also similar.[20] Hence, de Duve proposed that peptide-like molecules called "multimers" may have been synthesized on early Earth, served as early

20. Aside from thioesters, ATP, and their derivatives, acetyl phosphate ($C_2H_5O_5P$) is considered a viable candidate for the primordial energy currency (Whicher et al. 2018).

catalysts, and participated in protometabolism. A crucial component of his hypothesis was that thioesters provided the requisite energy for the formation of multimers. We shall not go into further details in this book but will point the reader in the direction of de Duve (1991) instead. Computational analysis of biochemical reaction networks suggests that thioesters (in place of phosphates), perhaps in conjunction with a variant of the canonical rTCA cycle tackled below, were possibly the cornerstones of ancient metabolisms (Goldford et al. 2017, 2019). On the other hand, experiments indicate that thioesters and thioacetic acid are markedly unstable in hydrothermal environments and are thus unlikely to accumulate abiotically and reach desirably high concentrations (Chandru et al. 2016).

2.6.3 The rTCA cycle

We have seen that the early versions of the iron-sulfur world accorded a starring role to the reverse tricarboxylic acid (rTCA) cycle, although its subsequent formulations partly moved away from this theme. However, apart from the original iron-sulfur world hypothesis, many other studies have also suggested that the rTCA cycle, or one of its variants, comprised the first noteworthy protometabolic pathway on Earth (Hartman 1975). Harold Morowitz was one of the most perspicacious exponents of this viewpoint, and we direct the reader to his thoughtful analyses undertaken jointly with Eric Smith (E. Smith & Morowitz 2004, 2016).

The reactions and products of the rTCA cycle are depicted in Figure 2.10. One of the points that is patently evident from inspecting this figure is that the precursors of all classes of biomolecules discussed earlier can be synthesized from some of the rTCA intermediates. The five compounds—acetate, pyruvate, oxaloacetate, succinate, and α-ketoglutarate—have therefore been dubbed the *standard universal precursors* for biosynthesis. Thus, the appeal of the rTCA cycle in terms of providing an origination point for biosynthesis is apparent. The rTCA cycle is autocatalytic in nature, as it commences with citric acid and results in the formation of two molecules of citric acid. If all the reactions are balanced out, the net result is quite simple (Orgel 2008):

$$CO_2 + 4H_2 \rightarrow CH_3COOH + 2H_2O \tag{2.61}$$

However, in presenting this simplified scheme, we have not listed essential molecules such as ATP, acetyl-CoA (acetyl coenzyme A), and ferredoxin;

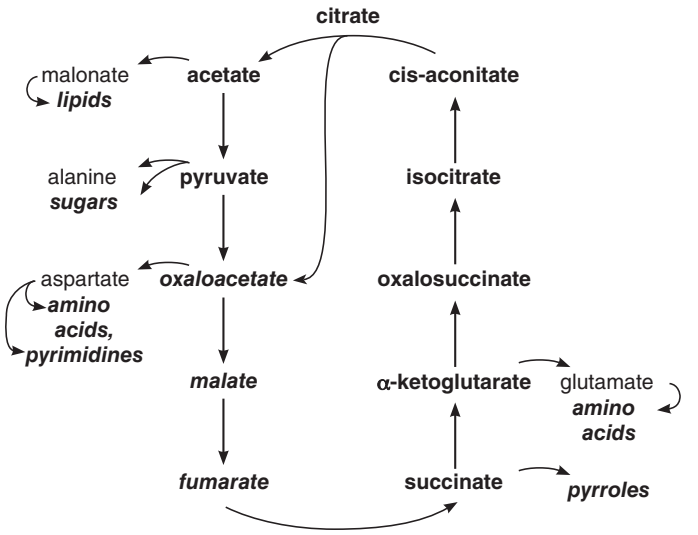


Figure 2.10 The basic products of the rTCA cycle (**bold**) and the major classes of biomolecules (***bold italic***) that can be synthesized from an appropriate starting point (normal). (© National Academy of Sciences. *Source:* Eric Smith and Harold J. Morowitz [2004], Universality in intermediary metabolism, *PNAS* 101[36]: 13168–13173, fig. 1.)

setting aside acetyl-CoA for the time being, we note that ferredoxin is an example of an iron-sulfur protein mentioned in Section 2.6.2. For a more accurate rendition of the final reaction, the reader may consult Fuchs (2011) and Berg (2011).

Apart from the fact that it produces an array of valuable precursors for further biosynthesis, the centrality of the rTCA pathway has also been proposed on phylogenetic and geochemical grounds. In the former context, the complete rTCA network has been documented in many clades of bacteria. If the reactions of the rTCA cycle are run in reverse, with a few subtle modifications, one ends up with the well-known Krebs cycle.[21] The Krebs cycle is ubiquitous in aerobic organisms, since it constitutes an essential component of cellular respiration. In addition, segments of the rTCA cycle have been documented in other prokaryotes. Thus, viewed collectively, there are cogent arguments in favor of the hypothesis that this metabolic pathway has ancient origins. The Krebs cycle and its inverted doppelganger, the

21. The Krebs cycle also goes by the names of the citric acid cycle and the tricarboxylic acid (TCA) cycle.

rTCA cycle, ought not be viewed in isolation for reasons explained shortly hereafter.

From a geochemical perspective, (2.61) illustrates that CO_2 and H_2 serve, effectively speaking, as the reactants. Detailed in situ experiments, in conjunction with laboratory analyses, have established that these two compounds were available in the vicinity of hydrothermal vents. One of the striking aspects about the rTCA cycle, which lends it further credibility, stems from the fact that the Gibbs free energy of formation (from CO_2 and H_2) is negative for all its intermediates at STP, and probably in hydrothermal conditions as well, suggesting that their formation ought to be favored on thermodynamic grounds. This feature is rare because it exhibits a marked contrast with other pathways, as they involve compounds with positive Gibbs free energies (e.g., formate). There are several other arguments having to do with the natural emergence of the rTCA cycle in a hydrothermal environment, which we shall not discuss here.

Apart from the rationales delineated above, there are reasons for contending that the rTCA cycle strikes a welcome balance between yielding the desired chemical complexity and avoiding the descent into chaos by virtue of its inherent selectivity; a detailed exposition of this topic is provided in Smith and Morowitz (2016). Despite these advantages, the rTCA cycle has been critiqued from different standpoints. First, the complete rTCA cycle is not documented in archaea, one of the two domains of prokaryotes (the other being bacteria). Hence, it has been averred that a sufficiently ancient metabolic pathway ought to be manifested in both domains. A potential explanation for this discrepancy is that the rTCA cycle, in an incomplete form, was interlaced with a different carbon fixation pathway at some primordial stage (a theme that we shall encounter soon), following which it underwent decoupling and gave rise to the current rTCA cycle.

The rTCA cycle could also suffer from a basic problem that we documented in Section 2.6.2—namely, the issue of low yields that are further amplified during the cyclic reactions. Apart from this general issue, other criticisms have been levelled against the rTCA cycle by Orgel (2008). It was argued, for example, that the completion of the cycle is contingent on a minimum of six different catalytic activities and that this scenario was unlikely. In this respect, the work by Muchowska et al. (2017) merits highlighting because it was demonstrated that six of the eleven reactions in the rTCA cycle were promoted by just two metal ions (Cr^{3+} and Zn^{2+}) and neutral iron, as opposed to five enzymes in extant biology. On a related

note, several recent intriguing studies dedicated to unraveling the role of nonenzymatic catalysts in facilitating the TCA cycle (i.e., the Krebs cycle) and other metabolic networks have sprung up (Muchowska et al. 2020), which are explicated below. One other striking facet of the rTCA cycle is that (α-ketoacid) analogs of the intermediates composing this pathway have been synthesized in spite of excluding catalysts and enzymes (Stubbs et al. 2020).

The antecedents of the modern TCA and rTCA cycles might have been interlinked. A detailed genomic and enzymatic investigation of the bacterium *Thermosulfidibacter takaii* by Nunoura et al. (2018) revealed that an ancient bidirectional TCA cycle may have enabled both autotrophic and heterotrophic lifestyles in early organisms and subsequently given birth to the current rTCA cycle; similar results were obtained for the bacterium *Desulfurella acetivorans* by Mall et al. (2018). Keller et al. (2017) showed that the synthesis of precursors of the TCA cycle was feasible by utilizing sulfate radicals as catalysts, while Springsteen et al. (2018) established that proto-metabolic analogs of the TCA cycle could be effectuated in the presence of ammonia and hydrogen peroxide. By carrying out a comprehensive analysis, Muchowska et al. (2019) were able to synthesize nine of the eleven intermediates comprising the TCA cycle using Fe^{2+} as a catalyst, while the inclusion of hydroxylamine (NH_2OH) and metallic iron yielded four proteinogenic amino acids.

Hence, the alleged constraints imposed by the necessity for multiple nonenzymatic catalysts in the rTCA cycle may not prove to be as detrimental as once presumed, although the geochemical availability of certain metals on early Earth is not well constrained. Another point, closely related to the first, is that unintended side reactions could disrupt the functioning of the cycle, although it seems plausible that the presence of nonenzymatic catalysts can mitigate this issue.

2.6.4 The acetyl coenzyme A pathway

The last metabolism network that merits explication is the acetyl coenzyme A (acetyl-CoA) pathway, which is also known as the Wood-Ljungdahl pathway. The majority of the carbon fixation routes, including the acetyl-CoA pathway, are built on coenzyme A. As one may guess from its name, coenzyme A is a cofactor, and the key point worth remembering is that it has an S-H bond, i.e., it is a thiol in chemical terms. We will use the notation

SHCoA for coenzyme A, where the "SH" denotes the thiol bond. Acetyl coenzyme A, on the other hand, is a thioester and is therefore characterized by the presence of an S-CO bond. The steps involved in the acetyl-CoA pathway are comparatively few, but the overall process is quite complex nonetheless. This presumably ancient pathway has been the subject of extensive reviews by Ragsdale and Pierce (2008), Fuchs (2011), and W. F. Martin (2020).

An important point deserving recognition is that, contra the rTCA cycle, the acetyl-CoA pathway is prevalent in both bacteria and archaea. The natural supposition would be that the same (or similar) pathway must materialize in both domains, but the respective architectures diverge in important respects while also sharing key commonalities. Instead of presenting all the intermediate products, we will briefly outline the net reactions for bacteria and archaea by adopting the analysis in M. J. Russell and Martin (2004). In the case of methanogens (archaea), the overall reaction is

$$4H_2 + 2CO_2 + \mathrm{SHCoA} \rightarrow CH_3COSCoA + 3H_2O, \tag{2.62}$$

where the first term on the right-hand side is acetyl-CoA. The Gibbs free energy at neutral pH is $\Delta G^0 \approx -59$ kJ/mol—that is, the reaction is exergonic in nature and can enable the synthesis of ATP. This reaction shares close similarities with (2.59), although there is no pyrite-pulling involved. In the above reaction, the reduction of CO_2 in the presence of a thiol takes place to yield a thioester bond with H_2 serving as the electron donor. In the case of acetogens (mostly bacteria), the overall reaction is easier to represent and is expressible as

$$4H_2 + 2CO_2 \rightarrow CH_3COOH + 2H_2O, \tag{2.63}$$

with a Gibbs free energy of $\Delta G^0 \approx -160.74$ kJ/mol at 343 K (70 °C). Although this reaction is highly exergonic, it still requires catalysts owing to the high activation energy required. Nevertheless, on thermodynamic grounds, organisms that fix carbon using the acetyl-CoA pathway have been evocatively described as being given "a free lunch that they are paid to eat" (Shock et al. 1998).

At this juncture, we wish to emphasize that proponents of the acetyl-CoA pathway do not necessarily claim that the first protometabolic networks were exactly identical to the acetyl-CoA pathway documented in

living organisms today. There are still a number of unresolved issues, most notably regarding the manifest differences in the pathways evinced by methanogens and acetogens. A cogent review of this topic, along with the hypotheses propounded to explain the empirical divergences, is delineated in Sojo et al. (2016). Now we shall take a closer look at some of the chief reasons advanced in support of the postulate that the acetyl-CoA pathway, or one of its variants, comprised the primordial metabolic pathway.

To begin with, we have already remarked that this pathway occurs in bacteria and archaea. As they are the earliest two domains of life, this datum lends credence to the hypothesis that the acetyl-CoA pathway is ancient. M. C. Weiss et al. (2016) undertook a very detailed phylogenetic analysis of millions of protein-coding genes in prokaryotic genomes. They concluded that the Last Universal Common Ancestor (LUCA) may have been thermophilic, may have been reliant on the acetyl-CoA pathway, and may have lived in a hydrothermal setting; an up-to-date summary of these findings is laid out in M. C. Weiss et al. (2018). Two caveats are worth highlighting here: (1) all the deductions of this study are not universally accepted (as of now), and (2) it is not feasible to trace the stages prior to the emergence of LUCA. The next point that bears mentioning is that the acetyl-CoA pathway is linear, as opposed to the rTCA cycle. Hence, it should not suffer from the potential issues related to declining yields sketched in Section 2.6.3. Furthermore, it has a fewer number of intermediate steps, thereby indicating it may be less susceptible to derailment. As we have seen, the overall reaction is exergonic, implying that any putative organisms can derive both their organic carbon and energy necessary for growth from the same pathway.

The last point in favor of the acetyl-CoA pathway is that most of the steps require (Fe,Ni)S proteins that closely resemble minerals that exist in hydrothermal vents, such as greigite (Fe_3S_4); one notable example is ferredoxin, which was encountered in Section 2.6.3. In addition, the reactants CO_2 and H_2 are abundantly available in these environments. Thus, from a geochemical perspective, most of the desired ingredients necessary for setting up the acetyl-CoA pathway already appear to exist. Varma et al. (2018) demonstrated that metals like iron, nickel, and cobalt reduce CO_2 to yield acetate and pyruvate—namely, the intermediary and final products of the acetyl-CoA pathway. Preiner et al. (2020b) experimentally corroborated the fixation of CO_2 (giving rise to formate and acetate, among others) by the minerals greigite, magnetite (Fe_3O_4), and awaruite

(Ni_3Fe) in alkaline environments at 100 °C. Hence, it does seem plausible that the acetyl-CoA pathway could have functioned in the absence of enzymes.

The criticisms leveled against the acetyl-CoA pathway relative to other carbon fixation pathways appear to be fewer in number. However, a couple of points deserve to be mentioned. Experiments have recently revealed that simple species containing the thioester functional group (e.g., thioacetic acid) are relatively susceptible to hydrolysis under hydrothermal conditions (Chandru et al. 2016). Hence, it is unclear whether thioesters (e.g., acetyl-CoA) can accumulate until sufficient concentrations are attained through abiotic channels. The other point, which is fairly generic, is that we are still some way off from definitively establishing, when viewed from an experimental standpoint, the synthesis of increasingly complex organic molecules eventually culminating in the production of the requisite biomolecules via this pathway.

Clearly, this is not as much of an issue for the rTCA cycle conceptually speaking, because we saw that it gives rise to the precursors of biomolecules. The natural question that emerges is whether the two pathways can be linked in some way. Several groups have indeed converged on this promising solution from multiple perspectives. As noted previously, the synthesis of acetate and pyruvate—core products of the acetyl-CoA pathway—has been accomplished in the laboratory. Figure 2.10 reveals that both these compounds are also an integral part of the rTCA cycle. Varma et al. (2018) drew on this basic observation as well as their empirical findings and prior theoretical studies (e.g., Braakman & Smith 2012; Camprubi et al. 2017) to posit a protometabolic network that merged elements of the acetyl-CoA pathway and the rTCA cycle, as depicted in Figure 2.11. Despite this noteworthy development, we caution that these experiments were typically reliant on high concentrations of metals (in neutral or ionic states) whose abundances are poorly constrained on Hadean Earth.

2.7 WHAT ARE THE PLAUSIBLE SITES FOR ABIOGENESIS?

Many different environments have been proposed as sites for the origin of life, each of which have their own distinct advantages and disadvantages (Camprubí et al. 2019). Owing to the multitude of candidates, we will not provide a comprehensive list here, opting instead to focus on

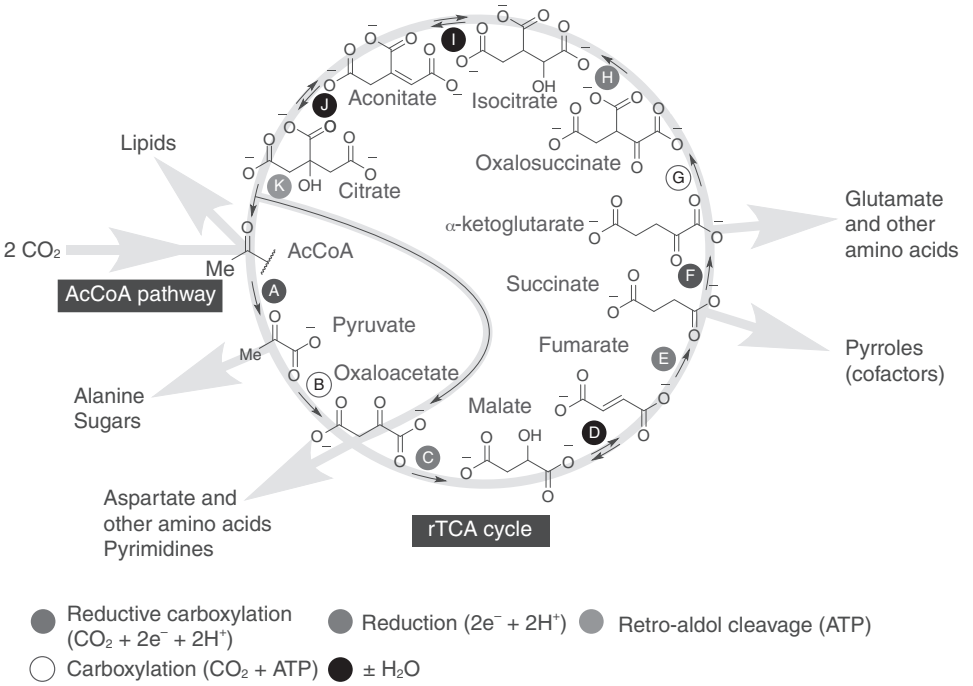


Figure 2.11 The acetyl-CoA pathway and the rTCA cycle have been juxtaposed to yield a hypothetical protometabolic network. Bear in mind that step *A* is common to both pathways. The shaded icons signify the different types of chemical mechanisms by which the intermediaries are synthesized. The biomolecules that can be produced from the appropriate precursors are also depicted: steps *A* and *F* involve reductive carboxylation; steps *B* and *G* involve carboxylation; steps *C*, *E*, and *H* involve reduction; steps *D*, *I*, and *J* involve addition or removal of H_2O; and step *K* involves a particular version of ATP-dependent cleavage. (© Macmillan Publishers Limited. *Source:* Sreejith J. Varma, Kamila B. Muchowska, Paul Chatelain, and Joseph Moran [2018], Native iron reduces CO_2 to intermediates and end-products of the acetyl-CoA pathway, *Nature Ecology and Evolution* 2: 1019–1024, fig. 5.)

a few possibilities. Thus, some of the intriguing microenvironments that we exclude are superheated cloud droplets in the upper reaches of the atmosphere (Woese 1979), aerosols (tiny droplets) at the ocean-atmosphere interface (E. C. Griffith et al. 2012),[22] pumice (volcanic rock) rafts in the oceans (Brasier et al. 2011), gel environments in aqueous settings (Trevors &

22. We will, however, briefly describe the reasons why aerosols are perceived as potential sites for abiogenesis in Section 5.7.1.3.

Pollack 2005), pores and fractures of rock deep beneath the Earth's surface, and sediments at the bottom of the ocean (Westall et al. 2018). We will also not tackle sea ice as a prospective setting for abiogenesis, as we have already gone over many of its advantages, such as eutectic freezing, stabilization of organics, and natural gradients in temperature and pH, to name a few.

Although identifying *all* of the necessary and sufficient conditions for the origin of life is not wholly feasible at this stage—one might, in fact, even declare it a fool's endeavor—some of the most plausible criteria are adumbrated below. Although this list resembles the one unveiled in Section 1.3, the enterprise has been pursued on account of its significance and to retain the modular nature of the book.

- While the availability of liquid water may seem obvious on account of water's centrality as a solvent for biochemistry, we have also seen that many organic compounds undergo rapid hydrolysis resulting in a conflict that has been termed the *water problem* or the *water paradox*. For this reason, a couple of environments discussed later assign a minimal (or nonexistent) presence to water. It must, however, be noted that once life-as-we-know-it has originated, the availability of water would serve as one of its essential requirements. It is due to this fact that many searches for life on exoplanets are centered on a "follow the water" approach.
- Life-as-we-know-it requires access to the standard bioessential elements, preferably in the form of water-soluble compounds: carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur. The availability of phosphorus in soluble form is a well-known conundrum (the phosphate problem) that was touched on earlier, but this difficulty is rendered irrelevant if prelife and the earliest lifeforms did not use phosphorus. In low-metallicity systems—that is, stellar environments wherein the abundances of elements aside from hydrogen or helium are much lower than the solar values—the availability of carbon and other bioessential elements may also comprise a limiting factor. Beyond these five elements, many of the transition metals play an important role in biological functions as catalysts.
- The recognition that life is a far-from-equilibrium phenomenon has been appreciated since Ludwig Boltzmann in the nineteenth century, as noted in Section 1.1. However, in order for

the processes associated with living systems to be driven into this state of disequilibrium, they must be embedded within a larger system that is dissipating disequilibrium.[23] To put it somewhat differently, many of the fundamental reactions responsible for the emergence and sustenance of life are endergonic (uphill) in nature, owing to which they must be coupled to reactions that are exergonic (downhill) (Branscomb et al. 2017). The reason we have embarked on this brief detour is because this criterion tends to be loosely classified under availability of energy sources and gradients for prebiotic synthesis.

- The appropriate range of temperatures is also very important. If the temperatures are too low, it might inhibit reactions from occurring at low rates, although this is not ineluctably the case. Similarly, if the temperatures are too high, the stability of most biomolecules is compromised and they undergo rapid decomposition. Hence, the temperature of the environment represents an important factor, and the same applies to other parameters such as the pressure, pH, and so on.
- The presence of minerals seems to be a vital, and probably essential, ingredient for the origin of life; although, it is conceivable that certain transition metals (in their pure form) could play a similar role. Apart from their well-known propensity to serve as catalysts, minerals can also assist in many other respects, such as stabilizing chemical compounds, as noted in Section 2.4.3.
- In most cases, the concentration of prebiotic compounds is too low to enable them to undergo polymerization (the concentration problem) and give rise to biopolymers like nucleic acids and proteins. Hence, it appears as though mechanisms for amplifying the concentrations of these compounds are an essential requirement.

Broadly speaking, in the event that we exclude microenvironments that lie at the ocean-atmosphere interface, most sites of abiogenesis that have been proposed are surface- and land-based. However, these hypotheses do

23. As this topic is somewhat technical, we recommend that the reader check out the analysis of dissipative structures by Kondepudi & Prigogine (2015).

not always articulate a crucial implicit assumption—namely, that landmasses were continuously present on Hadean Earth. While there is some tentative evidence from zircons that continental crust did exist during the Hadean era, it is very unclear what percentage of Earth's surface was covered by continents (Hawkesworth et al. 2017) and volcanic islands (Bada & Korenaga 2018); in addition, we have little information about its composition. If, as suggested by some authors, only a minimal fraction of the Earth's surface was composed of landmasses during the Hadean, this category of scenarios would necessitate reexamination.

Yet another factor that may adversely impact surface-based abiogenesis models stems from the much higher UV flux, due to the absence of ozone and a very active young Sun, that would have driven the photolysis of biomolecules. The high flux of large impactors during this period could have also posed difficulties to surface-based life on land. Hence, when we subsequently encounter surface- and land-based hypotheses for the origin of life, these factors should be duly appreciated.

2.7.1 Hydrothermal vents

Hydrothermal vents are prevalent near regions of volcanic activity on the seafloor. They represent openings in the rock from which hot water (via geothermal heating) emerges. In actuality, there are two broad different classes of hydrothermal vents (Baross & Hoffman 1985). The vents from the first category are called black smokers and seem relatively less pertinent insofar as abiogenesis is concerned, since the breakdown of biological molecules is elevated by their high temperatures (reaching up to 673 K). The second class of hydrothermal vents are of more interest to us. The first example in this category, the Lost City hydrothermal field (LCHF), was discovered in 2000. The LCHF consists of chimneys about 30 to 60 m in height and has a number of interesting properties. The hydrothermal fluid is characterized by its highly alkaline nature (pH of 9 to 11), lower temperatures (313 to 363 K), and high concentrations of dissolved H_2 and CH_4. The reader may consult M. J. Russell and Hall (1997), W. Martin et al. (2008), M. J. Russell et al. (2014), and Sojo et al. (2016) for in-depth analyses of hydrothermal vents and why they are perceived in some quarters as prime environments for the reification of abiogenesis. A general overview of an alkaline low-temperature hydrothermal vent and its primary geochemical attributes is presented in Figure 2.12.

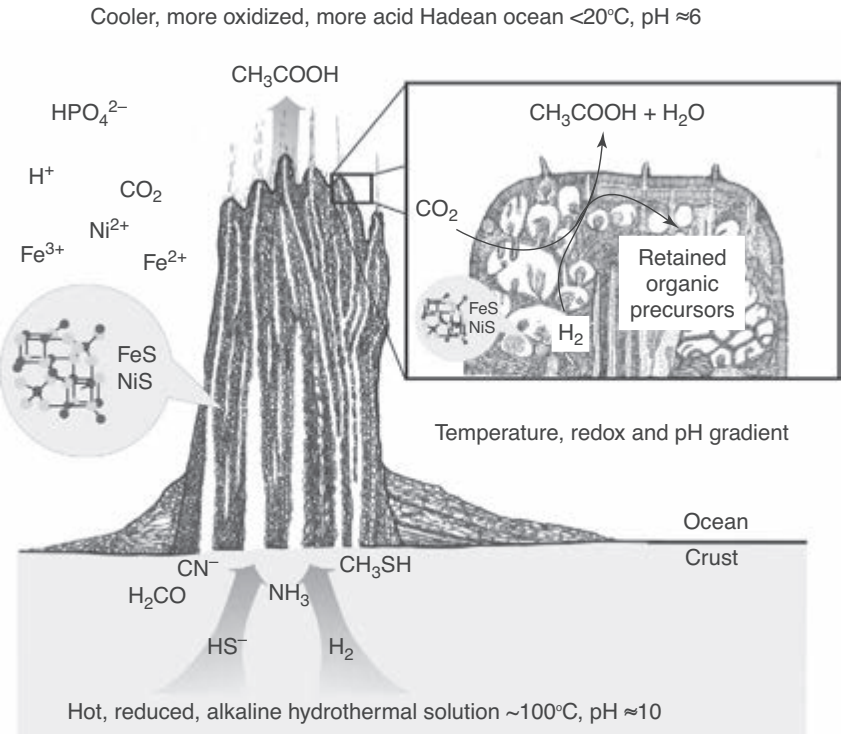


Figure 2.12 A schematic representation of a low-temperature alkaline hydrothermal vent. The environment contains metal sulfides (catalysts), carbon dioxide and hydrogen (reactants), in tandem with steep thermal, reduction potential (redox), and pH gradients. (© Elsevier. *Source:* Michael J. Russell and William Martin [2004], The rocky roots of the acetyl-CoA pathway, *Trends in Biochemical Sciences* 29[7]: 358–363, fig. 3.)

We encountered one of the foremost pros for hydrothermal vents in Section 2.6.4, where we delineated the chief arguments as to why the acetyl-CoA pathway is advantageous compared to other carbon-fixing pathways. Furthermore, it was noted that phylogenetic evidence favors the proposition that LUCA inhabited a hydrothermal setting and that its metabolism was closely linked to the acetyl-CoA pathway. The next aspect we need to evaluate is whether the requisite raw materials for the origin of life are available in sufficient abundance. We begin by observing that there exists ample evidence for iron-sulfur deposits in the vicinity of past and present hydrothermal vents; minerals containing other important metals like nickel and cobalt have also been documented. These compounds are important from the standpoint of serving as catalysts.

Of the panoply of minerals present on Hadean Earth, the whimsically (but aptly) named *green rust* has attracted interest of late (Russell 2018). Green rust, named on account of its color, signifies a category of chemical compounds comprising the two valence states of iron (Fe^{2+} and Fe^{3+}), the hydroxide anion (OH^-), and another negative ionic species. It has been proposed that green rust effectively functioned as an engine to maintain ionic disequilibria and power endergonic reactions. The mineral fougerite is an example of green rust, albeit an imperfect one because other metals occur in lieu of iron at some sites. However, it was argued that these very imperfections enhance the prospects for catalysis and bestow further chemical flexibility (Duval et al. 2019).

The next major ingredients are CO_2 and H_2. These compounds are important because several hypotheses regarding abiogenesis in hydrothermal vents are based on CO_2 and H_2 as the initial electron acceptor and donor, respectively, and invoke a mechanism known as flavin-based electron bifurcation (Buckel & Thauer 2018). Hadean Earth probably had a CO_2-rich atmosphere with a higher concentration of dissolved CO_2 in the oceans. It has been estimated that the Earth's oceans were conceivably mildly acidic (i.e., with a pH of ~ 6) and that the concentration of dissolved CO_2 was ~ 10 mM. Molecular hydrogen was produced through the serpentinization of the ocean's crust: this reaction, which essentially involves the oxidation of iron, was eloquently christened "life's mother engine" by M. J. Russell et al. (2013). The oceanic crust contains iron-magnesium silicate minerals (e.g., olivine) in which iron's oxidation state is +2. At high temperature and pressure, seawater circulating within the crust oxidizes the iron (to a state of +3) and is subsequently reduced to H_2.

The mechanics behind this exothermic reaction are rather complex, owing to which we shall present a simplified version below; for those who seek to learn more, the full details are delineated in Sleep et al. (2011).

$$3FeO + H_2O \rightarrow Fe_3O_4 + H_2 \tag{2.64}$$

On the basis of the evidence from current hydrothermal vents, it is suspected that the concentration of H_2 would have been ~ 10 mM. Apart from this source of H_2, radioactivity may have contributed to the generation of H_2 via the radiolysis of water, although the maximum concentration was possibly an order of magnitude lower. Thus, from the standpoint of CO_2 and H_2 availability, there are no major hurdles. In addition, we note

that other compounds like NH_3 and CH_4 have been documented in vent systems at concentrations of ~ 1 mM, but it is unclear whether they were synthesized by means of serpentinization or a different process; this caveat is expressly applicable to CH_4 because serpentinization under laboratory conditions has produced very low yields at times (Etiope & Whiticar 2019). Formate, an important prebiotic compound, has also been observed in the LCHF discharges at a concentration of ~ 0.1 mM.

The next major advantage associated with hydrothermal vents is a consequence of the fact that the Hadean ocean's pH was perhaps ~ 6, whereas that of the alkaline low-temperature vents is $\lesssim 11$. Thus, it is evident that there exists a pH gradient of Δ (pH) $\lesssim 5$ between the exterior and interior of the vent. From a biological standpoint, this fact is relevant because pH gradients are routinely used by prokaryotes for carbon fixation via the acetyl Co-A pathway; the mechanism is loosely akin to the generation of hydroelectric power through the controlled release of water from a dam. A unique feature of the pH gradients observed across hydrothermal vent membranes is that their magnitude and polarity—that is, higher pH inside (vents) than outside (ocean)—are both commensurate with the gradients associated with biological cells. In particular, the novel mechanism (chemiosmosis) underlying the synthesis of ATP in cells, which was first proposed by Peter Mitchell in 1961, entails the movement of protons across the membrane, quite reminiscent of water flowing across a turbine. The paradigmatic shift engendered by this discovery led to Mitchell being awarded the Nobel Prize in 1978.

Hence, this striking similarity has been invoked to suggest that the reduction of CO_2 with H_2 may have been facilitated by the readily accessible geochemical pH gradients in these environments. In the absence of enzymatic action, the routes by which this reduction occurred are unknown, but a number of hypotheses are being tested and evaluated. For example, a recent hypothesis is predicated on the datum that H_2 becomes more reducing in the alkaline conditions of these vents as per the Nernst equation; in a nutshell, this equation states that the reduction potential changes by approximately 0.059 V per unit pH. On the other hand, since the CO_2 in the ocean has a lower pH, it is more susceptible to being reduced. The reduction of CO_2 by H_2 to yield formate was successfully accomplished in a laboratory microfluidic reactor under the action of pH gradients (spanning several units) and metal sulfide minerals (Hudson et al. 2020), both of which are found at alkaline hydrothermal vents; this demonstration arguably constitutes a major development vis-à-vis abiogenesis at hydrothermal vents.

Although we have primarily emphasized the role of pH gradients here, it must be borne in mind that steep thermal and reduction potential gradients also exist in hydrothermal vent environments and could play analogous roles (Ooka et al. 2019).

Even though most hypotheses concerning abiogenesis in hydrothermal vents are usually classified as metabolism-first hypotheses, it must be recalled that the synthesis of many prebiotic molecules is feasible, and often thermodynamically favorable, at hydrothermal conditions. It is not surprising, therefore, that laboratory studies have yielded many of the standard amino acids, RNA nucleobases, and lipids. Recently, Ménez et al. (2018) analyzed rock samples from hydrothermal vents that purportedly confirmed the abiotic synthesis of the proteinogenic amino acid tryptophan at nanomolar concentrations, although this result has been subsequently disputed. In addition, the spontaneous assembly of short RNA oligomers as well as vesicles in controlled laboratory settings has been documented (Jordan et al. 2019). We may be predisposed to assume that these environments lack a suitable concentration mechanism, as they are situated within the ocean, but recall from Section 2.4.4 that porous media are effective in concentrating biomonomers, facilitating their polymerization, and potentially even favoring the selection of longer polymers.

Hence, on account of all these reasons, it is easy to see why a compelling case can be made for hydrothermal vents as the sites of abiogenesis. Lastly, one other rationale offered by its advocates merits a mention. If the precursor of either the rTCA or acetyl-CoA pathways (or variants thereof) arose in this environment, it would naturally ensure *continuity* between protometabolism mediated by geochemical constraints and the carbon fixation pathways in prokaryotes today (methanogens and acetogens). It has been argued that some of the other routes to abiogenesis proposed in the literature (e.g., the cyanosulfidic metabolism in Section 2.3.1) are fundamentally discontinuous with living organisms today. However, this apparent discrepancy ought not be regarded as being insurmountable—protometabolic networks might have functioned very differently from LUCA, as the latter only represents the most recent common ancestor of living organisms.

Yet several uncertainties still persist. First, the number of laboratory studies, especially when compared to Miller-Urey-type experiments, are relatively few. The biomolecules or their precursors that have been reported in these investigations were often synthesized from unrealistically high concentrations of reactants and under conditions that did not accurately simulate

hydrothermal vents. Another important point is that the thermal stability of biomolecules generally declines sharply with temperature, implying that they would be subject to fairly rapid decomposition. Furthermore, the abundance of important chemical species, such as reactive nitrogen, phosphates, and thioesters, is subject to much uncertainty. The efficacy of putative protomembranes to harness pH gradients and carry out the synthesis of prebiotic compounds was challenged by J. B. Jackson (2016), with a rebuttal tendered in Lane (2017a). Finally, we observe that the lifetime of low-temperature vents is $\sim 10^5$ years: a long timescale when viewed in absolute terms, but relatively short by geological standards. In a similar vein, the timescale over which a high pH gradient is maintained could be millions of years, and it may have therefore been necessary for life to have originated during this relatively short-lived period.

To summarize, a diverse array of positive features have been identified in connection with the origin of life in hydrothermal vents. On the other hand, a number of unknowns and potential issues also exist. Needless to say, a clear picture of this model's strengths and weaknesses is likely to emerge only after detailed experimental and theoretical studies are undertaken in the future.

2.7.2 Warm little ponds

The notion that life may have originated in warm little ponds (WLPs) has a storied history. It was first proposed by no less a personage than Charles Darwin himself, in a famous letter to his close friend Joseph Hooker—a well-known botanist in his own right—in 1871. A portion of his letter is reproduced below owing to its remarkably prescient nature.[24]

> It is often said that all the conditions for the first production of a living organism are now present, which could ever have been present. But if (and oh! what a big if!) we could conceive in some warm little pond, with all sorts of ammonia and phosphoric salts, light, heat, electricity, &c., present, that a proteine compound was chemically formed ready to undergo still more complex changes, at the present day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures were formed.

---

24. *To J. D. Hooker, 1 February [1871],* Darwin Correspondence Project, https://www.darwinproject.ac.uk/letter/DCP-LETT-7471.xml

Thus, Darwin not only identified a particular geological environment but also delineated some reasons why this setting would have been advantageous for the origin of life. Among other things, he also underscored the importance of free energy sources, nitrogen and phosphorus (ammonia and phosphates, respectively), in the context of life and proposed that protein-like compounds may have been the foundation of prelife. The phrase "warm little pond" is evidently rather broad in scope and consequently encompasses a wide range of environments, such as alkaline lakes (Toner & Catling 2019), which might have facilitated the emergence of cyanosulfidic metabolism (expounded on in Section 2.3.1). Hence, we will restrict ourselves to elucidating a single example below.

The premise that life originated at geothermal fields, also variously termed *hydrothermal fields* or *hydrothermal pools,* has attracted its share of supporters. Loosely speaking, geothermal fields are localized regions associated with volcanism; geysers and hot springs fall under this category. Notable locations where geothermal fields are prevalent include Yellowstone National Park and the Uzon caldera in the Kamchatka Peninsula. There are a number of reasons why these sites have been proposed as the sites of abiogenesis. Mulkidjanian et al. (2012) adopted the approach that cells maintain chemical (and biological) continuity—a "chemistry conservation principle," to be precise—allowing us to therefore trace the settings in which the earliest lifeforms evolved. From this supposition, it was inferred that the earliest habitats were characterized by a high K^+/Na^+ ratio that was consistent with the value observed in the cytoplasm of modern cells: high concentrations of zinc and manganese ions as well as phosphates. The first criterion was argued to be essential and used to eliminate marine environments, owing to their low K^+/Na^+ ratios. Mulkidjanian et al. came to the conclusion that only geothermal fields plausibly fulfilled all the criteria. However, we note that the chemistry conservation principle remains a point of contention, and both the postulate and its attendant conclusions have been questioned.

A different line of reasoning for geothermal fields was proposed by David Deamer and colleagues. A succinct summary of the central arguments can be found in Deamer & Georgiou (2015) and Damer & Deamer (2020). First, these environments are known to possess three distinct interfaces—to wit, atmosphere-water, atmosphere-mineral, and mineral-water. The importance of multiple interfaces stems from the fact that they are in a state of fluctuation as a result of precipitation and evaporation, thereby driving

wet-dry cycles and probably thermal and chemical gradients. We have already encountered the importance of all these factors. For instance, wet-dry cycles in conjunction with deliquescent (i.e., spontaneously accruing water vapor) minerals lining the pools facilitate mobilization and polymerization of ribonucleotides and promote the self-assembly of vesicles (Becker et al. 2018; Hud 2018; T. D. Campbell et al. 2019).

Of even higher importance, presumably, is the suite of recent laboratory experiments conducted by Becker et al. (2019) that culminated in the synthesis of all the RNA nucleosides and nucleotides from a mixture of small molecules, ribose, and phosphate-containing minerals via the sustained action of wet-dry cycles; the significance of this result is manifestly apparent in view of the difficulties surrounding RNA nucleotide synthesis, as explicated in Section 2.5.2. Another noteworthy point concerns the putative selection of those vesicles that are more robust (i.e., less susceptible to decomposition) during the wet-dry cycles, thereby suggesting that this mechanism may serve as an agent of selection. In addition, geothermal fields are believed to have lower concentrations of Ca^{2+} and Mg^{2+} relative to seawater, which was identified as being potentially important because these divalent cations can inhibit the self-assembly of protocells. On this note, the assembly of certain amphiphilic compounds into vesicles has been demonstrated, ceteris paribus, in water samples from geothermal fields but *not* in sea water (Milshteyn et al. 2018).

An intriguing, albeit ambiguous, recent discovery on Mars merits attention before moving forward. The Mars Exploration Rover *Spirit* came across millimeter-sized opaline silica (amorphous hydrated SiO_2) structures in the Gusev crater that exhibit remarkable similarities, in terms of their infrared spectral features, with silica deposits at hot springs and geysers in El Tatio, Chile, and Utah, USA (Ruff et al. 2020). Furthermore, the nodular silica structures from El Tatio that display the greatest resemblance to their Martian counterparts were formed through a combination of abiotic and biological processes, thus raising the tantalizing possibility that the Martian silica deposits might likewise have a biogenic origin (Ruff & Farmer 2016). On a related note, the earliest evidence for land-based life in terrestrial hot springs seemingly dates back to $\sim$3.5 Ga, on the basis of analysis of purported stromatolites and microbial "fabrics" from the Dresser Formation of the Pilbara Craton in Western Australia (Djokic et al. 2017).

Another point, which has hitherto not been sufficiently appreciated, is that geothermal fields may share connections with volcanic activity. Earlier,

we saw that the latter plays an important role in supplying water-soluble phosphorus compounds (Section 2.2.5) and in generating carbonyl sulfide, a gas that facilitates polymerization (Section 2.4.5). It is conceivable, therefore, that these advantages would be inherited by geothermal fields located in the vicinity of active volcanoes. The access to UV radiation is an example of a mixed benefit—on the one hand, it could function as an efficient source in driving the prebiotic synthesis of organic compounds,[25] but on the other hand, it promotes the photodissociation of the same compounds. Patel et al. (2015), in their discussion of cyanosulfidic protometabolism (see Section 2.3.1), have also proposed a scenario that involves a network of streams and pools and UV light. The basic idea is that the sequential addition of chemical compounds in the laboratory would be manifested in this setting when streams with different reactants and reaction histories deposit their products sequentially within pools. While it is undeniable that the delivery of compounds displays a temporal ordering, manifold combinations chosen at random are not guaranteed to facilitate the synthesis of the precursors of proteins, nucleic acids, and lipids, unlike the outcomes in laboratory studies.

Several generic negative points are associated with warm little ponds. To begin with, one of the chief disadvantages in comparison to hydrothermal vents is that pH values are on the lower end of the spectrum and could thus give rise to enhanced hydrolysis of biopolymer bonds; the prospects for using a pH gradient in protometabolic pathways are much lowered accordingly. Second, the water paradox encountered previously is still operational. This issue may be mitigated to some degree in the event that prebiotic reactions occurred in pores; another advantage is that thermal gradients, if present, are theoretically capable of facilitating the concentration and polymerization of biologically relevant organic molecules. Finally, these pools are shallow compared to oceans. This basic fact increases the likelihood of organics (e.g., nucleobases) being depleted due to seepage of the aqueous solution through the underlying base.

By drawing on empirical data, Pearce et al. (2017) proposed that the seepage rate on early Earth was ~ 1 m/yr. Hence, most of the organic compounds in warm little ponds would be lost over a span of ~ 10 years since $e^{-10} \approx 4.5 \times 10^{-5}$, provided that there were no continuous sources

25. In particular, Rimmer and Shorttle (2019) have recently demonstrated via detailed numerical models that important feedstock molecules (e.g., HCN) are possibly formed at high (~ 0.001 to 1 M) concentrations in surface hydrothermal environments.

replenishing them; we have chosen an *e*-folding timescale of about 1 yr for a shallow meter-sized pool. To tackle this issue more accurately, let us suppose that the following simplifications can be made, although they represent idealizations. First, we ignore the roles of precipitation and evaporation in modifying the concentrations of feedstock molecules like HCN. Second, we assume that the seepage rate is constant; in reality, the base may get saturated and the seepage would come to a halt. Third, although the chemical compounds are being constantly depleted via seepage, we operate under the premise that the pond is replenished by external water sources and its depth $\mathcal{H}_w$ is therefore roughly constant. In this event, choosing HCN as the chemical compound of interest, (2.11) must be modified as follows:

$$\frac{d\mathcal{C}_{HCN}}{dt} = \mathcal{A}_w \mathcal{S}_{HCN} - \frac{\mathcal{C}_{HCN}}{\tau_h} - \frac{\dot{\mathcal{Z}}\mathcal{C}_{HCN}}{\mathcal{H}_w}, \tag{2.65}$$

where the last term on the right-hand side quantifies the loss due to seepage, with $\dot{\mathcal{Z}}$ denoting the seepage rate (in m/yr). As before, we solve for the steady-state value of $\mathcal{C}_{HCN}$ by setting the right-hand side to zero and then determine the concentration $\phi_{HCN} = \mathcal{C}_{HCN}/(\rho_w \mathcal{A}_w \mathcal{H}_w)$. If we specialize to the case of shallow ponds at moderate temperatures, it can be verified easily that $\mathcal{H}_w/\tau_h \ll \dot{\mathcal{Z}}$. In this regime, we end up with

$$\phi_{HCN} = 10^{-9}\,\mathrm{M}\left(\frac{\mathcal{S}_{HCN}}{10^{-6}\,\mathrm{mol\,m^{-2}\,yr^{-1}}}\right)\left(\frac{\dot{\mathcal{Z}}}{1\,\mathrm{m/yr}}\right)^{-1}. \tag{2.66}$$

The revised estimate derived above may be used in Section 2.3 instead of (2.12), but the latter is more conservative (and arguably robust) since what we know about the actual seepage dynamics of warm little ponds during the Hadean era is minimal. Nevertheless, if we do use (2.66), it becomes apparent that maintaining high concentrations of HCN, or more complex organics, over long timescales is difficult and independent of the depth for shallow water bodies.

2.7.3 Beaches

Beaches have often been viewed as promising environments for the origin of life, as reviewed in Lingam and Loeb (2018b). We begin by noting that they possess a variety of natural interfaces—to wit, water-air, land-water, and air-land. The existence of these interfaces enables the natural sustenance

of gradients that are widely considered beneficial and probably necessary for life's emergence. The action of tides would have powered natural wet-dry cycles, whose importance is evident from the preceding sections. Another unusual gradient associated with beaches is that freshwater, either in the form of groundwater or carried by rivers and streams, would interact with the saline waters of the oceans. This feature is anticipated to set up vertical and / or horizontal gradients in salinity (and probably pH). In addition, gradients in moisture and sunlight are also present at such locations. Apart from their putative role in driving disequilibria, the aforementioned gradients would facilitate fluid circulation and the exchange of materials as well as catalysis.

The next facet worth highlighting concerns the plentiful availability of minerals in tidal environments, if modern Earth is anything to go by. The evaporation of seawater would have left behind an array of minerals, especially since there are reasons for supposing that the oceans had higher salinity on early Earth. In turn, the ready access to minerals would provide a variety of potential catalysts for prebiotic reactions. Moreover, there are some grounds for believing that borate minerals, capable of stabilizing ribose (Section 2.5.2), could have accumulated at beaches, although their abundance remains an open question because of our limited understanding of Hadean geology, especially insofar as plate tectonics is concerned.

The last salient point is that radioactive minerals (e.g., monazite and uraninite) may have accumulated at beaches through a physical mechanism known as hydrodynamic sorting. This phenomenon is important in two respects. First, the decay of radioactive elements can enable prebiotic synthesis and also give rise to the formation of free radicals that serve as catalysts. Second, as we have seen in Section 2.3.4, radioactivity provides an excellent avenue for producing significant quantities of formamide. The importance of formamide stems from its dual ability to serve as a feedstock molecule and a prebiotic solvent. Thus, beaches open up the possibility that life might have originated in a solvent other than water and provide a plausible route by which this solvent was formed. Figure 2.13 is a conceptual illustration of the possible synthesis of formamide.

Some of the potential issues associated with this environment are common to other settings as well, owing to which we shall desist from elaborating on them once again. Instead, we will highlight one barrier that is unique to tide-mediated environments like beaches. After the presumed impact with Theia that led to the Moon's formation, the initial Moon-Earth

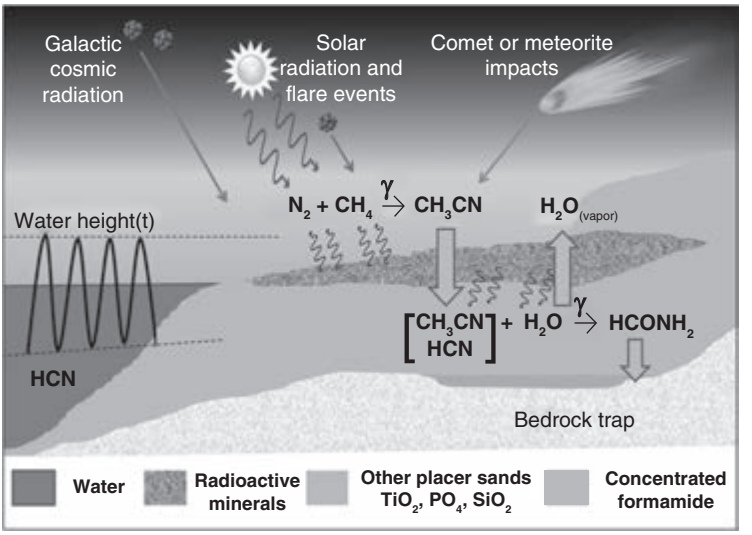


Figure 2.13 How formamide ($HCONH_2$) may have been synthesized from precursors like hydrogen cyanide (HCN) and acetonitrile (CH_3CN), which were in turn produced by a wide array of energy sources and prebiotic pathways. (CC-BY 4.0. *Source:* Zachary R. Adam, Yayoi Hongo, H. James Cleaves II, Ruiqin Yi, Albert C. Fahrenbach, Isao Yoda, and Masashi Aono [2018], Estimating the capacity for production of formamide by radioactive minerals on the prebiotic Earth, *Scientific Reports* 8: 265, fig. 4.)

distance was conceivably $a_{\text{moon}} \sim 4R_\oplus$, as opposed to today's $a_{\text{moon}} \sim 60R_\oplus$. The height of the tide (H_t) is given by

$$H_t \approx \frac{15}{8} \frac{M_{\text{moon}}}{M_\oplus} \frac{R_\oplus^4}{a_{\text{moon}}^3}, \tag{2.67}$$

where $M_{\text{moon}} \approx 0.01 M_\oplus$ is the mass of the Moon, and a_{moon} is the Moon–Earth distance.[26] The key thing to note here is the strong dependence of H_t on a_{moon}. In the immediate aftermath of the Moon's formation, the tides on Earth would have been around 3500 times higher than today—that is, we obtain $H_t \sim 2.3$ km from the above formula. Since we are fairly certain that life did originate after the Moon formed, (2.67) reveals that the tidal amplitudes were higher, perhaps much more so, than today. Hence, the strength

26. For the sake of simplicity, the above formula neglects a second-degree correction associated with the rigidity of the planet.

of the lunar tides might have influenced the likelihood of abiogenesis on beaches. If, for instance, the tidal forces were stronger by a factor of 10, it may not have posed problems, but a factor of 100 could have done so. Looking beyond the Earth, planets orbiting low-mass stars, like those in the recently discovered TRAPPIST-1 planetary system, are possibly subjected to tidal forces that are orders of magnitude greater than those experienced by the Earth.

Apart from beaches endowed with radioactive minerals, it is also conceivable that natural nuclear reactors were the sites of abiogenesis as they would have promoted even faster synthesis of formamide and other organic compounds. The positives must be counterbalanced by the fact that we have no knowledge of how common these reactors were, although it must be acknowledged that the inventory of radioactive minerals was higher than today (albeit not necessarily at the surface).

2.7.4 Semiarid environments: Intermountain dry valleys

In order to surmount some of the major paradoxes associated with the emergence of RNA and its precursors (Benner 2014), such as the asphalt problem (formation of organic "tar") and the water problem, Steven Benner and colleagues proposed that the formation of RNA occurred in dry intermountain valleys (Benner et al. 2012) or semiarid environments with an intermittent supply of water (Benner et al. 2019); the overall synthesis takes place in a discontinuous fashion. For the sake of brevity, we shall delve into the details of the former, but many of the basic ideas are readily portable to the latter.

Figure 2.14 illustrates the various steps involved in the formation of oligonucleotides by starting from a mixture of N_2, H_2O, and CO_2 as well as some CH_4. The first step entails the production of the basic feedstock molecules HCN and formaldehyde via electrical discharges or some other route. These two species accumulate in an aquifer by means of precipitation. The pH of the aquifer is alkaline (around 10 to 11) and it contains reducing agents as a result of serpentinization reactions involving olivine and other minerals (see Section 2.7.1). The aquifer also contains borate minerals that stabilize ribose (Kim et al. 2016), after its synthesis by way of the formose reaction involving formaldehyde. HCN undergoes hydrolysis in the aquifer to yield formamide, whose importance we specified in Section 2.3.4. Ammonium formate (NH_4HCO_2) is also present in this aqueous solution by virtue of the hydrolysis of formamide.

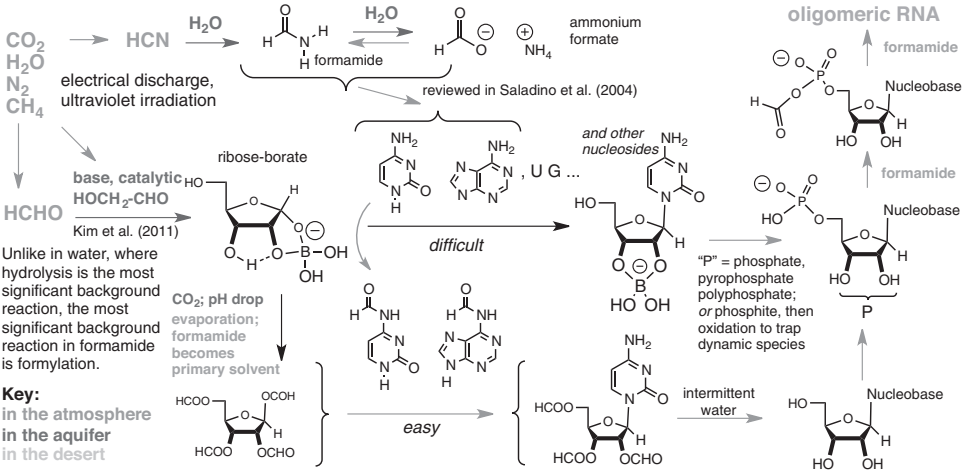


Figure 2.14 How a mixture of N_2, H_2O, and CO_2 eventually leads to the formation of nucleobases and sugars, nucleosides, and finally the nucleotides and oligonucleotides of RNA. (© American Chemical Society. *Source:* Steven A. Benner, Hyo-Joong Kim, and Matthew A. Carrigan [2012], Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA, *Accounts of Chemical Research* 45[12]: 2025–2034, fig. 5.)

The stage is set for the last environment, namely, the dry intermountain valley (or other semiarid locales). The valley is presumed to receive runoff from the aquifer, and the pH is ~6 as a consequence of buffering by atmospheric CO_2. In this setting, formamide serves as the precursor for producing nucleobases, and borate-ribose is dissociated to release ribose. The nucleobases are assembled together with ribose to yield ribonucleosides in the presence of formamide as a solvent, although this reaction can occur (albeit with much difficulty) even in the absence of formamide. Owing to the presence of reducing agents (supplied from serpentinization), phosphorus is available as soluble phosphite, thereby circumventing the phosphate problem. Finally, the phosphorylation of nucleosides leads to the synthesis of ribonucleotides that undergo subsequent oligomerization to yield RNA.

Several unknowns are associated with this proposal, among which some of them apply equally to other candidates as well. For example, many pathways culminating in the production of the nucleobases from HCN or formamide typically rely on fairly high concentrations of these compounds, and it remains unclear whether this requirement is realizable. With regard to the above mechanism, however, a few specific factors stand out.

The availability of borate and molybdate minerals constitutes an essential requirement. In actuality, we do not have an unequivocal understanding of the mineral inventory of the Hadean environment, nor even the area of exposed semiarid landmasses in light of the substantial uncertainties surrounding the growth of continental crust over time. Next, formamide plays a variety of roles in this hypothesis, ranging from feedstock molecule to solvent. In later iterations, however, viable alternatives to formamide were identified: for instance, in the reaction network of Benner et al. (2019), urea comprises one of the key ingredients. In all cases, the abundance of biomolecular building blocks in semiarid settings remains unresolved.

2.7.5 The planet as a global reactor

> All the world's a stage,
> And all the men and women merely players.
>
> —William Shakespeare, *As You Like It*

The preceding discussion illustrated that different environments have their own unique strengths and weaknesses. Many of them offer compelling reasons why they may have served as the sites of abiogenesis, but each comes with its own share of difficulties, perhaps even insurmountable ones. We might then ask ourselves whether some of these barriers could be overcome by viewing the myriad environments as the components of a gigantic chemical network. To paraphrase Shakespeare, perhaps all the world was a chemical reactor and all its environments were players in this planetary theater.

A meticulous discussion of how the origin of life is best envisioned as a planetary process—namely, one that is intimately connected to the flow of chemicals and energy on the planet—can be found in Stüeken et al. (2013), E. Smith and Morowitz (2016), and Sasselov et al. (2020). The basic idea is that the multiple environments of Hadean Earth would have possessed their own diverse array of energy sources and gradients, supplies of prebiotic compounds, catalysts, and mechanisms for their concentration. Some of these environments would have been either purely aerial, aquatic, or terrestrial in nature, with others lying at the interfaces of land, water, and air. Clearly, in order for these settings to interlink with one another, the presence of (sub)surface circulation mechanisms is of paramount importance. On Earth, some of the routes include atmospheric and ocean circulation, the

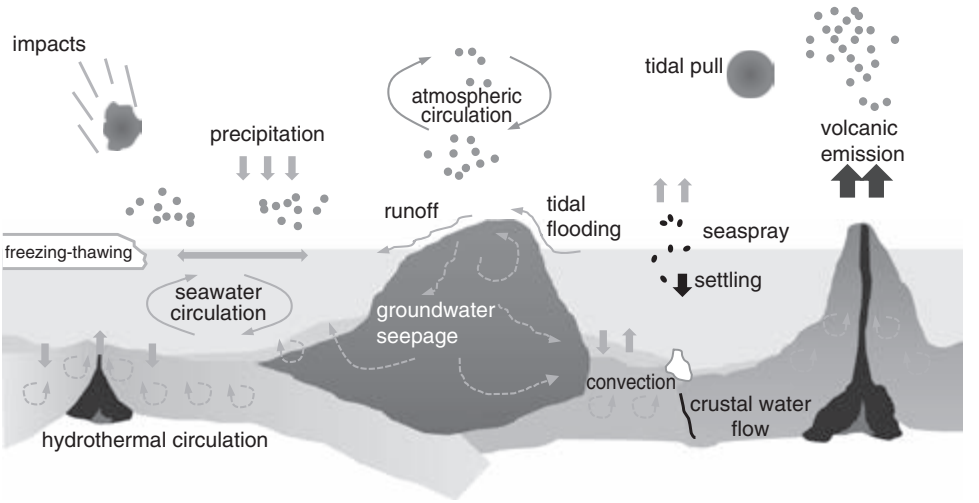


Figure 2.15 The various geochemical environments that could have served as the sites of abiogenesis, along with the multiple transport systems responsible for the circulation of prebiotic compounds. (© Blackwell Publishing Ltd. *Source:* E. E. Stüeken, R. E. Anderson, J. S. Bowman, W. J. Brazelton, J. Colangelo-Lillis, A. D. Goldman, S. M. Som, and J. A. Baross [2013], Did life originate from a global chemical reactor? *Geobiology* 11[2]: 101–126, fig. 5.)

runoff from rivers to oceans, volcanism, and asteroid and comet impacts, to name a few. Figure 2.15 presents the different environments and circulation systems.

If one subscribes to this approach, it becomes apparent that the diversity of local environments and transport processes linking them plays a crucial role in shaping the likelihood of life originating on a given planet or moon. How can we quantify this diversity in a heuristic fashion? It might be instructive to draw on the field of biological diversity and its attendant indices, with the proviso understanding that this is *not* an exact one-to-one mapping. For the sake of simplicity, we will first set aside variables such as catalysts and circulation systems, because they can be easily incorporated into the formalism without major loss of generality.

Let us denote the set of all environments that exist prior to abiogenesis on a planet by $\mathbb{S}$. The total number of distinct environments is N_e, whereas the total number of sites belonging to a particular environment j (e.g., number of beaches) is N_j. The total number of sites for abiogenesis in the planet (N_t) is given by $\sum_{j=1}^{N_e} N_i \equiv N_t$. The first metric that comes to mind is to

directly count the total number of different environments that exist, which amounts to calculating N_e. In ecology, this metric is known as the *species richness* and represents the number of species in a given sample. In order to illustrate how the species richness can be utilized, let us contrast the Earth against an aquatic world with only oceans (and ice) on the surface. Our planet would ostensibly possess the same set of environments as this world, except for the fact that additional habitats on land are accessible. We have already encountered possibilities like geothermal fields, intermountain valleys, and beaches in this regard. Thus, it might be plausible that the Earth's value of N_e is about an order of magnitude higher than ocean worlds. In turn, this may imply that the likelihood of life originating on our planet is higher compared to aquatic worlds, provided that all other factors are held fixed.

However, this metric does not tell us the whole story. Consider, for example, the following two cases: (1) ninety-eight sites of environment A and two sites of environment B and (2) fifty sites each of environments A and B. Intuitively speaking, we might be inclined to believe that (2) represents a more promising planet for abiogenesis compared to (1), despite the fact that they have the same value of $N_e = 2$; needless to say, this statement is highly idealized and should not be taken literally. We introduce the notation $P_j = N_j/N_t$, which represents the proportional abundance of environment j. One of the most well-known diversity measures is the Shannon index (S_p),

$$S_p = -\sum_{i=1}^{N_e} P_j \ln P_j, \tag{2.68}$$

and it is easy to verify that $S_p \approx 0.1$ for case (1), whereas $S_p \approx 0.7$ for case (2). Thus, we see that this metric reflects our earlier intuition that case (2) was more diverse than case (1). There are a number of other diversity measures in the literature, with one such metric being the Simpson index (D_p), also known as the Herfindahl index in economics, defined as follows:

$$D_p = \sum_{i=1}^{N_e} P_j^2. \tag{2.69}$$

An important point regarding D_p is that it increases when the diversity decreases. Hence, to reverse this trend, some widely used versions of Simpson's index include $1 - D_p$, $-\ln D_p$, and $1/D_p$. Since the last measure among this trio is arguably the most common, we will therefore utilize

$Q_p = 1/D_p$. For cases (1) and (2), we find $Q_p \approx 1$ and $Q_p = 2$, respectively, once again affirming the fact that (2) is more diverse than (1). A consummate exposition and analysis of the various metrics used for measuring biological diversity can be found in Magurran (2004).

Despite the advantages associated with this global picture in addition to its conceptual appeal, there is a potentially critical stumbling block. In the scenario where life arose in a single environment, all of the requisite ingredients (e.g., catalysts and energy sources) would be spatially localized, thereby increasing the efficacy of chemical reactions. In contrast, if reactants, products, and perhaps catalysts have to be assembled and transported across greater distances, they are more susceptible to both dilution and degradation. Thus, a global network may confer the advantage of chemical diversity on the one hand while reducing the effectiveness of prebiotic synthesis on the other hand.

A final point is that life must have originated in a *particular environment,* unless one subscribes to the very unlikely outcome in which life emerged de novo in more than one environment at the same time. Hence, even though the global reactor could have been responsible for the supply of chemical compounds to sundry environments, only one of them would have served as the actual setting for abiogenesis. Most of the candidates for the origin of life we have discussed hitherto, barring hydrothermal vents to an extent, are reliant on the external delivery of feedstock molecules and consequently reflect the notion that different environments were coupled to one another.

2.8 MATHEMATICAL MODELS RELATING TO THE ORIGIN OF LIFE

It must be emphasized at the outset that reading the contents of this section is optional. Hence, readers may skip directly to the chapter conclusion (Section 2.9) if they so desire. The primary aim behind collating together a handful of disparate mathematical models is to illustrate how they could shed some light on crucial subjects associated with the origin of life such as replication, natural selection, (genetic) information, and homochirality.

2.8.1 The transition from prelife to life

How exactly does one distinguish prelife from life? This is clearly a loaded question, as it goes to the very root of defining *life*. The latter, as we saw in Chapter 1, is a contentious issue with no clear answer. However, let

us focus only on a very narrow, but important, aspect of life—namely, its ability to replicate using templates; in contrast, we will suppose that prelife is not endowed with this feature, although it may be capable of mutation and selection. Our discussion of this topic mirrors the analysis undertaken by M. A. Nowak and Ohtsuki (2008). Numerous studies along these lines have been pursued, as reviewed in Takeuchi and Hogeweg (2012) and Menor-Salván (2018), but we shall work with the following model due to its relative simplicity.

Let us suppose that there exist only two different building blocks (i.e., monomers) 0 and 1 and their activated versions $0^\star$ and $1^\star$; it is the latter that possess the requisite energy, broadly speaking, to take part in polymerization. We introduce the notation j for a particular sequence, and polymerization occurs when $0^\star$ or $1^\star$ are added to the right end of the sequence j as follows:

$$\begin{aligned} j + 0^\star &\to j0 \\ j + 1^\star &\to j1 \end{aligned} \tag{2.70}$$

Thus, every sequence j has a precursor j', formed by eliminating the last "bit" in the sequence, and two successors $j0$ and $j1$, as seen from (2.70). It is helpful to consider an explicit example: j is the sequence 0111, from which we find j' is 011 and the two successors are 01110 and 01111 ($j0$ and $j1$, respectively). The rate of production of j from j' is denoted by r_j, while the rates of formation of $j0$ and $j1$ from j are r_{j0} and r_{j1}, respectively. Lastly, we assume that all sequences undergo decay at a constant rate k_d. The abundance of sequence j is given by x_j, and the time evolution of this abundance is given by

$$\frac{dx_j}{dt} = r_j x_{j'} - \left(k_d + r_{j0} + r_{j1}\right) x_j. \tag{2.71}$$

We can solve for the equilibrium by setting the right-hand side to zero, which yields x_j as a function of $x_{j'}$. If this equation is repeated recursively, we end up with

$$x_j = b_j b_{j'} b_{j''} \ldots b_\sigma, \tag{2.72}$$

where σ is the monomer (0 or 1) that initiated this sequence. Here, we have introduced the auxiliary notation $b_j = r_j / \left(k_d + r_{j0} + r_{j1}\right)$. It is helpful to consider the symmetric case $r_0 = r_1 = \alpha_0/2$ and $r_j = r_0$ for all other sequences

j; here, α_0 and r_0 represent fiducial values that are freely specifiable. It is left as an easy exercise for the reader to verify that

$$x_n = \frac{\alpha_0}{2r_0}\left(\frac{r_0}{k_d + 2r_0}\right)^n, \tag{2.73}$$

with x_n denoting the abundance of a given sequence with length n. The total abundance of *all* sequences of length n is merely $X_n = 2^n x_n$. Using these relations, the average sequence length turns out to be

$$\bar{n} = \frac{\sum_{n=1}^{\infty} nX_n}{\sum_{n=1}^{\infty} X_n} = 1 + \frac{2r_0}{k_d}, \tag{2.74}$$

which is independent of the formation rates of the monomers.

Hitherto, we have dealt with systems that do not have the capacity for replication. Hence, we modify the dynamical equations for prelife by including an additional term to account for this new ability:

$$\frac{dx_j}{dt} = r_j x_{j'} - x_j\left(k_d + r_{j0} + r_{j1}\right) + \theta_0 x_j\left(f_j - \phi\right) \tag{2.75}$$

One of the key variables in the above expression is the fitness f_j, i.e., the rate at which the sequence j is being replicated. The quantity ϕ represents an extra decay rate and is introduced for the purposes of balancing the extra sequences being produced by replication. In other words, it can be determined from $\sum_i x_i\left(f_i - \phi\right) = 0$, consequently leading us to $\phi = \sum_i f_i x_i / \sum_i x_i \;\forall\, i$. It is therefore the average fitness of the population. The quantity θ_0 is very important because it quantifies whether prelife dominates over life or vice versa. If we retain only the last term on the right-hand side of (2.75), it transforms into a special case of the pivotal quasispecies equation introduced by Manfred Eigen (who received the Nobel Prize in 1967) and Peter Schuster in the 1970s. The quasispecies model has proven to be very successful as a conceptual model for describing how Darwinian evolution acts on self-replicating "species."[27] The conceptual and theoretical details behind this approach can be found in the classic treatise by Eigen and Schuster (1979).

---

27. As per certain theoretical models, the emergence of such self-reproducing entities and self-organization has been argued to be a generic property of evolutionary dynamics (Fontana & Buss 1994).

The chief quantity of interest ($\mathcal{G}_j$) corresponds to

$$\mathcal{G}_j = \theta_0 \left(f_j - \phi\right) - \left(k_d + r_{j0} + r_{j1}\right). \tag{2.76}$$

To understand the importance of this variable, we will sketch a heuristic approach as the exact result depends on numerical simulations, which can be found in M. A. Nowak and Ohtsuki (2008). For the sake of argument, let us suppose that we drop the first term on the right-hand side of (2.75). In this event, it can be seen that the solution to this equation acquires the form $\exp\left(\mathcal{G}_j t\right)$. Thus, the sign of $\mathcal{G}_j$ constitutes an effective determinant for determining whether the population of replicators increases or decreases exponentially over time. By solving for $\mathcal{G}_j = 0$, we can determine the threshold θ_c at which there is a critical transition from prelife to life. For $\theta_0 < \theta_c$, the abundance of replicators is very low and $\phi \approx 0$ is a reasonably valid assumption borne out by numerical simulations. We can therefore solve for θ_c in terms of the other variables.

Hitherto, we have implicitly concerned ourselves only with error-free replication. Let us assume that we are interested in a replicator whose sequence contains n monomers. During the addition of each monomer, the probability of introducing an error (i.e., mutation) is denoted by P_m. During the n steps involved, the probability of *not* introducing an error is $(1 - P_m)^n$. In this event, the fitness f_i must be replaced by the modified fitness $f_i\,(1 - P_m)^n$ in (2.75). We can repeat the above calculation by demanding that the modified net reproduction rate $\mathcal{G}_j$ is positive and neglecting ϕ as it will be very small at the critical threshold. Hence, we end up with

$$(1 - P_m)^n > \frac{k_d + r_{j0} + r_{j1}}{\theta_0 f_j}. \tag{2.77}$$

After taking the natural logarithm of this expression and using the fact that $\ln(1 - x) \approx -x$ when $x \ll 1$, we obtain

$$P_m < \frac{1}{n} \ln\left(\frac{\theta_0 f_j}{k_d + r_{j0} + r_{j1}}\right), \tag{2.78}$$

implying that the critical error rate is inversely proportional to the length of the sequence. This is closely connected to, but *not* the same as, the famous *error threshold* introduced by Eigen that related the maximum sequence length to the error rate.

The above analysis was both abstract and highly simplified. Nonetheless, two broad conclusions can be tentatively drawn. It is only when the replication rate exceeds a certain threshold that life (endowed with replication) is likely to be selected over prelife. Furthermore, in the presence of mutations, there exists an error threshold, which is inversely proportional to the sequence length, that must be exceeded for life to emerge. In both instances, it is observed that a critical transition occurs in the crossover from prelife to life. Evidently, actualizing this transition in the laboratory comprises a daunting challenge, and we may never witness an unambiguous eureka moment—to wit, this crossover could become apparent only after the threshold has been crossed (Schwille 2017).

2.8.2 Thermodynamical constraints on self-replication

One of the most basic facts that stands out when it comes to self-replication in living organisms is its irreversible nature; in other words, a single organism is more likely to divide into two as opposed to the converse, two organisms spontaneously merging to yield one lifeform. In 2013, Jeremy England used nonequilibrium statistical mechanics to study the heat dissipated during self-replication. As the paper necessitates a knowledge of nonequilibrium statistical mechanics, a prerequisite *not* assumed for this book, we will restrict ourselves to sketching some of the salient intermediate steps and analyzing the final result. For a complete derivation, the reader is directed to England (2013); on a related note, a kinetic and thermodynamic analysis of replication is presented in Gaspard (2016).

The conditional probability that a system is found in the microstate b at some time $t = \tau_f$ provided that it started at the microstate a at $t = 0$ is denoted by $\pi\ (a \to b)$. In the case of microscopic irreversible dynamics, the probability of a given trajectory and its inverse are related to the amount of heat dissipated—*entropy production,* to be more accurate—during this process via the relation

$$\frac{\pi\ (b \to a)}{\pi\ (a \to b)} = \langle \exp\left[-\beta \Delta Q(a \to b)\right] \rangle_{a \to b}, \tag{2.79}$$

where $\beta = 1/k_B T$ is the inverse temperature and $\Delta Q(a \to b)$ is the heat released during the transition $a \to b$. The appearance of $\langle \ldots \rangle$ on the right-hand side is a consequence of the fact that there exist multiple paths from a

to b and hence the weight, $\exp[-\beta\Delta Q(a\to b)]$, must be averaged over all possible trajectories from a to b; this weight closely resembles the Boltzmann factor from equilibrium statistical mechanics.

The preceding discussion is only centered on microstates, but we are interested in statements about macrostates. Suppose, for instance, that $\mathcal{A}$ denotes the macrostate in which one entity exists just prior to the onset of replication and $\mathcal{B}$ is the macrostate with two entities just after the end of self-replication. We are interested in the macroscopic conditional probabilities $\pi(\mathcal{A}\to\mathcal{B})$ and the reverse trajectory $\pi(\mathcal{B}\to\mathcal{A})$. The former, for example, refers to the probability that a macrostate $\mathcal{A}$ prepared at $t=0$ is found to be in the macrostate $\mathcal{B}$ at $t=\tau_f$ and is expressible as

$$\pi(\mathcal{A}\to\mathcal{B}) = \int_{\mathcal{B}} db \int_{\mathcal{A}} da\, p(a|\mathcal{A})\pi(a\to b)\,, \tag{2.80}$$

where $p(a|\mathcal{A})$ denotes the condition probability that the microstate a belongs to the macrostate $\mathcal{A}$. The two integrands in this expression represent the probability of finding the initial microstate a in the desired macrostate ($\mathcal{A}$) and the probability of moving from the initial microstate a to the final microstate b, which is in a different macrostate ($\mathcal{B}$). The integrals over da and db are undertaken to account for all possible combinations of a and b.

One can construct $\pi(\mathcal{B}\to\mathcal{A})$ in a similar fashion and thereby calculate the ratio $\pi(\mathcal{B}\to\mathcal{A})/\pi(\mathcal{A}\to\mathcal{B})$. After some simplifications, the following relation is obtained:

$$\frac{\pi(\mathcal{B}\to\mathcal{A})}{\pi(\mathcal{A}\to\mathcal{B})} = \left\langle \langle \exp[-\beta\Delta Q(a\to b)]\rangle_{a\to b} \exp\left(\ln\left[\frac{p(b|\mathcal{B})}{p(a|\mathcal{A})}\right]\right)\right\rangle_{\mathcal{A}\to\mathcal{B}} \tag{2.81}$$

In the above equation, $\langle\ldots\rangle_{\mathcal{A}\to\mathcal{B}}$ represents a suitably weighted average over all paths from $a\in\mathcal{A}$ to $b\in\mathcal{B}$. Upon undertaking further nontrivial algebra, we end up with the final inequality,

$$\beta\langle\Delta Q\rangle_{\mathcal{A}\to\mathcal{B}} + \Delta S_{\text{int}} + \ln\left[\frac{\pi(\mathcal{B}\to\mathcal{A})}{\pi(\mathcal{A}\to\mathcal{B})}\right] \geq 0, \tag{2.82}$$

where $\Delta S_{\text{int}} = S_{\mathcal{B}} - S_{\mathcal{A}}$ is the change in internal entropy for the forward reaction. An immediate consequence of (2.82) is that choosing $\mathcal{A}$ and $\mathcal{B}$ to be identical eliminates the last term on the left-hand side. In this limit, we

recover the second law of thermodynamics: the change in internal entropy plus the change in the entropy of the heat bath (first term on the left-hand side) is positive. In actuality the two macrostates $\mathcal{A}$ and $\mathcal{B}$ are arbitrary; from a practical standpoint, it is not wholly clear how one can identify which of the two macrostates a given microstate belongs to.

Subsequently, England (2013) specializes to the case where the entities under consideration have a replication rate k_r and decay rate k_l. Thus, in an infinitesimal time dt, the probability $\pi\,(\mathcal{A} \to \mathcal{B})$ is given by $k_r\,dt$, whereas the reverse probability $\pi\,(\mathcal{B} \to \mathcal{A})$ equals $k_l\,dt$. Thus, using this data in (2.82), the maximum value of k_r, denoted by $k_{\max}$, is given by

$$k_{\max} = k_l \exp\left(\beta \Delta q + \Delta s_{\text{int}}\right), \tag{2.83}$$

with Δq and Δs_{int} representing the heat dissipated during the replication process and the change in internal entropy, respectively. In deriving the above relation, the inequality $k_r > k_l$ was implicitly assumed to be valid. The above equation reveals that the replication rate increases with the decay rate and heat dissipated, but it decreases with the temperature. Thus, ceteris paribus, a replicator that fulfills these criteria is likely to be more favored. However, it must be appreciated that this analysis is concerned only with the thermodynamic aspects of self-replication. In other words, the role of (genetic) information does not enter the picture—when this factor is added, it should influence which entities are better at replication and undergo selection.

Finally, let us turn our attention to (2.82) once again and we will suppose that the change in internal entropy is negligible, while retaining the previous ansatz for $\pi\,(\mathcal{A} \to \mathcal{B})$ and $\pi\,(\mathcal{B} \to \mathcal{A})$. Hence, we end up with

$$\langle \Delta Q \rangle \geq 2.48\,\text{kJ/mol} \left(\frac{T}{298\,\text{K}}\right) \ln\left(\frac{k_r}{k_l}\right), \tag{2.84}$$

where we have dropped the subscript $\mathcal{A} \to \mathcal{B}$ for $\langle \Delta Q \rangle$. In the case of RNA, it was pointed out that there are grounds for assuming $k_l \sim \left(5\,\text{yr}\right)^{-1}$ (Section 2.4) and $k_r \sim (1\,\text{h})^{-1}$ (Section 2.5.2). When we substitute the above values into (2.84), we end up with $\langle \Delta Q \rangle \geq 26.5$ kJ/mol at 298 K. If we hold the ratio k_r/k_l fixed and repeat the same calculation at 373 K, which is characteristic of hydrothermal vents, we find $\langle \Delta Q \rangle \geq 33.2$ kJ/mol. As these estimates are quite close to the enthalpy of ~ 42 kJ/mol associated

with RNA replication, England (2013) argues that RNA operates close to the "limit of thermodynamic efficiency" insofar as its assembly is concerned.

The same calculation can be repeated for DNA by choosing $k_l \sim \left(3 \times 10^7 \text{ yr}\right)^{-1}$ while all other constants stay the same. At 25 °C, it is found that $\langle \Delta Q \rangle \geq 65$ kJ/mol for DNA self-replication, which exceeds the estimated enthalpy for the ligation reaction (England 2013). Thus, as per the predictions of this simple model, the self-replication of DNA would appear to be prohibited on thermodynamic grounds in the absence of external catalysts (e.g., enzymes), while this would not seem to be the case for RNA. The basic reason, from a mathematical standpoint, is due to the much higher stability of DNA against hydrolysis as compared to RNA.

2.8.3 Life and the role of information

Hitherto, the significance of information was implicit for the most part—that is, we have encountered it when discussing genes, DNA / RNA, and self-replication in general (Yockey 2005). However, it is almost certain that the storage, processing, and transmission of information would be one of the hallmarks of even extraterrestrial life. The integral role of information in living systems has been acknowledged in the classic texts by Schrödinger (1944) and von Neumann (1966).

The latter treatise, in particular, is justly celebrated for introducing the concept of a *universal* constructor (UC). While the UC does constitute an example of a self-replicating machine, its functions encompass not only "mere" copying but also the construction of other objects; on this basis, it has been proposed that living cells share some of the properties of a UC. Despite recent breakthroughs concerning the thermodynamics of information, top-down causation, and the long history of information-theoretic analyses in realms adjacent to abiogenesis, our understanding of how life and information relate to one another still remains at a nascent stage; reviews pertaining to this emerging subject include Nurse (2008), Adami and LaBar (2017), S. I. Walker (2017), and Bartlett and Beckett (2019).

Due to the relative paucity of comprehensive analyses in this subject, we will restrict ourselves to chronicling the idea of "functional complexity" that was elucidated by the Nobel laureate Jack Szostak in the 2000s (Szostak 2003; Hazen et al. 2007). Let us suppose that we have an RNA sequence consisting of L bases; the total number of sequences that are theoretically allowed is $N = 4^L$. However, in order to carry out a given biochemical

function (e.g., enzymatic activity), not all of them are equivalent. It is clear that the fraction of RNA sequences (P_e) that encode for this particular task ranges between N^{-1} and 1. Let us denote the total number of such sequences by N_e, in which case the P_e is given by

$$P_e = \frac{N_e}{N} \tag{2.85}$$

and the functional-information content is defined as

$$I_e \equiv -\log_2 P_e = -\log_2 \left(\frac{N_e}{N} \right), \tag{2.86}$$

where the last equality followed after using (2.85). Hence, from the above formula, it is apparent that I_e ranges between 0 and $2L$ bits. In contrast, note that the total information content is $I = \log_2 N = 2L$, implying that $I_e \leq I$. Another metric of interest is $L_\star = -\log_4 P_e = I_e/2$. Thus, we can interpret $L_\star$ as the effective length of the polymer required to encode the functional information. The key thing to bear in mind regarding functional information is that it represents the property of an ensemble of sequences (of a given length) and not that of a particular molecule.

Let us now consider a specific example—namely, binding to ATP with relatively high affinity. For $L = 70$, it has been shown that $P_e \sim 10^{-11}$ (i.e., the fraction of 70-mers of RNA that bind to ATP is about 10^{-11}). Using these results, we find $I = 140$ bits, whereas $I_e \approx 36.5$ bits and $L_\star \approx 18.3$. Although (2.86) depends only on the fraction P_e, there are empirical grounds for believing that it also varies with the length L. Numerical simulations seem to indicate that the functional information decreases when the sequence length increases, but it is not clear yet whether this observed trend would continue ad infinitum.

2.8.4 How did homochirality originate?

Up to this stage, we have explored how amino acids and sugars may be synthesized, as they are important components of proteins and nucleic acids, respectively. However, one crucial point that was left unsaid concerns the fact that amino acids and sugars are chiral—that is, when these molecules are superimposed on their mirror images, they are not identical; the two mirror images are known as enantiomers. Every species on Earth that we

know of is characterized by homochirality, with virtually all of the twenty standard amino acids existing in levorotatory (left-handed) form, while the sugars utilized in biology are nearly always dextrorotatory (right-handed).

For the time being, setting aside the question of how homochirality arose, two questions are worth articulating. Is life in the Universe predestined to be based on left-handed amino acids (L-amino acids) and right-handed sugars (D-sugars)? Is it possible for life on Earth to function with a mixture of left- and right-handed molecules? The answer to the first question is unclear and the proposed solutions are very complex. One of the remarkable consequences of electroweak symmetry breaking, a pioneering discovery in twentieth-century particle physics, is that the two enantiomers have an infinitesimal energy difference between them. It is therefore conceivable that electroweak parity violation is linked to biomolecular homochirality, although the evidence for this premise does not appear to be conclusive (Blackmond 2019; Pályi 2020). Hence, the converse hypothesis—namely, that life on our planet employs L-amino acids and D-sugars solely on account of random events—cannot be ruled out.

The answer to the second question is more straightforward. In a well-known experiment, Joyce et al. (1984) studied the template-directed synthesis of oligonucleotides and found that this process occurred readily if the monomers were of the same chirality as the template, whereas the oligomerization was inhibited if the monomers were of the opposite chirality. On the other hand, if the monomers were selected from a racemic mixture (equal amounts of left- and right-handed molecules), it was found that monomers with the opposite handedness were incorporated, eventually leading to the curtailment of the reaction. Research undertaken after this seminal study has furnished strong grounds for contending that extant life on Earth necessitates homochirality. The putative advantages associated with homochirality have also been extensively investigated, and the reader is referred to Ruiz-Mirazo et al. (2014) and Pályi (2020) for in-depth summaries.

If we accept that homochirality was essential for the origin of life on Earth, this raises the question of how it originated. The basic premise is that a slight enantiomeric excess (EE) was present initially, which underwent amplification owing to a number of potential reasons. The initial EE could have been generated in outer space by dint of circularly polarized radiation that preferentially decomposed one of the two enantiomers; one such potential source for circularly polarized radiation is electromagnetic scattering by dust grains. A careful analysis of the Murchison meteorite

has demonstrated the existence of an EE for L-amino acids; this discovery is noteworthy because known life on Earth *does* rely almost exclusively on L-amino acids. Alternatively, the chiral surfaces of minerals are quite effective at discriminating between left- and right-handed molecules and may therefore assist in chiral selection. In 2016, experiments suggested that an excess of L-alanine was synthesized under conditions simulating high-velocity impacts by meteorites. Lastly, natural stochastic fluctuations could engender a small EE due to the action of asymmetric forces.

Thus, it is apparent that the seed EE can be generated in a number of ways. Over the past couple of decades, a vast number of chemical and physical mechanisms have been identified in the laboratory that amplify this seed EE to very high levels. Two notable examples are the Soai reaction and Viedma ripening (Saito & Hyuga 2013). The former is a celebrated autocatalytic chemical reaction, in which a virtually pure enantiomer is produced from a nearly racemic mixture (Soai et al. 1995). Viedma ripening, originated by C. Viedma (2005), is a physical process during which a racemic mixture undergoes precipitation to form crystals such that each enantiomer preferentially interacts with other molecules of the same handedness, giving rise to small quantities of enantiomerically pure states at the end (Sögütoglu et al. 2015). Apart from these two mechanisms, it has been shown that racemic amino acids may undergo polymerization to yield homochiral oligopeptides. Another intriguing discovery concerns the synthesis of enantiomerically pure RNA precursors by starting with racemic building blocks in the presence of chiral amino acids that orchestrate the amplification of homochirality.[28] This pathway seems promising given that a slight excess of L-amino acids could occur naturally owing to various reasons delineated in the prior paragraph.

In light of the importance of autocatalysis in promoting homochirality, we will round off our discussion with the famous theoretical model proposed by F. Charles Frank (1953). The two phenomena underlying the Frank model are self-catalysis and cross-inhibition—that is, an enantiomer catalyzes its own production while inhibiting the production of the complementary enantiomer. Denoting the concentrations of two enantiomers by ϕ_1 and ϕ_2, the governing equations are

28. It is intriguing that realistic mixtures of D-sugars could have driven the enrichment of L-amino acid precursors, ostensibly coupling the enantiomeric evolution of different classes of chiral biomolecules (Wagner et al. 2017).

$$\frac{d\phi_1}{dt} = \left(k_p - k_i\phi_2\right)\phi_1 \tag{2.87}$$

and

$$\frac{d\phi_2}{dt} = \left(k_p - k_i\phi_1\right)\phi_2, \tag{2.88}$$

where k_p is the production rate and k_i is the cross-inhibition rate for the two enantiomers. This set of ordinary differential equations closely resembles the famous Lotka-Volterra (predator-prey) models in ecology. If we subtract (2.88) from (2.87), we find

$$\frac{d\left(\phi_1 - \phi_2\right)}{dt} = k_p\left(\phi_1 - \phi_2\right), \tag{2.89}$$

and this can be readily integrated to yield

$$\phi_1 - \phi_2 = \left(\phi_{01} - \phi_{02}\right)\exp\left(k_p t\right), \tag{2.90}$$

where ϕ_{01} and ϕ_{02} are the initial concentrations of the two enantiomers. We can also eliminate dt from (2.88) and (2.87), thereby obtaining

$$\left(\frac{k_p}{\phi_1} - k_i\right)d\phi_1 = \left(\frac{k_p}{\phi_2} - k_i\right)d\phi_2. \tag{2.91}$$

We integrate this equation and utilize (2.90) to obtain the final result:

$$\frac{\phi_1}{\phi_2} = \frac{\phi_{01}}{\phi_{02}}\exp\left[\frac{k_i}{k_p}\left(\phi_{01} - \phi_{02}\right)\left(\exp\left(k_p t\right) - 1\right)\right] \tag{2.92}$$

An inspection of (2.92) reveals a couple of interesting points. First, in the event that ϕ_{01} is exactly equal to ϕ_{02}, the ratio ϕ_1/ϕ_2 does not evolve over time. However, as we have remarked earlier, no mixture of enantiomers will be 100 percent racemic. Hence, excluding this case, we see that ϕ_1/ϕ_2 undergoes rapid amplification if $\phi_{01} - \phi_{02} > 0$. On the other hand, when $\phi_{01} - \phi_{02} < 0$, it is easy to verify that ϕ_2/ϕ_1 will undergo rapid amplification. Thus, as long as $\phi_{01} \neq \phi_{02}$, a nearly racemic mixture will be converted into a non-racemic one, and high levels of enantiomeric purity are achievable. Our statement is applicable even when the EE is extremely small

(i.e., provided that it's nonzero) on account of the fact that ϕ_1/ϕ_2 is a *double* exponential function of time, as seen from (2.92).

F. C. Frank (1953) presented some generalized versions of the above model, which we shall not delve into because many of the qualitative conclusions are preserved. Subsequently, a large body of literature has sprung up based on, or motivated by, the Frank model that leads to broadly similar results (Plasson et al. 2007; Gleiser & Walker 2012). One may rightly object that the dual requirements of autocatalysis and cross-inhibition are excessively stringent and unlikely to prevail in the real world. We can, therefore, eliminate cross-inhibition and incorporate the effect of noise (e.g., due to environmental perturbations) within the framework instead. It has been shown that the resultant class of theoretical models fortuitously gives rise to biological homochirality (Jafarpour et al. 2017).

In closing, we observe that homochirality is one of the most distinct signatures of biological systems on our planet. Hence, at least insofar as life-as-we-know-it is concerned, the detection of high EE characteristic of homochirality represents a genuine biosignature candidate (Glavin et al. 2020). There are multiple avenues by which high EE is detectable. For starters, in situ observations can avail themselves of chiral separation methods based on electrophoresis or chromatography. Without going too deeply into the technical details, the basic principle is to create a chiral environment that enables the two enantiomers to interact with it differently (stereospecific interactions), thereby enabling their separation and analysis (Stalcup 2010). It is much more difficult, but not impossible, to search for signatures of homochirality via remote sensing, as explained in Section 6.5.5. The best-known method entails the study of scattered circularly polarized light, which is anticipated to be produced by high concentrations of biomolecules, owing to a mechanism known as circular dichroism (Sparks et al. 2009). Hence, upcoming life-detection missions within and outside our Solar system may employ these techniques to search for homochirality in a gainful manner.

2.9 CONCLUSION

> This is the account of how all was in suspense, all calm, in silence; all motionless, still, and the expanse of the sky was empty.
>
> This is the first account, the first narrative. There was neither man, nor animal, birds, fishes, crabs, trees, stones, caves, ravines, grasses, nor forests; there was only the sky.

> The surface of the earth had not appeared. There was only the calm sea and the great expanse of the sky.
>
> There was nothing brought together, nothing which could make a noise, nor anything which might move, or tremble, or could make noise in the sky.
>
> There was only immobility and silence in the darkness, in the night.
>
> —Anonymous, *Popol Vuh: The Sacred Book of the Ancient Quiché Maya*

In this chapter, we have scrutinized a small subset of the countless questions and mysteries that surround life's etiology on Earth. We have seen how the building blocks of life are synthesized through a number of different avenues and how these blocks can undergo assembly to yield biopolymers like RNA. Recent breakthroughs in systems chemistry have paved the way toward a deeper understanding of how the first cells endowed with replication and metabolic functions emerged. At the same time, origin-of-life research has embraced its theoretical component, with interesting toy models hinting at how the transition from prebiotic chemistry to biology took place. Hence, there are grounds to be cautiously optimistic that we are reaching the end of the initial phase—to wit, explaining how the precursors of biomolecules were potentially synthesized, and assembled together, from a common starting point under conditions that were plausible on early Earth.

This optimism must be counterbalanced by the fact that oceans of ignorance still stretch ahead of us. Let us consider, for example, the RNA world. Even if we could somehow synthesize RNA from scratch in the laboratory, working with realistic feedstock molecules and pathways, deep questions still persist. How did we transition from the RNA world to the current one with DNA and proteins as the intertwined lead actors? How did the various subsystems of protocells self-assemble and interact together, and what were the potential coevolutionary processes that shaped them? What roles did environmental factors such as temperature, pH, and ionic concentration and composition play in the transition(s) to the first lifeforms? How can we utilize the likes of nonequilibrium statistical mechanics and information-theoretic concepts to refine our understanding of life's emergence? While there is promising research being undertaken to answer such questions, it may be safely said that these efforts are at a nascent stage.

If we look beyond the Earth, the aforementioned issues are automatically exacerbated by the fact that we possess no direct knowledge of extraterrestrial life yet. When confronting the question of life out there, we do not even know whether it is carbon-based, deploys water as the solvent for biochemistry, and relies on elements like sulfur and phosphorus, which are essential for biological systems on our planet. In consequence, it is fundamentally unclear whether alien life would rely on the same building blocks: for instance, which of the twenty amino acids encoded in the genetic code are actually universal? Although the quest for a universal biology is truly a formidable one, there are tentative grounds for asserting that it is no mere pipe dream, taking inspiration from the multiple interdisciplinary lines of investigation that have sprung up in the past decade.

Theoretical research motivated by physics and chemistry may help us better understand *why* the particular building blocks employed on our planet were thus selected during the process of life's emergence and the mechanisms through which they came together to give birth to a living entity (Coveney et al. 2012; Pérez-Villa et al. 2020). Laboratory experiments will evidently play a huge role in furthering our understanding of abiogenesis, especially if we are successful in synthesizing artificial living systems that do not resemble known biochemistry. This line of thought has a rich and multilayered history that spans more than a century, as evidenced by the following quotation (J. Loeb 1912, pp. 5–6):

> We all, however, desire to know how life originates and what death is, since our ethics must be influenced to a large extent through the answer to this question. . . . The gap in our knowledge which we feel most keenly is the fact that the chemical character of the catalyzers (the enzymes or ferments) is still unknown. Nothing indicates, however, at present that the artificial production of living matter is beyond the possibilities of science. . . . I believe that we must also follow out the other problem: namely, we must either succeed in producing living matter artificially, or we must find the reasons why this is impossible.

Lastly, advances in phylogenetics, aided by paleontology and geochemistry, ought to help us ascertain the timing of LUCA's existence, the environment in which it would have dwelt, and its physiology. Apart from

these approaches, numerous biosignatures have been identified for in situ and remote sensing missions; the latter will be covered in Chapter 6. Homochirality and "pathway complexity"—i.e., a graph-theoretic measure of the shortest route that leads to the assembly of a given compound from its constituent building blocks (S. M. Marshall et al. 2017)—might represent two interesting examples from this category.

The road ahead is undoubtedly arduous, and it would be quite foolish to downplay the myriad difficulties associated with the pursuit of universal biology, but the endeavor is not impossible prima facie. If we were to discover general principles by which life could arise in diverse environments, we may, perhaps, wend our way toward determining the frequency of abiogenesis in the future. Needless to say, if this goal were achieved, it would greatly assist in determining how and where to search for extraterrestrial life. The detection of life beyond Earth will, by its very nature, enable us to constrain the statistical likelihood and nature of abiogenesis, thereby enabling us to assess the rarity of biospheres in the Cosmos.

Chapter 3

THE EVOLUTIONARY HISTORY OF LIFE ON EARTH

I draw east to Mount Jieshi
to behold the dark green sea,
water spreading vast and calm.
Craggy islands surge and loom,
wooded thick with tree and brush;
vast grasslands stirring, lush,
ruffled by soughing autumn winds.
Breakers rise and smash and foam—
the paths of both the sun and moon
seem to spiral from those waves:
the luminescent Milky Way
seems churned within those depths.

—Cao Cao, *Behold the Dark Green Sea*

Our planet's complex and diverse biosphere, currently home to as many as one trillion species, is often taken for granted by humans. Hence, the epigraphs at the beginning and the end of the chapter—by Cao Cao and Li Qingzhao, respectively—are worth heeding, because they serve as fitting reminders of both the grandeur and the fragility of our rich and dynamic biosphere. It is thus necessary to maintain cognizance of the basic (yet profound) fact that our planet has undergone significant evolution over time. For instance, microscopic organisms were the sole inhabitants for most of its evolutionary history, with complex multicellular organisms originating much later. From a geological standpoint, the Earth was habitable since almost 4.5 Ga (Gyr ago), but oxygen rose to appreciable levels only around 2.4 Ga. This illustrates the dynamic nature of our planet, which has been shaped by geological and biological forces acting in concert.

Understanding the evolutionary history of the Earth is of paramount importance when it comes to searching for extraterrestrial life. First and

foremost, understanding the course of biological evolution on Earth serves as a reference point for identifying and quantifying signatures indicative of life (e.g., oxygen produced during photosynthesis). Second, determining how life evolved on our planet may help us gain a better understanding of whether complex life (both technological and nontechnological) is rare in the Universe. One of the best-known examples in this regard is the Drake equation (see Chapter 8), which seeks to quantify the number of extant technological species capable of interstellar communication. The factors f_l and f_i in this equation are of particular importance as they quantify the fraction of "habitable" planets on which life actually arises and the fraction of life-bearing planets on which (technological) intelligence emerges.

A better understanding of f_l and f_i, and how they depend on stellar and planetary factors, could assist in the selection of the most optimal target planets and stars in the context of searching for signatures of life, given the limited amount of resources (e.g., telescope time) available. Furthermore, the values that we assign to f_l and f_i will also play a role in determining how much funding we allocate to the search for microbial and technological life—that is, biosignatures and technosignatures, respectively. Suppose, for example, that the limits $f_l \sim 1$ and $f_i \to 0$ were valid. In this event, there would be grounds for surmising that the search for technological intelligence represents a highly unrealistic endeavor, while the search for microbial life would seem more promising. On the other hand, if $f_l \to 0$, the search for both microbial and technological extraterrestrial life may prove to be fruitless. In the most optimistic scenario, $f_l \sim 1$ and $f_i \sim 1$, the searches for both biosignatures and technosignatures would constitute valuable endeavors.

Thus, it can be cogently argued that there exist compelling reasons for studying the evolution of life on our planet. The Earth could therefore be viewed, particularly during its past eons, as an "exoplanet" in its own right. Before we embark on our journey through the major events in Earth's evolutionary history, it will behoove us to contemplate in some detail a couple of terms that we shall encounter repeatedly. The reader is cautioned, however, that the ensuing discussion is not as rigorous or comprehensive as the situation warrants, because of intrinsic constraints imposed by space and scientific expertise.

The first, and foremost, is the dichotomy between evolutionary *contingency* and *convergence,* which is sometimes expressed in terms of *chance*

and *necessity,* respectively.[1] As this is a topic that has been explored extensively, both by evolutionary biologists and philosophers of science, we will not tackle this subject in depth. The difference between these two standpoints is best understood by considering the famous "tape of life" metaphor popularized by Stephen Jay Gould in his book *Wonderful Life* (1989, pg. 48):

> I call this experiment "replaying life's tape." You press the rewind button and, making sure you thoroughly erase everything that actually happened, go back to any time and place in the past—say, to the seas of the Burgess Shale. Then let the tape run again and see if the repetition looks at all like the original.

As per Gould's formulation, contingency can be envisioned in terms of "an unpredictable sequence of antecedent states, where any major change in any step of the sequence would have altered the final result." In other words, the final outcome would be *contingent* on the steps that preceded it. The importance of contingency has also been emphasized by several biologists, most notably François Jacob and Jacques Monod (the 1965 Nobel Prize winners in medicine), who authored *The Possible and the Actual* (1982) and *Chance and Necessity* (1971), respectively. By and large, many evolutionary biologists have argued that $f_i \to 0$ (e.g., Mayr 1985) on the basis of the importance of contingency in biological evolution, neatly encapsulated by the following lines from G. G. Simpson (1964, p. 773):

> The fossil record shows very clearly that there is no central line leading steadily, in a goal-directed way, from a protozoan to man. Instead there has been continual and extremely intricate branching, and whatever course we follow through the branches there are repeated changes both in the rate and in the direction of evolution.

On the other hand, it must be appreciated that not all evolutionary pathways are ineluctably contingent. Over the past few decades, there has been mounting evidence that *convergent* evolution is widespread, and perhaps even ubiquitous to an extent. By *convergent evolution,* we refer to the phenomenon wherein organisms that are not closely related evolve similar traits

---

1. It is, of course, very likely that neither camp suffices to reveal the complete picture. Life on Earth, and elsewhere, probably entails both contingent and convergent aspects (Vermeij 2006).

"independently"; the mechanisms underpinning evolutionary convergence and even its definition and scope have been subject to much debate (Stayton 2015).[2] The world around us is replete with examples ranging from animal locomotion and vision to light-capturing devices in plants for photosynthesis. As we shall see later, it is plausible that high intelligence might also belong to this category. A detailed exposition of convergent evolution with myriad examples can be found in S. C. Morris (2003, 2015) and McGhee (2011). Proponents of convergent evolution have typically, but *not* always, championed relatively high values of f_i— including the limiting case of $f_i \sim 1$ (Cirković 2018b)—once life had gotten started.

The preceding discussion can be envisioned in terms of the behavior evinced by nonlinear dynamical systems. In chaotic models, the trajectories may diverge dramatically even when the initial conditions are in very close proximity. On the other hand, in systems with attractors (e.g., fixed points), a wide array of initial conditions will converge toward the same final state. Thus, in the same spirit, evolution might be characterized by both types of behavior, albeit to varying degrees of potency. An important point to appreciate is that the interplay between contingency and convergence should *not* be naively equated with randomness and determinism. As seen from the above example, chaos appears outwardly as though it is random, but it actually originates from deterministic dynamical systems.

The last point to highlight concerns the notion of improbable versus probable events. There is often a tendency in the literature to equate deterministic (or convergent) processes with a high probability and random (or contingent) mechanisms with a low probability. It is imperative to recognize that this equivalence is not generally valid. A deterministic process may function with a very low probability of occurrence; this could, for example, stem from the fact that a large number of preconditions must be satisfied. In contrast, a chance event with a low probability of success, if repeated enough times, will eventually yield a high probability of being actualized.

It is instructive to explore the latter point further. Let us suppose that an event has a probability of success denoted by P_0 (with $P_0 \ll 1$). The

2. A vital point worth appreciating is that the last common ancestor of different species collectively invoked as an example of evolutionary convergence might have already possessed certain essential adaptations that enabled the subsequent emergence of seemingly convergent features (Blount et al. 2018). Hence, it is conceivable that the outwardly independent evolution of similar characteristics is not wholly independent in actuality.

probability of this event not occurring in a single trial is $1 - P_0$. After n_t trials, the probability of nonoccurrence is generalized to $(1 - P_0)^{n_t}$. Hence, the probability P_n that the event will occur after n trials is

$$P_n = 1 - (1 - P_0)^{n_t} . \tag{3.1}$$

Let us now assume that we seek to achieve a high occurrence probability of $P_n = 0.999$. From the above equation, we obtain

$$n_t \approx \frac{6.9}{P_0}, \tag{3.2}$$

implying that the number of trials required for success is approximately inversely proportional to the probability of success in a single trial. Thus, in principle, given a sufficiently high number of trials, even highly improbable events have a high chance of coming true (Hazen 2017).

In this chapter, we will delineate some of the major events in the evolutionary history of the Earth and discuss their likelihood of occurrence on other worlds. We will adopt the moniker *major evolutionary events* (MEEs) henceforth when referring to them.[3] A summary of MEEs has been presented in Figure 3.1, and we recommend that readers consult this figure while perusing through this chapter. Although we use the word *evolutionary* in defining MEEs, it must be understood that most of them were shaped by the environmental conditions on Earth at that time. In the same spirit, most MEEs were accompanied by the emergence of new biological feedback mechanisms that regulated the subsequent environmental conditions on our planet. The two-way coupling between the "ecological theater and the evolutionary play," a phrase coined by G. Evelyn Hutchinson, has been meticulously surveyed by the likes of Odling-Smee et al. (2003), Sultan (2015), Laland et al. (2016), and Hendry (2017).

At the very outset, it must be emphasized that the mechanisms behind these events, and even their timing, is not yet clear. Hence, for instance, we cannot determine with absolute certainty whether the MEEs arose through a vast number of incremental and cumulative processes (each endowed with

3. Our nomenclature differs from the standard paradigm of major evolutionary transitions (METs), as we consider events that are not conventionally listed under the category of METs. We will, however, revisit the METs at a later juncture (Section 3.9.1).

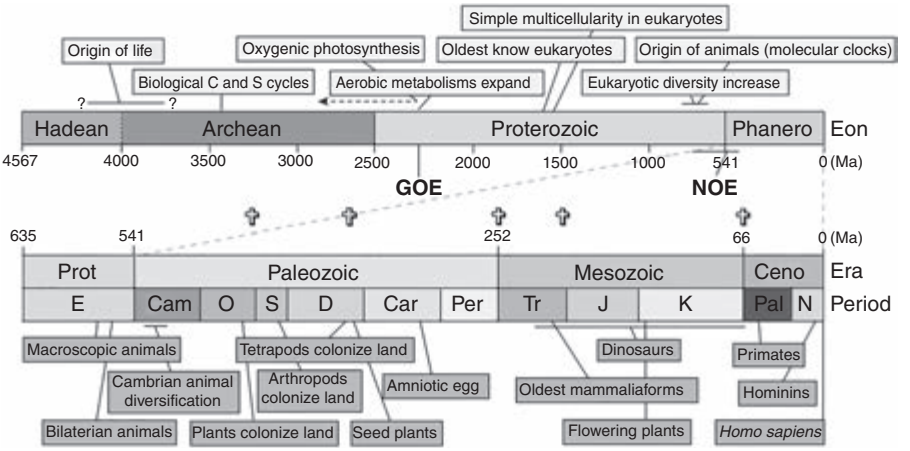


Figure 3.1 A timeline of the major evolutionary events in the history of the Earth. The crosses in the figure are signposts for the five major mass extinctions. Phanero = Phanerozoic (541 Ma-present); Prot = Proterozoic (2500–541 Ma); Ceno = Cenozoic (66 Ma-present); E, Ediacaran (635–541 Ma); Cam = Cambrian (541–485 Ma); O = Ordovician (485–444 Ma); S = Silurian (444–419 Ma); D = Devonian (419–359 Ma); Car = Carboniferous (359–299 Ma); Per = Permian (299–252 Ma); Tr = Triassic (252–201 Ma); J = Jurassic (201–145 Ma); K = Cretaceous (145–66 Ma); Pal = Paleogene (66–23 Ma); Neo = Neogene (23–2.6 Ma). Ma represents 1 Myr (10^6 yrs) ago, GOE and NOE represent the Great Oxidation Event and the Neoproterozoic Oxygenation Event, respectively, and the exact intervals for the geological epochs are subject to some uncertainty. *Source:* Andrew H. Knoll and Martin A. Nowak [2017], The timetable of evolution, *Science Advances* 3[5]: e1603076, fig. 1.)

a high probability of occurrence) or via singular and highly improbable events. Similarly, we cannot tell whether the same *functions* associated with the MEEs could have arisen through only a single channel or many paths. Hence, our subsequent exposition and the dates provided therein should be viewed with due caution.

3.1 THE ORIGIN OF LIFE

As we have already covered this topic in some detail in Chapter 2, we will not address it further. It suffices to recap at this stage that we have witnessed much progress in our understanding of how life originated on Earth, but diverse and significant challenges are yet to be overcome. Suppose, for example, that one subscribes to the RNA world paradigm, whereby RNA

fulfilled the dual functions of DNA and proteins in modern organisms. In Section 2.5, we saw that, notwithstanding the many advantages and accomplishments of the RNA world, the laboratory synthesis of RNA under geologically plausible conditions is tremendously difficult. Furthermore, even if the emergence of RNA can be taken for granted, the sophisticated molecular machinery that characterizes even the most basic unicellular microorganisms is many orders of magnitude more complex than "mere" biomolecules such as nucleic acids, proteins, and lipids at first sight. For this reason, several authors contend that the gap between nonlife and life is more profound than the differences between the smallest bacterium and the gargantuan blue whale.

At the same time, it must be appreciated that life originated relatively quickly on Earth. On the basis of the available evidence (Section 2.1), it can be concluded with near certainty that life existed on Earth by 3.7 Ga. As the Earth was potentially habitable as early as ~ 4.5 Ga, this specifies an upper bound of 800 Myr on the timescale for abiogenesis ($t_{0,\oplus}$); in reality, the actual value of $t_{0,\oplus}$ could be much smaller. The Earth is believed to sustain habitable climates, barring anthropogenic causes, for a total duration of $t_{H,\oplus} \sim 5.5$ to 6.5 Gyr (Goldblatt & Watson 2012; Wolf & Toon 2015; de Sousa Mello & Friaça 2020). Thus, we see that the timescale for the emergence of life on Earth was $\lesssim 10$ percent of the total habitability interval.

In view of this datum, the relatively rapid origination of life on our planet has been adduced for concluding that abiogenesis may be quite common in the Universe (Lineweaver & Davis 2002). However, an in-depth analysis of this topic using Bayesian reasoning has demonstrated that rapid abiogenesis on Earth does not furnish definitive evidence for either the rarity or commonality of abiogenesis elsewhere (Spiegel & Turner 2012). In order to systematically gauge the rate(s) of abiogenesis, the hunt for in situ markers of ancient life on Earth and other worlds, the pursuit of laboratory experiments and computational models for actualizing artificial life, and surveys of extrasolar planets for signatures of biological activity are mandatory; of this trio, it is plausible that exoplanet surveys hold more promise for constraining the rate(s) of abiogenesis (J. Chen & Kipping 2018).

To illustrate the points made in Spiegel & Turner (2012) using a simpler approach, we will parallel the reasoning in B. Carter (1983). Suppose that we denote a hypothesis by H and the data by D. As per Bayes' theorem, we have

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}, \tag{3.3}$$

where $P(H|D)$ and $P(D|H)$ are the conditional probabilities of the hypothesis given data and vice versa, respectively. $P(H)$ and $P(D)$ denote the probabilities associated with the hypothesis and the data. Let us now consider two different hypotheses, H_c and H_r, denoting that the origin of life is common or rare in the Universe. Using (3.3), it follows that

$$\frac{P(H_c|D)}{P(H_r|D)} = \frac{P(D|H_c)}{P(D|H_r)}\frac{P(H_c)}{P(H_r)}. \tag{3.4}$$

If we were to discover a second sample of life that has an origin independent of the presently known organisms on Earth, we would have grounds for believing that $P(D|H_c)/P(D|H_r) \gg 1$. To put it differently, the available data would be seemingly more consistent with H_c than with H_r.

On the other hand, given only one sample of life (which is our current status), it becomes nearly impossible to distinguish between H_c and H_r because of observation-selection effects. To understand why this is true, consider the extreme scenario wherein the probability of life's emergence is so low that it has emerged on a single planet, the Earth, and nowhere else. In this case, it becomes a virtual tautology to say that we (intelligent and conscious observers) must naturally find ourselves situated on the same planet. Thus, even if the origin of life were exceptionally rare, we would always find ourselves on a planet which has life (by definition), making it very difficult to identify whether abiogenesis is rare or common based on just a single data point.

Detecting a second sample of life is therefore highly valuable, arguably even necessary, from the standpoint of determining whether life is common or rare in the Universe. There are several avenues by which such a second sample can be detected, in principle, provided that it exists. The first possibility is that life might have had multiple origins on Earth, and some forms of life distinct from our own may survive in "shadow biospheres" (Davies et al. 2009). Such lifeforms could, perhaps, be identified if their constituent biomolecules are characterized by (1) chirality that is opposite to known life on Earth or (2) the presence of nonorganic biochemistry, solvents apart from water, or functional groups other than phosphates (e.g., arsenates).

Alternatively, a second sample might exist in certain subsurface environments on Mars, the clouds of Venus, and the oceans and lakes of Titan.

However, it must be recognized that any biospheres on Mars and Venus may have been contaminated with Earth-based life due to the exchange of life via rocky debris ejected during impact events (lithopanspermia), as expounded in Chapter 10. Another class of worlds that constitute intriguing candidates for detecting life within our Solar system are those endowed with subsurface oceans such as Europa and Enceladus; owing to their importance, they are separately addressed in Chapter 7. Finally, the last possibility is that future observations could discern signatures of life on planets outside our Solar system. There are a number of forthcoming telescopes with the capacity to search for signatures of life, and an impressive array of putative biosignatures have been identified in this realm. We will return to both of these issues in Chapter 6.

To round off our discussion, very little is known to date about how life originated on our planet, let alone on other worlds. Hence, we are not in a position to determine the frequency with which it can emerge on other worlds. However, given the vast jump in complexity during the transition from simple inorganic molecules to a fully functional cell, it seems very plausible that the origin of life can be regarded as a difficult evolutionary step.

3.2 THE DIVERSIFICATION OF BACTERIA AND ARCHAEA

In Chapter 2, we briefly encountered prokaryotes, which are basically unicellular organisms that lack organelles (specialized substructures within the cell) with membranes. It is widely held that the first living organisms were prokaryotes. The best-known examples of prokaryotes are bacteria, and indeed it was believed for a long time that they constituted the entirety of prokaryotes. However, in 1965, Emile Zuckerkandl and Linus Pauling (a double Nobel Prize laureate) proposed that the genetic sequences of different organisms could be analyzed to trace and demarcate their relationships with one another. In a landmark paper, Woese and Fox (1977) studied the RNA sequences in ribosomes (sites of protein synthesis) and demonstrated that prokaryotes comprise *two* domains: Bacteria and Archaea.

The latter had been hitherto viewed as "weird" bacteria, but subsequent studies have clearly established that archaea are very different from bacteria in multiple respects, as discussed below (Baker et al. 2020). Identifying which species of bacteria or archaea is the most ancient, or determining when the split between these two domains occurred, is a challenging endeavor for a

number of reasons. One of the foremost difficulties stems from the phenomenon of horizontal gene transfer (HGT), whereby prokaryotes were swapping genetic material with each other, possibly via viruses or through the intake of DNA fragments from dead cells. HGT occurs not only among prokaryotes but also between prokaryotes and eukaryotes (Husnik & McCutcheon 2018); we shall delve into the latter in Section 3.5. Despite the complications introduced by HGT, the field of phylogenetics has accomplished many breakthroughs in the twenty-first century, a few of which will be covered later.

Methanocaldococcus jannaschii stands out as the first archaeon whose genome was completely sequenced by scientists at the erstwhile Institute for Genome Research (TIGR) headed by Craig Venter in 1996. At that time, it was concluded that < 20 percent of its genes were shared by bacteria whose genomes were already known, and that a significant fraction of its genes (~ 50 percent) were unknown in both eukaryotes and bacteria. The major differences between archaea and bacteria include differences in the enzymes used for DNA replication and the lipids that make up cell membranes. A broad spectrum of hypotheses have been advanced to explain these differences, which we shall not address here. One promising approach suggests these differences can be traced to how the flux of protons into the cell (see Section 2.7.1) was utilized to fix carbon (i.e., synthesize organic compounds). For an exposition of this idea, the reader may consult Lane (2015).

The importance of prokaryotes in the context of Earth's evolutionary history primarily stems from the fact that they play a major role in the biogeochemical cycles of bioessential elements such as carbon, nitrogen, and sulfur. It is, in fact, not an exaggeration to say that the biological fluxes for five of the six bioessential elements (C, H, O, N, and S) are driven by redox reactions (involving electron transfer between species) catalyzed by microbes (Falkowski et al. 2008; Jelen et al. 2016). The various biochemical reactions driven by microbes and the abiotic processes that supply the requisite raw materials and energy are depicted in Figure 3.2. Let us consider some of these reactions in greater detail; we shall adopt the nomenclature and values delineated in Schlesinger and Bernhardt (2013).

The first reaction of import entails the production of methane (CH_4) via archaea known as methanogens by means of the acetyl-CoA pathway discussed in Section 2.6.4. In a simplified form, the overall reaction is expressible as

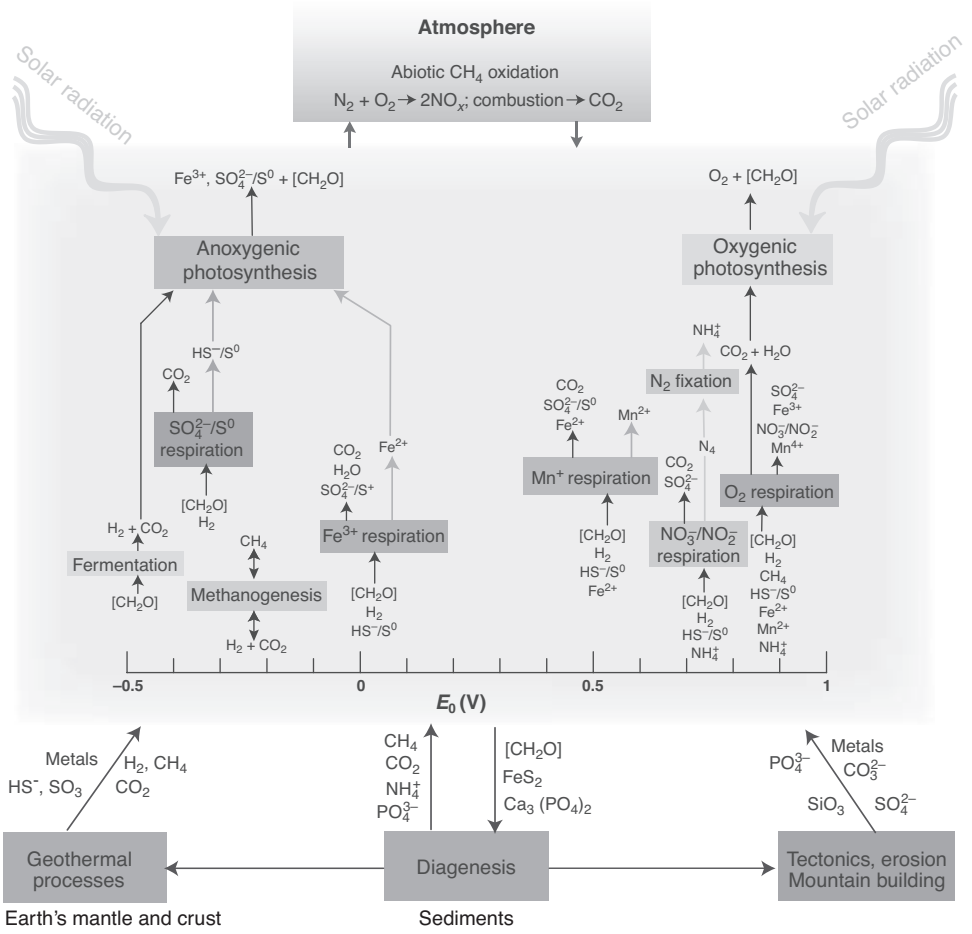


Figure 3.2 The inputs and outputs of energy and raw materials in the biosphere. The abiotic components are depicted at the top (atmosphere) and bottom (tectonic and geothermal), whereas the middle, light-shaded region depicts biochemical reactions driven by microbes. Broadly speaking, respiration involves the breakdown of organic compounds to release energy, whereas photosynthesis entails the synthesis of organic compounds using solar radiation. The reduction potential (E_0) is shown for different reactions and quantifies the propensity of chemical species to acquire electrons and thereby become reduced. (© American Association for the Advancement of Science. *Source:* Paul G. Falkowski, Tom Fenchel, and Edward F. Delong [2008], The microbial engines that drive Earth's biogeochemical cycles, *Science* 320[5879]: 1034–1039, fig. 1.)

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O, \tag{3.5}$$

and the overall reaction is thermodynamically favorable (exergonic). It must, however, be appreciated that this reaction proceeds in different stages, of which the initial step is thermodynamically uphill (endergonic). The methane produced via this process could have been used by other prokaryotes as the substrate for their biochemical reactions. Alternatively, it could have escaped into the atmosphere, where it underwent photodissociation by UV radiation. Methanogenesis, in addition to producing methane, was an important source of carbon fixation on the early Earth. It has been estimated that $\sim 4.1 \times 10^{10}$ kg / yr of organic carbon (C) was generated through this process (see Kharecha et al. 2005).

Next, let us consider nitrogen (N_2), which is known to be a very inert gas with a residence time of $\sim 10^9$ yr in the atmosphere. Given the importance of nitrogen in biomolecules (e.g., amino acids), it is truly remarkable that microbes have found a way of fixing nitrogen from the atmosphere by converting it into ammonia (NH_3) or ammonium (NH_4^+). The corresponding chemical reaction is expressible as

$$N_2 + 8H^+ + 8e^- + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i, \tag{3.6}$$

where ATP, ADP, and P_i denote adenosine triphosphate, adenosine diphosphate, and the phosphate group (PO_4^{3-}), respectively. The presence of ATP is necessary to power this chemical reaction as breaking the triple bond in N_2 requires a high amount of energy (~ 946 kJ / mol). The fixation of nitrogen depends on complex enzymes called nitrogenases that, in turn, necessitate elements such as iron and molybdenum for their functioning. The ammonia thus produced is oxidized by other microorganisms (both bacteria and archaea) to yield nitrite (NO_2^-) and nitrate (NO_3^-) as follows:

$$NH_3 + 1.5O_2 \rightarrow NO_2^- + H_2O + H^+ \tag{3.7}$$

$$NO_2^- + H_2O \rightarrow NO_3^- + 2H^+ + 2e^- \tag{3.8}$$

The nitrite derived from the preceding chemical reaction is used by certain bacteria to oxidize ammonium, via a process known as anammox (anaerobic ammonium oxidation), to release N_2 as a product.

$$NH_4^+ + NO_2^- \rightarrow N_2 + 2H_2O \tag{3.9}$$

Other bacteria are responsible for denitrification, whereby N_2 is produced from organic compounds (CH_2O) in the presence of nitrate ions via the reaction

$$5CH_2O + 4H^+ + 4NO_3^- \rightarrow 2N_2 + 5CO_2 + 7H_2O. \tag{3.10}$$

A set of chemical reactions mediated by microbes can also be written for the sulfur (S) cycle. For instance, sulfur-oxidizing bacteria can produce sulfate (SO_4^{2-}) from sulfur or hydrogen sulfide (H_2S). In turn, the sulfate is utilizable by sulfate-reducing bacteria for the purposes of respiration, thereby releasing H_2S into the environment.

Finally, we turn our attention to photosynthesis—to wit, the biological capture and conversion of solar energy into chemical energy. For the time being, we shall focus on photosynthesis that does not generate oxygen as the end product (i.e., anoxygenic photosynthesis), as we will cover oxygenic photosynthesis in Section 3.3. The first point to appreciate here is that a continuous free energy flux is necessary for the sustenance and growth of life. Solar (or stellar) radiation represents a readily available and plentiful energy source in this context. Hence, it is hardly surprising that many organisms on Earth have found ways to harness light energy. As a consequence, the majority of life on Earth is powered by photosynthesis either directly or indirectly.

Photosynthesis relies on a complex and interconnected series of cellular machinery that we shall not cover in much detail. The interested reader may consult Kiang, Siefert, et al. (2007) and Blankenship (2014) for in-depth overviews of this subject. The general equation for photosynthesis is given by

$$CO_2 + 2H_2A + h\nu \xrightarrow{\text{pigments}} CH_2O + H_2O + 2A, \tag{3.11}$$

where H_2A represents an electron donor (reducing agent) and CH_2O is shorthand notation for carbohydrates (organic compounds). Photosynthetically active radiation (PAR) represents the range of wavelengths commonly used by oxygenic photosynthetic organisms on Earth, lying roughly within 400 and 700 nm. However, certain organisms are capable of using wavelengths up to a maximum of $\sim 1\ \mu$m (i.e., in the near-infrared range) to carry out anoxygenic photosynthesis. Under ideal conditions,

about eight photons are required per carbon in photosynthesis, although real-world systems appear to necessitate ~ 8 to 12 photons.

Anoyxgenic photosynthesis is believed to have arisen quite early in the planet's history, as pointed out a few paragraphs hereafter. Although the exact origins of this mechanism are unclear, one notable class of proposals suggests that it evolved in shallow waters where the putative organisms were subjected to high doses of UV radiation due to the absence of an ozone layer. In order to combat the damaging effects of UV photons, a combination of pigments and proteins might have initially assisted in screening and protecting cells from UV radiation (Mulkidjanian & Junge 1997), after which the biological synthesis of organic compounds at longer wavelengths became feasible. Other hypotheses propose that primitive pigments for photosynthesis evolved in microorganisms dwelling near hydrothermal vents at the seafloor (W. F. Martin et al. 2018).

The machinery behind the photosystems responsible for photosynthesis is highly complex. Light-harvesting complexes (LHCs) facilitate the enhanced absorption of light, whereas the reaction center (RC) complex serves as the site for converting radiation into chemical energy through electron transfer mechanisms. There are two types of RC complexes: Type II and Type I. In anoxygenic photosynthesis, Type I RCs reduce iron-sulfur proteins (ferredoxins), whereas Type II RCs are deployed to reduce organic compounds called quinones. In contrast, oxygenic photosynthesis comprises *two* reaction centers associated with the so-called Type I and Type II photosystems (linked in series). The latter enables the extraction of electrons from water, whereas the former entails the reduction of the cofactor nicotinamide adenine dinucleotide phosphate (abbreviated as $NADP^+$), which serves as an electron carrier. As per most models, all RCs on Earth are believed to have evolved from a common ancestor, which might have possessed a structure that was intermediate between Type I and Type II; we address this topic further in Section 3.3.

At the heart of photosynthesis are pigments known as chlorophylls (Chls), which constitute the defining feature of RCs. In oxygenic photosynthesis, the chlorophylls represent a group of related molecules that share the same core structure but have different side chains. Another important class of pigments are the bacteriochlorophylls (BChls), which have been documented in a number of bacteria that perform anoxygenic photosynthesis. Although subtle differences between Chls and BChls do exist, the two groups are very similar in many respects, owing to which we shall treat

them in a unified manner. Given that Chls are complex molecules, it is not surprising that their biological synthesis is very intricate and appears to necessitate at least seventeen distinct steps. It has been shown that the structure of Chls is closely akin to that of the heme group—which is present, for example, in hemoglobin that serves as an oxygen carrier—implying that the biosynthesis of Chl was perhaps co-opted from the potentially older biochemical pathway of heme production.

Anoxygenic photosynthesis involves a plethora of electron donors. It is instructive to list the three pathways that are expected to have been the most prominent on Archean Earth. The first involves the use of H_2 as the reducing agent:

$$2H_2 + CO_2 \rightarrow CH_2O + H_2O \tag{3.12}$$

The amount of carbon fixed through this pathway on early Earth has been estimated to be $\sim 3.5 \times 10^{11}$ kg / yr. The second pathway is based on the use of hydrogen sulfide as the electron donor, and the corresponding reaction is

$$2H_2S + CO_2 \rightarrow CH_2O + H_2O + 2S. \tag{3.13}$$

The primary production through this pathway (i.e., the amount of carbon fixed) is expected to have been $\sim 3 \times 10^{10}$ kg / yr. The third pathway, which involves the use of Fe^{2+} as the reducing agent, is expressible as

$$4Fe^{2+} + HCO_3^- + 10H_2O \rightarrow CH_2O + Fe(OH)_3 + 7H^+. \tag{3.14}$$

The primary production associated with this pathway, according to some theoretical calculations, was the highest of various anoxygenic photosynthetic routes, with an estimated yield of $\sim 4 \times 10^{12}$ kg / yr (Canfield et al. 2006).

Finally, it must be noted that the fixation of carbon (i.e., the conversion into organic compounds) takes place through a number of different pathways in photoautotrophs.[4] The most widespread among them is the Calvin-Benson-Bassham (CBB) cycle, which we return to in Section 3.3 as all oxygenic photosynthetic organisms and some anoxygenic ones utilize this pathway. However, in place of the CBB cycle, the reverse tricarboxylic

4. Photoautotrophs are organisms that produce their own "food" via carbon fixation through metabolic networks involving photosynthesis.

acid (rTCA) cycle encountered in Section 2.6.3 forms the basis for carbon fixation in green sulfur bacteria (phylum *Chlorobi*). Apart from these two pathways, green non-sulfur bacteria (members of phylum *Chloroflexi*) adopt a third route known as the 3-hydroxypropionate cycle.

Thus, given the evidence so far, we can ask ourselves the question of whether anoxygenic photosynthesis could readily arise on other worlds. The preceding discussion has shown that a number of electron donors take part in photosynthesis and that the synthesis of organic compounds (carbon fixation) is feasible through at least three different routes on Earth. Therefore, we appear to have grounds for concluding that many different routes are possible for anoxygenic photosynthesis, thereby suggesting that its evolution may not be uncommon. However, we encounter a potential snag: the centrality of Chls in RCs. Fortunately, there exist other avenues through which the capture and conversion of light energy into chemical energy is rendered practical.

The most notable among them are certain pigments from the class of microbial rhodopsins. Apart from Chls and rhodopsins, other compounds such as carotenoids and melanin have been proposed as the cornerstones of potential light-capture mechanisms, but the evidence is arguably slender. Rhodopsins are photoreceptor proteins that are present in the retina and facilitate vision in environments with low photon fluxes. However, some microbial rhodopsins are known to function as light-driven proton pumps with the capacity to facilitate the production of ATP through a mechanism known as chemiosmosis (see Section 2.7.1). An interesting feature of select microbial rhodopsins is that their absorbance characteristically displays a distinctive peak around 570 nm, which is complementary to chlorophyll pigments because Chls are characterized by relatively low absorption of photons at wavelengths $\sim$ 500 to 600 nm.

The proton pumps in the category of microbial rhodopsins include proteorhodopsins, bacteriorhodopsins, xanthorhodopsins, and heliorhodopsins, of which the former are the most prevalent. Moreover, the fraction of heterotrophic bacteria endowed with proteorhodopsin genes in the ocean is $\sim$ 75 percent, and the amount of solar energy harvested by this pigment is probably comparable to that captured by chlorophyll a (Gómez-Consarnau et al. 2019). However, an important caveat must be pointed out at this juncture. No concrete evidence exists to date for rhodopsin autotrophy—that is, the fixation of carbon using microbial rhodopsins. A theoretical calculation by Larkum et al. (2018) indicates that the conversion of CO_2 to CH_2O

using proteorhodopsin photochemistry would necessitate ∼18 photons. This number is close to the estimate of ∼16 to 18 photons for organisms endowed only with Type II RCs. However, this value is approximately double that of the number of photons required for fixing carbon using oxygenic photosynthesis.

The final line of evidence that we wish to delineate concerns the timing of the evolution of anoxygenic photosynthesis, and other metabolic pathways, on our planet. However, as emphasized in Section 3.1, we must be careful in using timescales as a means of identifying for or against the plausibility of MEEs on other worlds. Furthermore, evidence for these pathways is mostly indirect as well as very scanty, and therefore subject to much controversy. With these cautionary statements out of the way, the geological record indicates that the metabolic diversification of bacteria and archaea, including anoxygenic photosynthesis, occurred quite early in Earth's evolutionary history. In many cases, we see that the evidence for metabolic pathways emerges within a few 100 Myr after the first potentially reliable evidence for life (at 3.7 Ga).

The earliest evidence for anoxygenic photosynthesis can be traced to at least 3.5 Ga and possibly even further back to 3.8 Ga. In particular, evidence from the 3.4 Ga Buck Reef chert in South Africa (filamentous mats) and the 3.5 Ga Apex chert in Western Australia (confirmed microfossils) has been interpreted to indicate that H_2-based and / or sulfur-based anoxygenic photosynthesis was already existent in this period (J. M. Olson 2006; Schopf et al. 2018). Along similar lines, the analysis of hydrothermal vent precipitates from the Dresser Formation (Pilbara Craton, Western Australia) and carbon isotope ratios of microfossils from the Apex chert suggests that methanogens had already evolved by ∼3.5 Ga, and molecular clock evidence seems to push this date back even further. The analysis of fossil microbial mats in the Strelley Pool Formation (Pilbara Craton) seemingly indicates that microorganisms that metabolized sulfur and iron existed by 3.4-3.5 Ga. When it comes to nitrogen fixation, isotopic evidence has been interpreted in favor of its existence by at least 3.2 Ga. Further references pertaining to this topic can be found in Lingam and Loeb (2019f).

To summarize our discussion, more than one channel exists for facilitating the harnessing of light energy by biological pigments and driving the synthesis of organic compounds through multiple pathways using different electron donors. Moreover, the origin of anoxygenic photosynthesis after the emergence of life also appears to have been quite rapid, perhaps even

nearly concurrent, on our planet. While none of these reasons ought to be regarded as definitive, a case can be made that microorganisms might quickly evolve a diverse array of metabolic pathways on other worlds. If this diversification were to occur, one of the consequences is that the fluxes of bioessential elements and their global biogeochemical cycles would be regulated by the microbial biosphere.

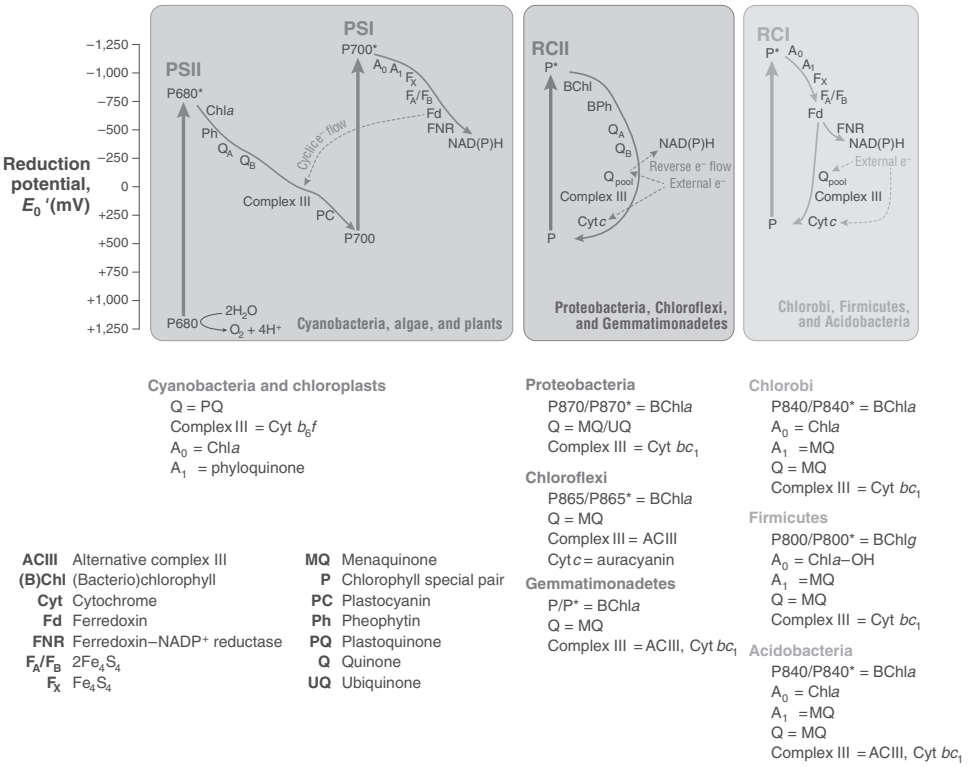
3.3 OXYGENIC PHOTOSYNTHESIS

We have discussed the origins and mechanisms of anoxygenic photosynthesis. In contrast, oxygenic photosynthesis is given by

$$CO_2 + 2H_2O + h\nu \xrightarrow{\text{pigments}} CH_2O + H_2O + O_2, \tag{3.15}$$

and the presence of H_2O on both sides of the reaction indicates that water serves as both reactant and product. Detailed summaries of the mechanisms underlying oyxgenic photosynthesis and their evolution are found in Hohmann-Marriott and Blankenship (2011) and Fischer et al. (2016). We have identified oxygenic photosynthesis as a MEE in its own right because of two different reasons. The first stems from the fact that the electron donor in this case (water) was relatively plentiful and readily accessible when compared to the other electron donors encountered in anoxygenic photosynthesis (L. M. Ward et al. 2019). Equally importantly, it enabled the production of oxygen, which transformed the subsequent evolutionary and geological history of the planet, as outlined further in Section 3.4.

As seen in Figure 3.3, one of the chief differences between anoxygenic and oxygenic photosynthesis is that the latter involves two distinct photosystems. One of the chief points worth appreciating here is that the reaction center of photosystem II in oxygenic photosynthesis is composed of P680 (chlorophyll a molecule), which loses an electron when it undergoes excitation via light energy. The resultant chemical species ($P680^+$) is an extremely powerful oxidizing agent, as seen from its high reduction potential of $\sim$ 1.25 V. It is this property, above all else, that enables the oxidation of water during oxygenic photosynthesis. Hence, it is instructive to focus on how water oxidation is facilitated by the water-oxidizing complex (WOC) in photosystem II; alternative contenders for the WOC on other worlds are touched on in Lingam and Loeb (2020e).



Cyanobacteria and chloroplasts
Q = PQ
Complex III = Cyt b_6f
A_0 = Chl*a*
A_1 = phyloquinone

Proteobacteria
P870/P870* = BChl*a*
Q = MQ/UQ
Complex III = Cyt bc_1

Chloroflexi
P865/P865* = BChl*a*
Q = MQ
Complex III = ACIII
Cyt *c* = auracyanin

Gemmatimonadetes
P/P* = BChl*a*
Q = MQ
Complex III = ACIII, Cyt bc_1

Chlorobi
P840/P840* = BChl*a*
A_0 = Chl*a*
A_1 = MQ
Q = MQ
Complex III = Cyt bc_1

Firmicutes
P800/P800* = BChl*g*
A_0 = Chl*a*–OH
A_1 = MQ
Q = MQ
Complex III = Cyt bc_1

Acidobacteria
P840/P840* = BChl*a*
A_0 = Chl*a*
A_1 = MQ
Q = MQ
Complex III = ACIII, Cyt bc_1

ACIII Alternative complex III
(B)Chl (Bacterio)chlorophyll
Cyt Cytochrome
Fd Ferredoxin
FNR Ferredoxin–NADP⁺ reductase
F_A/F_B $2Fe_4S_4$
F_X Fe_4S_4
MQ Menaquinone
P Chlorophyll special pair
PC Plastocyanin
Ph Pheophytin
PQ Plastoquinone
Q Quinone
UQ Ubiquinone

Figure 3.3 *Left box:* The energetics for oxygenic photosynthesis along with the reduction potentials. Large negative values indicate strong reductants (electron donors), whereas strong oxidants (electron acceptors) are epitomized by large positive values on this reversed scale. Photosystem II produces photooxidants with high reduction potentials that are necessary for driving the oxidation of water. In contrast, anoxygenic photoautotrophs with type I (*right box*) or type II (*middle box*) reaction centers lack this pivotal ability. (© Annual Reviews. *Source:* Woodward W. Fischer, James Hemp, and Jena E. Johnson [2016], Evolution of oxygenic photosynthesis, *Annual Review of Earth and Planetary Sciences* 44: 647–683, fig. 1.)

All reaction centers (RCs), barring the one associated with photosystem II, carry out photooxidation reactions entailing a single electron. In contrast, the photosystem II reaction center is endowed with the capacity for four-electron chemistry, which is necessary for water oxidation:

$$2H_2O \rightarrow O_2 + 4H^+ + 4e^- \qquad (3.16)$$

The WOC constitutes the only known biological water oxidation catalyst to date and consists of a tetramanganese cluster at its core. The assembly of this

cluster requires four Mn^{2+} ions. During five photocycles, a total of five electrons are liberated from the four ions to produce the initial oxidation state of the WOC. During four successive photocycles, the WOC enables photosystem II to split two water molecules in accordance with (3.16). Thus, to put it differently, two photons are consumed for the oxidation of a single water molecule. In light of the central role of manganese in oxygenic photosynthesis, it has been proposed that Mn^{2+} may have served as an electron donor for anoxygenic photosynthesis prior to the advent of oxygenic photosynthesis (Lingappa et al. 2019). This hypothesis appears plausible given that both shallow and deep ocean waters during the Archean era (2.5–4.0 Ga) contained high levels of Mn^{2+}. Subsequently, photosynthesis based on Mn oxidation (already endowed with two photosystems) might have evolved the WOC through mechanisms such as gene duplication.[5]

Given that anoxygenic photosynthesis uses only one photosystem, whereas its oxygenic counterpart uses two of them, the natural question is: How did the latter arise? Three natural possibilities present themselves:

- The ancestor of all photoautotrophs had two photosystems, but the anoyxgenic variants lost one or the other.
- The ancestral photoautotroph had just one photosystem initially, but it diverged into two different systems via a process like gene duplication.
- The two photosystems evolved independently in two different lineages, after which they were both incorporated in the same lineage via horizontal gene transfer (HGT).

Of these three, the first appears to be unlikely prima facie because it requires significant gene losses that are not commensurate with available records. The mechanism of HGT in conjunction with the fact that the two RCs associated with the photosystems are outwardly very distinct from one another has made it difficult to definitively identify the steps involved in the origin of oxygenic photosynthesis.

The primary function of the light-dependent reactions is the conversion of light energy into chemical energy leading to the formation of ATP

5. During DNA replication, some genes are inadvertently duplicated due to errors. The two copies of the gene may therefore acquire different functions over multiple generations.

and NADPH, the latter of which is the reduced version of $NADP^+$ introduced in Section 3.2. The chemical energy thus generated is utilized for the purpose of carbon fixation. Oxygenic photosynthesis relies on the Calvin-Benson-Bassham (CBB) cycle to fix carbon. This autocatalytic network comprises eleven different enzymes that are responsible for catalyzing a total of thirteen reactions (Raines 2003). Owing to this complexity, we will restrict ourselves to the net reaction

$$3CO_2 + 6NADPH + 6H^+ + 9ATP \rightarrow$$
$$G3P + 6NADP^+ + 9ADP + 3H_2O + 8P_i, \qquad (3.17)$$

where most of the corresponding compounds have been defined above and just after (3.6). G3P refers to glyceraldehyde 3-phosphate, which is a 3-carbon anionic metabolic product with the formula $C_3H_7O_6P$. The main enzyme involved in this cycle that initiates the chemical reactions is known as ribulose 1,5-bisphosphate carboxylase oxygenase (RuBisCO). It has been estimated that > 99.5 percent of all carbon fixed on our planet involves RuBisCO, owing to which it constitutes the most abundant protein. However, RuBisCO has a number of drawbacks associated with its use. For starters, the enzymatic activity of RuBisCO is slower compared to its counterparts from other pathways. More importantly, RuBisCO is (in)famous for also functioning as an oxygenase, during which O_2 is consumed as a reactant and CO_2 is generated as a product. This reaction is wasteful from the standpoint of energy expenditure and the efficient utilization of CO_2.

These difficulties have led to numerous studies of why RuBisCO and the CBB cycle are employed for carbon fixation instead of other pathways (e.g., the rTCA cycle), which are used in anoxygenic photosynthesis (Section 3.2). One of the primary reasons may stem from the fact that many of the alternative pathways involve proteins and cofactors that are sensitive to oxygen, which is the chief byproduct of oxygenic photosynthesis. However, this is unlikely to provide a complete solution, as there exist other enzymes that ought to be compatible with aerobic (oxygenic) metabolism. Hence, it remains an open question whether the ubiquity of the CBB cycle in oxygenic photosynthesis represents a "frozen accident" or the most optimal solution (under the given circumstances) selected by evolution.

Finally, we turn our attention to when oxygenic photosynthesis arose on Earth. Many textbooks and papers often erroneously claim that cyanobacteria were responsible for the evolution of oxygenic photosynthesis, but it

is more accurate to state that ancestral cyanobacteria or somewhat closely related groups, neither of which were necessarily identical to their modern-day counterparts, constituted the first oxygenic photosynthetic organisms. The geological and phylogenetic biomarkers for oyxgenic photosynthesis have been subject to manifold interpretations, and the interval between the earliest and most recent putative lines of evidence spans almost 2 Gyr (3.8 Ga vs. 1.9 Ga). The arguments for and against these dates are marshaled and briefly reviewed below.

The earliest possible date of > 3.7 Ga was based on the notion that uraninum (U) is relatively soluble in oxidizing fluids, whereas thorium (Th) does not fulfill this property, thereby leading to changes in the U-Th-Pb ratio. The high abundance of U relative to Th in the 3.8 Ga Isua supracrustal belt in Greenland was consequently interpreted as evidence for oxygenic photosynthesis. The issue with such analyses of trace elements is that the geochemical cycles during this epoch are poorly understood, and the data are not very robust as they display variation even at the same geological site. The oldest reasonable fossil evidence that we have for cyanobacteria dates to ~ 1.9 Ga and was derived from the Belcher Islands, Canada. The geological record reveals that there was a fairly distinctive rise in O_2 levels at ~ 2.4–2.5 Ga (discussed in Section 3.4), which is conventionally attributed, at least partially, to oxygenic photosynthesis.

If we naively take the mean of ~ 3.7 Ga and ~ 1.9 Ga, we end up with a timing of ~ 2.8 Ga. Many studies, albeit not devoid of controversy, appear to indicate that oxygenic photosynthesis arose several 100 Myr prior to the aforementioned upsurge in atmospheric O_2 (Catling & Zahnle 2020). For instance, it has been proposed that the formation of Mn oxides requires substantial concentrations of free oxygen, implying that their discovery would serve as a proxy for oxygenic photosynthesis. There is indirect evidence for Mn oxides from > 2.95 Ga rocks of the Sinqeni Formation, Pongola Supergroup, South Africa, implying that oxygenic photosynthesis may have been around since nearly 3 Ga (Planavsky et al. 2014). However, it should be noted that this class of hypotheses is not universally accepted, because other papers have proposed that oxygenic photosynthesis evolved immediately before (i.e., $\lesssim 100$ kyr) the documented rise in oxygen levels (L. M. Ward et al. 2016).

The reader is directed to Schirrmeister et al. (2016) for a succinct review of ancient cyanobacteria insofar as putative microfossils together with biological and geochemical markers are concerned. Henceforth, we

will assign a fiducial value of ~2.7 Ga for the emergence of oxygenic photosynthesis, but this estimate could easily differ by roughly 0.5 Ga either way; in other words, our chosen value accordingly embodies one potential timeline among many other candidates. In fact, the most recent molecular clock analyses, *inter alia,* seemingly favor the origin of photosystems capable of water oxidation in the Archean—to wit, by ~3 Ga or even earlier (Cardona et al. 2019; Sánchez-Baracaldo & Cardona 2020).

Oxygenic photosynthesis has ostensibly evolved only a single time on our planet. A number of nontrivial requirements had to be fulfilled for its emergence: (1) the photosystem with high reduction potential that could oxidize water, (2) the WOC that catalyzed the oxidation of water to O_2, and (3) the coupling of two different photosystems. Even if we presume that oxygenic photosynthesis on other worlds necessitated the same three criteria, it does not automatically follow that molecular machinery *exactly* identical to that found in Earth's biology would evolve on other worlds. Despite the potential existence of multiple paths toward oxygenic photosynthesis, the possibility that it arose only once on our planet—and perhaps not very soon after the emergence of life—might indicate that it represents a difficult evolutionary step that is not readily achievable on exoplanets and exomoons.

3.4 THE RISE OF OYXGEN AND THE GREAT OXYGENATION EVENT

In our journey through Earth's evolutionary past, we saw that oxygen was produced as a consequence of oxygenic photosynthesis. We will therefore discuss two different, but related, aspects. First, what were the consequences of rising levels of oxygen in the environment? Second, at what point did the oxygen levels in the atmosphere rise discernibly?

3.4.1 The significance of oxygen

It is no exaggeration to say that the rise in oxygen levels dramatically, and irreversibly, changed the face of the Earth. As the causal effects are multifarious, we refer the reader to Knoll (2015, 2021) and Judson (2017) for the relevant details. We will restrict ourselves to a brief exposition, except for one particular aspect that we elaborate on in detail.

To begin with, we observe that the buildup of oxygen in the atmosphere led to the formation of ozone (O_3) via the reaction

$$3O_2 \leftrightarrow 2O_3. \tag{3.18}$$

The formation of the ozone layer facilitated a significant reduction in the amount of ultraviolet (UV) radiation that reached the surface of the planet, particularly at shorter wavelengths. The biologically effective irradiance, which is essentially the convolved product of the UV flux at the surface and the DNA action spectrum that encapsulates the degree of UV damage at different wavelengths, was estimated to be ~ 600 times on Archean Earth (at 3.9 Ga) relative to the present-day value (Rugheimer et al. 2015). The most dramatic difference is evinced for UV-C radiation (viz., wavelengths of $\sim 120-280$ nm): the surficial UV-C flux was ~ 0.9 W / m^2 at 3.9 Ga, as opposed to the negligible value of $\sim 2 \times 10^{-16}$ W / m^2 today. Thus, the radiation environment on Earth became less harsh after the development of the ozone layer. It is therefore plausible that the UV protection accorded by ozone assisted in the development of complex life on our planet, although there were clearly numerous other factors involved.

Another major consequence of the rise in oxygen levels was the rapid diversification of minerals. This stems from the fact that significantly elevated partial pressures of O_2 in the atmosphere permitted elements (with the capacity for multiple oxidation states) to exist in the form of various minerals at the surface. It has been estimated that over 2500 minerals currently prevalent on Earth were unlikely to have originated in anoxic environments—that is, those endowed with negligible oxygen (Hazen & Ferry 2010). The diversity of Earth's minerals can therefore be attributed (in)directly to biological processes (e.g., oxygenic photosynthesis), owing to which certain rare minerals may serve as signposts of life in the process of searching for biosignatures in situ.

The rise in atmospheric oxygen enabled the creation of new habitats that were oxygen rich. Previously, most of the existing niches had been anoxic or micro-oxic (with low levels of oxygen) at best. The emergence of oxic habitats provided fresh opportunities for microorganisms to colonize them, thereby contributing to an increase in biological diversity; this process was conceivably aided by the concomitant jump in metabolic complexity (Raymond & Segrè 2006). In this context, it is worth noting that some aerobic prokaryotes are more robust in terms of their tolerance toward extreme

levels of pH and salinity in comparison to their anaerobic counterparts. There are also grounds for supposing that the buildup of oxygen led to an increase in the abundance of oxidants such as nitrates and sulfates, which served as the basis for certain microbial metabolic pathways.

Oxygen is very reactive in nature, originating from the fact that its electronegativity (i.e., the measure of an element's tendency to attract electron pairs) is second only to that of fluorine. While the high reactivity of oxygen has proven to be a boon for evolution, as we shall see below, it also constitutes a potent bane in many respects. The chief reason stems from the fact that oxygen has a strong propensity for abstracting electrons from enzymes and stimulating the formation of reactive oxygen species (ROS) like hydrogen peroxide (H_2O_2) and superoxides (anions with the group O_2^-). These species can cause significant damage to DNA and proteins, especially those with iron-sulfur clusters (encountered, e.g., in Section 2.6.3). In order to combat oxidative stress, organisms had to evolve new adaptive strategies, possibly via the assimilation of preexisting systems (exaptation) endowed with serendipitous anti-oxidative properties. It seems reasonable to conclude that the buildup of oxygen would result in the exertion of strong selection pressure.

Finally, it must be recognized that oxygen constitutes an energy-rich source. Aerobic respiration yields approximately an order of magnitude more energy than anaerobic respiration per unit quantity of food (e.g., carbohydrates and fats). The amount of energy released per electron transfer, wherein O_2 serves as the electron acceptor, is higher than that of other elements, barring chlorine and fluorine. As the latter duo are not widely prevalent in the Universe, this fact lends credibility to the notion that molecular oxygen might be necessary for the emergence of complex life in the Universe. We shall delve into some caveats regarding this statement later, but one of them merits a special mention—highly oxygenated environments may stimulate the overproduction of ROS, thereby instigating oxygen toxicity and stymieing (or suppressing altogether) the evolution of complex life (Lingam 2020). In our subsequent analysis of the advantages stemming from oxygen, we adopt the line of reasoning outlined in Catling et al. (2005).

We denote the number of organisms within a given volume whose mass is greater than m by n and employ the subscripts O and A to signify aerobic and anaerobic organisms, respectively. We introduce the quantity $N(m) = \partial n/\partial m$, which is governed by the mass balance equation

$$\Gamma_m N(m) + \frac{\partial\,(N(m)\dot{m})}{\partial m} = 0, \tag{3.19}$$

where Γ_m is the death rate (units of inverse time) and $\dot{m} = dm/dt$ is the growth rate. If we interpret $N(m)$ as the particle number density, $m \to x$, $\Gamma_m \to \partial/\partial t$, and $\dot{m} \to v$, we see that the above equation is analogous to the well-known continuity equation in fluid mechanics. Let us suppose that all organisms with mass m are prey for other organisms with mass $m_1 \geq m$. In this event, we have

$$\Gamma_m = \int_m^\infty N\,(m_1)\,P\,(m, m_1)\;\, dm_1, \tag{3.20}$$

where $P\,(m, m_1)$ is the probability per unit time that an organism of mass m_1 will eat an organism with mass m. The growth rate for organisms of mass m, which eat other organisms with lower mass, is in turn given by

$$\dot{m} = G \int_0^m N\,(m_1)\,m_1 P\,(m, m_1)\;\, dm_1, \tag{3.21}$$

with G representing the ecological gross efficiency—that is, the quantity of new biomass produced per unit intake of food. The power-law ansatzen $P\,(m, m_1) \propto (m\,m_1)^\beta$ and $N(m) \propto m^{-\alpha}$ are adopted henceforth. We leave it to the reader to substitute these scalings into (3.20) and (3.21) and verify that the convergence of integrals requires $1 + \beta < \alpha < 2 + \beta$. It can also be shown that $\Gamma_m \propto m^{-\gamma}$ and $\dot{m} \propto m^\delta$, where $\gamma = \alpha - 2\beta - 1$ and $\delta + \gamma = 1$.

Using the preceding scalings in (3.19) leads to a quadratic equation in α that yields

$$\alpha = \delta + \frac{\sqrt{1 + 8G} - 1}{2G}. \tag{3.22}$$

We will henceforth use $\delta = 0.75$, which follows from a combination of empirical experiments and theoretical models. The value of G is roughly 0.3 for small aerobic lifeforms, whereas the corresponding estimate for anaerobic organisms is around four times lower due to the relatively high efficiency of aerobic metabolism. Thus, plugging these values into (3.22) leads us to $n_O \propto m^{-1.1}$, which is mostly in agreement with laboratory and field studies that have yielded $n_O \propto m^{-1}$. In contrast, for anaerobic organisms, we end up with $n_O \propto m^{-1.5}$.

Although the difference in exponents comes across as being rather small (about 0.5), this translates to significant differences in the abundances of bigger organisms. Denoting the size of the organism by a, the mass is given by $m \propto a^3$. On the basis of these relations, we see that $n_O/n_A \propto a^{3/2}$. Hence, a size difference of two orders of magnitude—for example, 2 μm versus 200 μm—implies that the number of aerobic organisms dominates over their anaerobic counterparts by three orders of magnitude, provided that the proportionality constant in n_O/n_A is of order unity. Hence, this heuristic argument suggests that the number of anaerobic lifeforms becomes very small as one approaches greater sizes, implying that there may be a general tendency (with some exceptions) for larger organisms to rely on aerobic respiration.

Next, we ask what is the largest size that an organism can attain for a given level of atmospheric oxygen. Let us first consider the case where O_2 is transported to all cells within the organism solely via diffusion. For the sake of simplicity, we will assume that the organism is spherical—changing the geometry alters the result typically by a factor that is close to unity—with a radius a. The organism consists of an inner sphere with radius r separated from the exterior by a membrane of thickness x. The volume and area of the inner sphere are $V = 4\pi r^3/3$ and $A = 4\pi r^2$. By applying Fick's law, we have

$$-K\frac{\partial \mathcal{P}}{\partial x} = J, \tag{3.23}$$

where K is the permeability constant (in m^2 of O_2 Pa^{-1} s^{-1}), $\mathcal{P}$ is the pressure (units of Pa), and J is the diffusion flux. The latter is expressible as $J = \mathcal{M}V/A$, where $\mathcal{M}$ denotes the metabolic rate (in m^3 of O_2 s^{-1} m^{-3} of tissue), while V and A are the volume and area introduced earlier. From $x = a - r$, the utilization of the chain rule leads to $d\mathcal{P}/dx = -d\mathcal{P}/dr$. Using this relation along with the prior definitions of V and A yields

$$\mathcal{P}_{ex} - \mathcal{P}_{in} = \int_0^a \frac{\mathcal{M}r}{3K}\, dr = \frac{\mathcal{M}a^2}{6K}, \tag{3.24}$$

with $\mathcal{P}_{ex}$ denoting the external (atmospheric) pressure and $\mathcal{P}_{in} \approx 0$. Hence, we end up with

$$a_{\max} = \sqrt{\frac{6K\mathcal{P}_{ex}}{\mathcal{M}}} \sim 1.8 \times 10^{-3}\ \mathrm{m} \sqrt{\frac{\mathcal{P}_{ex}}{\mathrm{PAL}}}, \tag{3.25}$$

where the last equality follows from assuming that the above variables have values comparable to those of Earth-based organisms. Note that PAL denotes the present atmospheric level of O_2—that is, 1 PAL is 0.21 atm ($\sim 2 \times 10^4$ Pa). In other words, from the above equation, it can be seen that $\mathcal{P}_{ex} \sim 6 \times 10^3$ Pa is necessary to ensure that the organism size can reach $\sim 10^{-3}$ m. One of the smallest known animals on Earth, a microscopic parasite from the genus *Myxobolus,* has a characteristic length of $\sim 2 \times 10^{-5}$ m. As per (3.25), an organism of this size may potentially exist in an environment with an oxygen content of $\sim 10^{-4}$ PAL. The green alga *Ostreococcus tauri,* among the smallest free-living eukaryotes, has a size of only $\sim 10^{-6}$ m. By deploying (3.25), we find that the corresponding O_2 content ought to be $\sim 3 \times 10^{-7}$ PAL, which is extremely low.

Next, suppose that the organism is endowed with blood circulation and the diffusion of O_2 takes place across a layer (akin to the epidermis, the outermost layer of the skin) of thickness ℓ. We will assume that O_2 can enter from both the top and bottom hemispheres. Let us focus solely on the top hemisphere, since the same arguments would apply to the lower hemisphere by symmetry. In this case, Fick's law yields

$$-K\left[\frac{(\mathcal{P}_b - \mathcal{P}_{ex})}{\ell}\right] = J, \tag{3.26}$$

where $\mathcal{P}_b$ represents the average pressure of O_2 in the blood, while the factor inside the square brackets encapsulates the constant pressure gradient across the layer. As before, we have $J = \mathcal{M}V/A$, with $V \approx 2\pi a^3/3$ and $A \approx 2\pi a^2$ denoting the volume and area of the hemisphere, respectively. Thus, substituting these relations into (3.26) yields

$$a_{\max} = \frac{3K}{\mathcal{M}\ell}(\mathcal{P}_{ex} - \mathcal{P}_b) \sim 4.8 \times 10^{-2}\,\text{m}\left(\frac{\mathcal{P}_{ex}}{\text{PAL}}\right), \tag{3.27}$$

with the last equality arising from the fact that $\mathcal{P}_b$ is taken to be the average of the oxygen pressure in the arteries ($\sim \mathcal{P}_{ex}/2$) and veins (~ 0 atm). The other quantities have been chosen to equal the characteristic values for animals. Hence, it follows from (3.27) that organisms may attain a maximum size of $\sim 3 \times 10^{-2}$ m when the atmospheric oxygen is ~ 60 percent PAL. This size is comparable to the smallest mammals on Earth, such as the bumblebee bat (*Craseonycteris thonglongyai*) and the Etruscan shrew (*Suncus etruscus*). For a basic circulatory system to exist, it has been estimated that the minimal size

should be a few millimeters. Choosing $a_{\max} \sim 2 \times 10^{-3}$ m implies, by way of (3.27), that oxygen levels of ~ 4 percent PAL could suffice for organisms with this size to function.

Thus, the preceding discussion would appear to suggest that high oxygen levels might have been a prerequisite for the evolution of macroscopic complex life, which in turn shares causal links with the advent of oxygenic photosynthesis (Payne et al. 2011). However, several caveats must be acknowledged. To begin with, the putative organisms we consider herein do *not* have sophisticated respiratory organs—for example, alveoli (air sacs) in mammalian lungs—that greatly enhance the accessible area for the diffusion of O_2. Second, we have invoked geometric arguments in our analysis that do not fully capture real-world complexity.[6] In this context, it should be noted that we have neglected the role of temperature on body size, despite its undoubted significance (Angilletta 2009). Third, the assumption that all of the atmospheric oxygen is actually accessible by organisms is an idealization. Fourth, we observe that certain sponges (e.g., *Halichondria panicea*) are capable of both survival and growth at very low oxygen levels of ~ 0.5–4 percent PAL. Finally, organism size is manifestly a function of not only environmental parameters but also evolutionary innovations—for example, the origin of eukaryotes and complex multicellularity, explored later. In fact, empirical evidence for the maximum size indicates the possible existence of two discrete jumps corresponding to these two crucial evolutionary breakthroughs.

The question of what were the minimum oxygen levels necessary for the increase in organismal size and complexity is patently difficult to answer. Notwithstanding the aforementioned limitations of our analysis, it seems reasonable to suppose that organisms with relatively large sizes and rapid locomotion would require high oxygen content, perhaps commensurate with that of modern-day Earth. We will return to this question in Section 3.6.

3.4.2 The Great Oxygenation Event

Previously, we alluded only in passing to the rise in oxygen levels that took place around 2.4 Ga. The consequences of this event were profound, as

6. One cannot help but be reminded of the "spherical cow" metaphor associated with theoretical physics, wherein toy (i.e., highly simplified) models are employed to unravel complex phenomena; the purpose of these abstractions is to make the calculations tractable.

reflected in its nomenclature: the Great Oxygenation Event (GOE). The causes and timing of the GOE constitute the subjects of active research, owing to which it must be recognized that our following discussion will undergo, in all likelihood, revisions in the future.

Arguably, the most unambiguous evidence for the GOE has been derived by virtue of mass-independent fractionation (MIF) of sulfur isotopes. Loosely speaking, mass-independent fractionation of isotopes refers to any mechanism that causes the separation of isotopes (fractionation) to take place, in which the fractionation is not proportional to the mass difference between isotopes. Sulfur MIF prior to the onset of the GOE depended on UV photolysis (dissociation of molecules by light) of SO_2 in the troposphere (lowest region of the atmosphere). The key point here is that the UV photons must have wavelengths < 230 nm. This requirement is practical only when the atmosphere lacked even a thin ozone layer. Hence, the discovery of sulfur MIF prior to the onset of the GOE constrains the available O_2 to have been $< 10^{-5}$ PAL. Subsequently, the degree of sulfur MIF was lowered after the GOE, enabling us to constrain the timing of the latter event.

Apart from sulfur MIF, isotopes of carbon, sulfur, nitrogen, iron, molybdenum, selenium, and chromium from sediments are consistent with a distinct change in the oxygen content of our atmosphere. Collectively, they offer strong grounds for asserting that the onset of the GOE can be dated to approximately 2.4 billion years ago, although this date may be subject to fluctuations of ± 0.1 Gyr. For example, the analysis by Gumsley et al. (2017) indicates that the onset of the GOE took place 2.43–2.46 Ga based on U-Pb dating of the Ongeluk Formation in South Africa. This study also concluded that (1) the GOE was roughly coeval with the first Paleoproterozoic (2.5–1.6 Ga) global glaciation (Snowball Earth) event and (2) the rise in O_2 levels was not monotonic but characterized by oscillations that continued for the next ~ 200 Myr. On the other hand, a more recent publication by Warke et al. (2020) founded on sulfur isotope measurements from the Seidorechka and Polisarka Formations in northwest Russia suggests that the GOE predated the Paleoproterozoic Snowball Earth and that the former occurred 2.43–2.50 Ga.

It was initially presumed that the GOE drove the rise in O_2 levels such that the atmospheric content was ~ 10 percent PAL during the mid-Proterozoic (1.8–0.8 Ga). However, many isotope proxies seem to indicate that this period was characterized by relatively low levels of oxygen that

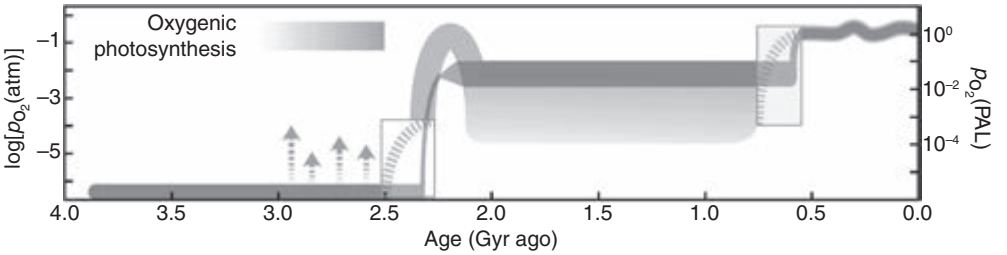


Figure 3.4 A record of the atmospheric oxygen (O_2) levels on Earth over time. The dark-shaded region demarcates the classical view that oxygen increased in two discrete steps. The broader, light-shaded region reflects the emerging model for the dynamic evolution of atmospheric oxygen content. The arrows indicate potential whiffs of oxygen, although their timing, duration, and magnitude are not tightly constrained. (© Macmillan Publishers Limited. *Source:* Timothy W. Lyons, Christopher T. Reinhard, and Noah J. Planavsky [2014], The rise of oxygen in Earth's early ocean and atmosphere, *Nature* 506[7488]: 307–315, fig. 1.)

were < 1 percent PAL (Planavsky et al. 2018), although the exact status is not conclusively settled (Lenton 2020). Finally, one other point that we wish to raise concerns the purported "whiffs" of oxygen that have been tentatively identified as having existed prior to the GOE. The basic idea is that transient O_2 accumulation in the atmosphere facilitated the oxidation of sulfide minerals, which contain trace metals such as molybdenum and rhenium. It is believed that these metals were released into rivers, and thence into oceans, as a consequence, thus leading to the enrichment of certain trace metals. However, the reasons behind the observed metal enrichments (e.g., in 2.5-Gyr-old shales in Western Australia) are not well understood and may be linked to other factors such as geochemical cycling and postdepositional processes. A summary of Earth's oxygen levels over time, modulo the many uncertainties, is sketched in Figures 3.4 and 3.5.

Turning our attention to oxgenic photosynthesis, in Section 3.3 we saw that the timing of oxygenic photosynthesis is very variable, since the available evidence for its presence before 2.4 Ga is not universally accepted. One notable class of hypotheses operates under the premise that oxygenic photosynthesis arose just prior to the GOE. Consequently, it is held that oxygenic photosynthesis led to rapid oxygenation of the atmosphere and the onset of the GOE, possibly over a timescale of $\sim 10^5$ yrs, which is short by geological standards (L. M. Ward et al. 2016). Many hypotheses, in contrast, posit that oxygenic photosynthesis arose a few 100 Myr before the GOE at the minimum and interpret the proposed evidence for whiffs of oxygen as

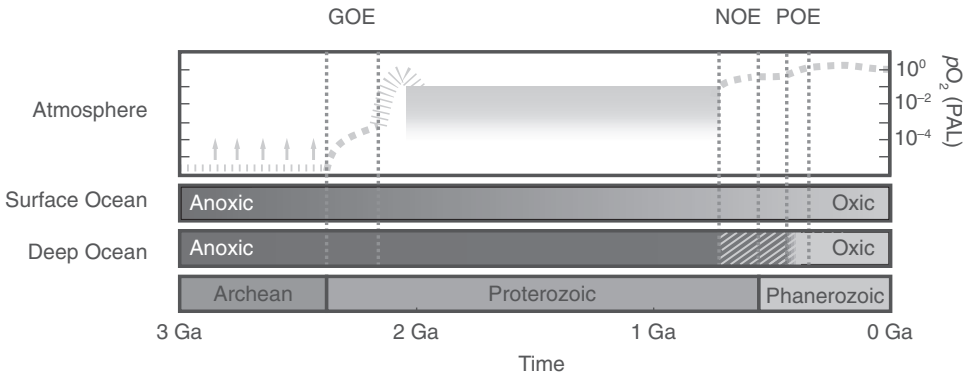


Figure 3.5 The evolution of Earth's atmospheric and oceanic O_2 reservoirs with time. GOE, NOE, and POE represent the Great Oxygenation Event, Neoproterozoic Oxygenation Event, and Paleozoic Oxygenation Event, respectively. The deep ocean is believed to have exhibited fluctuations in oxygenation in between the NOE and POE. (© American Association for the Advancement of Science. *Source:* Lewis J. Alcott, Benjamin J. W. Mills, and Simon W. Poulton [2019], Stepwise Earth oxygenation as an inherent property of global biogeochemical cycling, *Science* 366[6471]: 1333–1337, fig. 1.)

lending further credence to the idea that the former considerably predates the latter. If we do subscribe to the notion that there was a sizable gap between oxygenic photosynthesis and the GOE, then the asynchronicity of these two major events must be explained. A number of geochemical mechanisms have been advanced to this effect; a detailed overview can be found in Catling and Kasting (2017).

On Earth, the buildup of O_2 in the atmosphere is regulated by a number of sources and sinks. The primary sources for O_2 were the burial of organic matter and reducing agents such as pyrite, an iron sulfide mineral. The burial of organic matter quantifies the amount that escapes either being consumed via respiration or oxidative decay. In other words, it can be viewed as the difference between the rates of O_2 production via oxygenic photosynthesis and O_2 removal by means of aerobic respiration and decay. The major sinks for oxygen on Earth before it built up to significant levels were the production of reducing gases from metamorphic and volcanic activity.[7] It is instructive to define the parameter Δ_{O_2},

7. Metamorphism refers to the alteration of minerals in rocks because of geological processes, often (but not always) involving high temperatures and pressure.

$$\Delta_{O_2} \equiv \frac{\text{O}_2 \text{ source flux}}{\text{O}_2 \text{ sink flux}} = \frac{F_{\text{burial}}}{F_{\text{metamorphic}} + F_{\text{volcanic}}}, \tag{3.28}$$

where F_{burial}, $F_{\text{metamorphic}}$, and F_{volcanic} are the fluxes (in moles / yr) of the burial of reductants (also organic matter) and the production of reducing metamorphic gases and volcanic gases, respectively. The consumption of O_2 to oxidize Fe^{2+} in the ferruginous (iron-rich) oceans of the Archean era, thereby potentially instigating the formation of banded iron formations (BIFs), ought to be included as an additional sink in the denominator of (3.28), depending on the context.

Even on modern-day Earth, it is not easy to constrain the flux of organic carbon burial. Hence, estimating Δ_{O_2} is not an easy enterprise, but there are cogent reasons for supposing that $\Delta_{O_2} > 1$ on modern-day Earth, with $\Delta_{O_2} \approx 1.8$ representing a likely value. The importance of Δ_{O_2} stems from the fact that $\Delta_{O_2} > 1$ enables the buildup of oxygen in the atmosphere, whereas $\Delta_{O_2} < 1$ may lead to an anoxic atmosphere because the O_2 being produced by the sources is entirely consumed by its sinks. When it comes to how, why, and when the Earth transitioned from $\Delta_{O_2} < 1$ to $\Delta_{O_2} > 1$, there are two generic classes of explanations: (1) increased sources of O_2 and (2) decreasing sinks of O_2.

Let us begin with examining (1), an increase in O_2 sources. Most hypotheses in this category propose that there was an increase in the burial of organic carbon, thus enhancing the flux of O_2 (Eguchi et al. 2020). One such example claims that there was an increase in the carbon sequestered in the mantle due to efficient subduction (i.e., the process by which tectonic plates sink into the mantle). This process can, in principle, drive up the rates of organic burial, thus effectively increasing the numerator of (3.28). The second scenario relies on an increase in the area of continents, leading to higher delivery rates of dissolved phosphorus to the oceans, thus promoting greater primary productivity and resulting in higher carbon burial rates. However, a potential issue with this line of reasoning is that continental growth appears to have slowed by ~ 3 Ga, and there is some evidence that at least 70 percent of the continental crust (by volume) had formed by this period.

Many models, however, are more appropriately classified under (2), as elucidated in Kadoya, Catling, et al. (2020). The first possibility is that F_{volcanic} in (3.28) reduced over time. A number of proposals have been discussed in this setting. One notable avenue contends that the amount of SO_2

outgassed in the past was lower. The premise naturally implies, in turn, that a smaller amount of H_2 would have been consumed for the reduction of SO_2 via

$$5H_2 + FeO + 2SO_2 \rightarrow FeS_2 + 5H_2O. \quad (3.29)$$

In other words, an excess of H_2 (reducing gas) may have existed at first, thereby contributing to a higher value of F_{volcanic} that eventually diminished over time. Another candidate from this category is the hypothesized transition from submarine (underneath the sea surface) to subaerial (surface-based) volcanism. The emissions associated with the former tend to comprise more reduced gases in comparison to the latter. Hence, it is conceivable that the transition was accompanied by a decrease in the flux of reductants, consequently facilitating the rise in oxygen levels.

The other option is that the flux of reducing gases produced via metamorphism ($F_{\text{metamorphic}}$) declined over time. The crucial point is that key gases produced via metamorphic processes include H_2 and CH_4. The basic idea is quite simple: the hydrogen inherent in H_2 and CH_4, the latter through UV photolysis, escaped to space on account of its relatively light weight. This mechanism would have led to an oxidation of the Earth's crust. As a result, subsequent metamorphic activities would tend to release fewer quantities of reducing gases, thus engendering a reduction in $F_{\text{metamorphic}}$ over time and potentially driving the onset of the GOE. Some of the unknowns associated with this class of models are the efficiency of hydrogen escape and the coverage of continental crust at this stage.

Serpentinization (Section 2.7.1) is another important geological process that generates H_2. Hotter mantle temperatures in the past may have been indirectly responsible for increased serpentinization. If so, as the flux of H_2 generated accordingly declined over time with the cooling of the mantle, this possibly assisted in the onset of the GOE. Several hypotheses seeking to explain the rise in oxygen not unsurprisingly exhibit deep connections with biological functions. For instance, it was proposed that the magnitude of available nickel (Ni) declined over time, partly because of mantle cooling. As Ni plays a central role in the enzymes of methanogens, which made their appearance in Section 3.2, this trend could have contributed to declining methane fluxes, consequently facilitating the buildup of oxygen in the atmosphere.

Finally, a recent biogeochemical model presented in Laakso and Schrag (2017) merits a mention. This work builds on earlier proposals that a global

glaciation event—to wit, a Snowball Earth episode—might have preceded the GOE and was responsible for the onset of the latter by triggering a transient increase in O_2 levels. The transient increase takes place owing to a combination of two factors. The first is the extensive ice cover that serves to decouple the atmospheric oxygen from its sinks in the ocean and land, thus contributing to the increase in O_2. The second is the predicted enhancement in post-glaciation CO_2 levels that promote increased weathering of the continent, thus resulting in higher fluxes of phosphorus to the oceans. As phosphorus controls primary productivity over long timescales (see Section 7.6), this results eventually in higher organic carbon burial rates, consequently leading to a spike in the production of oxygen. The net result of this transient increase is that the atmospheric O_2 content shifts from one stable state (with oxygen $< 10^{-7}$ PAL) to another (with oxygen between 10^{-3} and 10^{-1} PAL). This framework is qualitatively similar to other biogeochemical models that are steadily gaining traction in the literature, which are predicated on the concept that multiple (meta)stable states exist, with the GOE instantiating the abrupt transition between two such states.

We will round off the analysis by furnishing an example of a dynamical system with more than one steady state by adopting the methodology in Knoll and Nowak (2017). Let us denote oxygenic and anoxygenic photosynthetic organisms by the subscripts O and A, respectively. In the case of the latter, we consider microbes that use Fe^{2+} because of the relatively high primary production associated with this pathway, as noted in the paragraph following equation (3.14). Let us denote the population density by X. The governing equations are

$$\frac{dX_A}{dt} = X_A\left(R_A C_A - \mathcal{K} X_T\right) + M_A \tag{3.30}$$

$$\frac{dX_O}{dt} = X_O\left(R_O - \mathcal{K} X_T\right) + M_O, \tag{3.31}$$

where M denotes the migration rates from ecological niches where the two species are not in competition with each other, and R denotes the reproduction rates for the two microbes. In the above equations, the extinction rate term $-\mathcal{K} X_T$ encapsulates the competition for scarce resources, such as the limiting nutrient phosphorus, and regulates the total abundance $X_T = X_A + X_O$. The first term on the right-hand side of (3.30) is proportional to the concentration of iron (Fe^{2+}), denoted by C_A, that is related to

the oxygen concentration (C_O) via

$$C_A \approx \frac{\mathcal{C}}{1+\zeta C_O}, \tag{3.32}$$

where $\mathcal{C}$ denotes the concentration in the absence of oxygen and ζ is a constant factor that controls C_A. In the formal limits $C_O \to 0$ and $C_O \to \infty$, we see that $C_A \to \mathcal{C}$ and $C_A \to 0$, respectively. To complete our system of equations, the oxygen concentration is regulated by

$$\frac{dC_O}{dt} = \mathcal{G} X_O - \frac{C_O}{\tau_O}, \tag{3.33}$$

where the first term on the right-hand side signifies the production of oxygen via oxygen photosynthesis and is therefore proportional to X_O and the second term denotes the sink term that is proportional to the quantity of available O_2. We will assume that $\mathcal{G}$ and τ_O are constant, although it is more appropriate to model them as slowly varying functions of time.

We leave it as an exercise for the reader to determine the steady states by setting the right-hand side of (3.30), (3.31), and (3.33) to zero. The stability of these points can be duly analyzed by applying the standard techniques of nonlinear dynamics. It turns out that the steady states for X_O obey a quartic (fourth-order) equation, implying that either two or four real-valued solutions could exist for this system. In contrast, if we consider the limit $M_i \to 0$, with $i \in \{A, O\}$, the results are much simplified. We end up with the sole nontrivial steady state:

$$X_O = \frac{1}{\zeta \mathcal{G} \tau_O}\left(\frac{\mathcal{C} R_A}{R_O} - 1\right); \quad \frac{X_A}{X_O} = \frac{\zeta \mathcal{G} \tau_O}{\mathcal{K} R_O}\left(\frac{\mathcal{C} R_A}{R_O} - 1\right)^{-1} - 1. \tag{3.34}$$

It is essential to recognize that the above system ought not be regarded as a wholly realistic and accurate model for the GOE. The reason is because the dynamical nature of multiple oxygen sources and sinks has not been fully incorporated into the framework, and neither has the coupling between the biogeochemical cycles of oxygen and limiting nutrients (e.g., phosphorus).[8]

8. Recent numerical models increasingly support the viewpoint that such phenomena can explain the rise in atmospheric O_2 without the need for invoking the likes of tectonic or biological switches (Alcott et al. 2019; Ozaki et al. 2020).

Nonetheless, this toy model suffices to illustrate how multiple steady states could be readily manifested in fairly simple dynamical systems, and might provide conceptual insights pertaining to the onset of the GOE.

3.4.3 Summary

We have argued that the rise in atmospheric oxygen (O_2) due to the evolution of oxygenic photosynthesis had a number of profound consequences for life on Earth. Subsequently, we discussed how oxygenic photosynthesis was followed by a distinctive increase in the oxygen content of Earth's atmosphere around 2.4 Ga, a transition widely known as the Great Oxygenation Event. Naturally, the immediate question that presents itself is whether the accumulation of oxygen is likely, or even inevitable, in the atmospheres of other worlds.

Although this question lacks a definitive answer as of now, it appears as though there are two fundamental requirements that ought to be fulfilled. The advent of oxygenic photosynthesis is regarded by virtually all researchers as an indispensable precondition for the rise in atmospheric oxygen on worlds where abiotic channels are not predominant. It may be argued that the second condition is that oxygen sources and sinks must evolve over time, as the ensuing dynamics would alter Δ_{O_2} in (3.28) accordingly. Given the range of mechanisms we have outlined in this context, it seems plausible that worlds with oxygenic photosynthesis would eventually witness an increase in atmospheric oxygen content prima facie. However, as we shall explore in Section 4.3.5, this is not necessarily the case for all planets and depends on the properties of the host star.

3.5 EUKARYOTES

It is safe to claim that eukaryotes are loosely synonymous with complex life on Earth; to be precise, complex life is principally eukaryotic in nature, even if not all eukaryotes exhibit the same degree of complexity. We have already encountered prokaryotes (bacteria and archaea), but eukaryotic cells are strikingly different in many respects. For starters, the genetic material in eukaryotes is stored in an internal compartment, the nucleus, that is separated from the material inside the cell (cytoplasm) by means of a double-layer membrane. The eukaryotic cell has many other membrane-bound

organelles, of which the mitochondria (and plastids) are the most important. Eukaryotic cells are, on average, larger than their prokaryotic counterparts by about an order of magnitude. Eukaryotic cells are also more flexible owing to the presence of a cytoskeleton made of protein filaments (e.g., microtubules) that plays a key role in maintaining and altering cell shape as well as facilitating cell division.

There are several other functional differences between eukaryotes and prokaryotes. For instance, the process of cell division unfolds very differently between these two groups. Eukaryotic cells are notably capable of sexual reproduction, which is essentially predicated on the rearrangement and exchange of genetic material. Eukaryotes are also distinguished by their propensity for complex regulation of gene expression, responsible for characteristics such as cellular differentiation, which is discussed in Section 3.6. The flexibility of eukaryotic cells alluded to earlier enables phagocytosis, the engulfment of solid particles (or even other lifeforms), thus giving rise to internal vesicles called phagosomes.

As a consequence of the aforementioned factors, there exists a veritable chasm in the cellular and functional complexity of eukaryotes and prokaryotes. This outwardly profound gap is further exacerbated by the fact that true intermediates between eukaryotes and prokaryotes, dating from the early stages of eukaryote evolution, have not been documented in depth.[9] Hence, explaining how, why, and when the origin of eukaryotes (eukaryogenesis) occurred is highly challenging, owing to which there is no single theory that can be regarded as being definitive, complete, and incontrovertible. Our definition of *eukaryogenesis* encompasses all of the evolutionary events that took place between the first eukaryotic common ancestor (oldest ancestor with its only living descendants being eukaryotes) and the last eukaryotic common ancestor (most recent ancestor of all living eukaryotes), dubbed FECA and LECA henceforth. Due to these inherent challenges, it must be acknowledged that our account is neither exhaustive nor set in stone.

---

9. In particular, the path from the first to last common ancestor of eukaryotes is poorly understood, and was termed an "event horizon" (Poole & Gribaldo 2014), although the semantics underpinning this phrase is not identical to what one encounters in the physics of black holes.

3.5.1 Mitochondria and eukaryogenesis

We will focus on mitochondria herein, as they are responsible for one of the deepest splits between hypotheses that seek to explain eukaryogenesis: Did mitochondria emerge early in eukaryotic evolution, or did they arise at a late stage? Mitochondria serve as the sites of respiration and ATP synthesis and have therefore been dubbed the "powerhouses" of eukaryotic cells. All known eukaryotes seemingly possess mitochondria or once had (and thereafter lost) them at some stage; the eukaryotic microbe *Monocercomonoides* constitutes an example of the latter. Some textbooks tend to associate mitochondria exclusively with aerobic respiration. This statement is not correct because anaerobic mitochondria have been unearthed, and so have other mitochondria-related organelles that function in the absence of oxygen, the best known of which are hydrogenosomes (Embley & Martin 2006).

In 2010, the importance of hydrogenosome-like organelles was underscored when they were extracted from the cells of purported obligately anaerobic metazoa (i.e., animals with the capacity for survival in *exclusively* anoxic environments) from the phylum *Loricifera* (Danovaro et al. 2010; 2016). However, this significant claim remains disputed (Bernhard et al. 2015). Another such example of an obligately anaerobic metazoan is the species *Henneguya salminicola,* a parasite from the phylum *Cnidaria,* which is believed to have lost its mitochondrial genome as well as the genes in the nucleus involved in the replication of mitochondria (Yahalomi et al. 2020); in this respect, it resembles the eukaryote *Monocercomonoides* broached in the prior paragraph.

Despite the structural and functional diversity of mitochondria and their related organelles, phylogenetic analyses strongly indicate that they share a common evolutionary origin. We shall therefore refer to these different organelles collectively as mitochondria henceforth. Most mitochondria and homologous organelles, with certain exceptions such as most hydrogenosomes, have retained their own genomes, even after endosymbiosis (defined in the next paragraph). A number of hypotheses have been advanced to explain this observation, of which the most promising is arguably based on the notion that genes must be situated in close proximity to their products (e.g., proteins) for ensuring direct and rapid control of their biosynthesis—the regulation of gene expression is particularly important in this instance because the encoded proteins play a key role in the smooth functioning

of the electron transport chain—namely, a series of intracellular protein complexes that governs the flow of electrons from donors to acceptors.

In endosymbiosis, the union of two cells (one inside the interior of the other) leads to the emergence of novel lineages. The most radical aspect of endosymbiosis is that this mode of evolution is not gradual and does not depend on the standard paradigm of point mutations. Endosymbiotic hypotheses have a rich and controversial history that dates back to the nineteenth century, but the most important early pioneer in this regard was the Russian biologist Konstantin Mereschkowsky, who proposed that plastids and the nucleus (which we address later) were derived from endosymbiosis (Mereschkowsky 1905). The idea that mitochondria were acquired via endosymbiosis might have been first propounded by Paul Portier and Ivan Wallin in the early twentieth century (cf. Sato 2019). Endosymbiosis was subsequently revived by Lynn Margulis in a seminal paper that advocated the separate origin of plastids and mitochondria via two independent endosymbiosis events (L. Sagan 1967). Margulis's work is notable for the prominent role that she accorded to endosymbiosis in shaping evolution on Earth and for her visionary synthesis of cellular, biochemical, and paleontological evidence into a cohesive whole (Lazcano & Peretó 2017). This justly celebrated paper was apparently rejected by more than a dozen journals, patently revealing that revolutionary hypotheses do not always receive the befitting reception from the scientific community. In the same year, Jostein Goksøyr published a lesser-known paper that also postulated a separate origin for plastids and mitochondria via endosymbiosis (Goksøyr 1967). A meticulous historical and scientific overview of this subject was furnished in Archibald (2014). Figure 3.6 illustrates how endosymbiosis, in tandem with other phenomena, probably played a vital function in the inception of eukaryogenesis.

There are many endosymbiotic hypotheses for the origin of mitochondria depending on the nature of the host cell and the endosymbiont (both of which can be either archaea or bacteria) and the mechanism by which the latter was assimilated in the former (Zachar & Szathmáry 2017; Roger et al. 2017).[10] Models also differ on the timing of this symbiosis, which

10. In fact, certain models assign an important role to viruses in the context of endosymbiosis, eukaryogenesis, and early eukaryote evolution (Koonin 2016; Forterre & Gaïa 2016); they are, however, excluded from our conspectus despite their possible relevance.

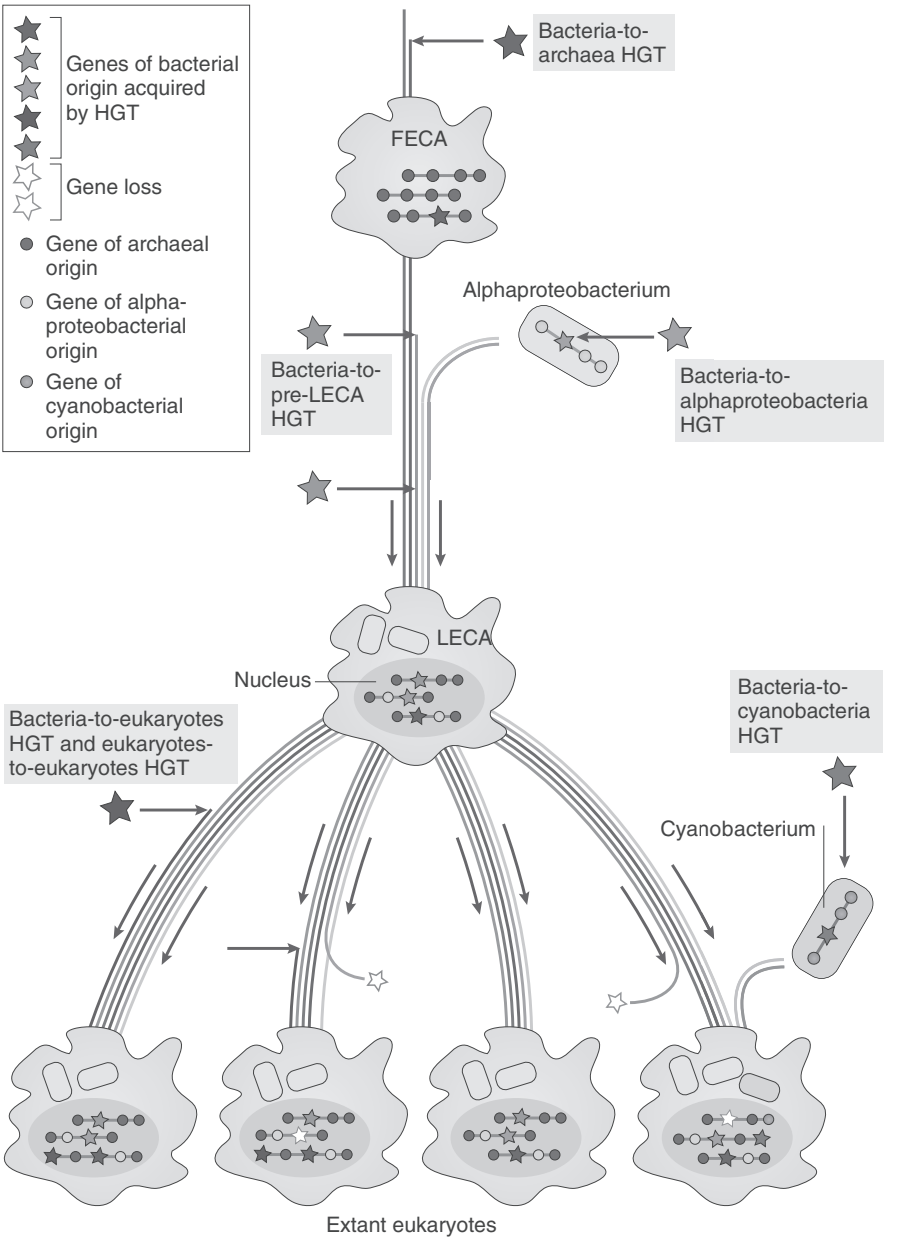


Figure 3.6 Scenarios for the incorporation of bacterial genes in eukaryotes starting with an archaeal host cell. LECA represents the last eukaryotic common ancestor, FECA is the first eukaryotic common ancestor, and HGT denotes horizontal gene transfer. While a number of genes are attributable to α-proteobacterial (mitochondrial) and cyanobacterial (plastid) ancestors, other genes were conceivably acquired via prokaryote-prokaryote, or rarely through eukaryote-eukaryote, HGT. (© Macmillan Publishers Limited. *Source:* Laura Eme, Anja Spang, Jonathan Lombard, Courtney W. Stairs, and Thijs J. G. Ettema [2017], Archaea and the origin of eukaryotes, *Nature Reviews Microbiology* 15: 711–7723, fig. 4.)

will be discussed shortly hereafter. The reader may consult W. F. Martin et al. (2015) for an in-depth review of endosymbiotic models for the acquisition of mitochondria, plastids, and the nucleus. Figure 3.7 presents the various endosymbiotic models that have been advanced to explicate the origin of mitochondria. However, while there exists broad consensus that mitochondria originated as a direct consequence of endosymbiosis, several dissenting viewpoints—both historically and even up to the present day—have argued in favor of an autogenous origin (i.e., "origin from within") for mitochondria and other organelles (Baum 2015).

One of the most comprehensive and famous "mitochondria-early" models of endosymbiosis was expounded by W. Martin and Müller (1998). As per this hypothesis, primitive eukaryotes sans mitochondria never existed. On the basis of the biochemistry of energy metabolism, the authors contended that the host was an archaeon that was anaerobic and strictly dependent on H_2 for the synthesis of organic products, whereas the endosymbiont was a facultative anaerobic bacterium. In other words, the latter possessed the capacity for both aerobic respiration (in the presence of O_2) and fermentation (anaerobic metabolism), with the latter releasing H_2 as a waste product. A central element of this hypothesis is that endosymbiosis was necessary because the organic molecules manufactured by the archaeon are consumed by the bacterium to generate hydrogen, which in turn provides the input for the archaeon to synthesize organic matter. Hence, the hydrogen hypothesis represents an example of (anaerobic) microbial syntrophy, an intricate process that has been characterized as "obligately mutualistic metabolism" (B. E. L. Morris et al. 2013). The hydrogen hypothesis appears to be consistent with the geochemical changes on Earth as well. As H_2 levels declined, partly on account of the GOE, the host may have clung closer to the endosymbiont for gaining greater access to the H_2 produced by it, eventually assimilating the latter.

One of the chief advantages claimed by this hypothesis is that it provides an explanation for the existence of both aerobic and anaerobic mitochondria (and related organelles). Moreover, the latest evidence from phylogenetics strongly supports the view that the host cell was archaeal in nature and that the endosymbiont was related to modern α-proteobacteria (a diverse class of bacteria with extreme variations in metabolic capacity and genome size) to some degree. With regard to the former, phylogenetic analyses indicate that a group (technically a superphylum) of archaea known as Asgard archaea (Zaremba-Niedzwiedzka et al. 2017; Spang et al. 2018; Castelle &

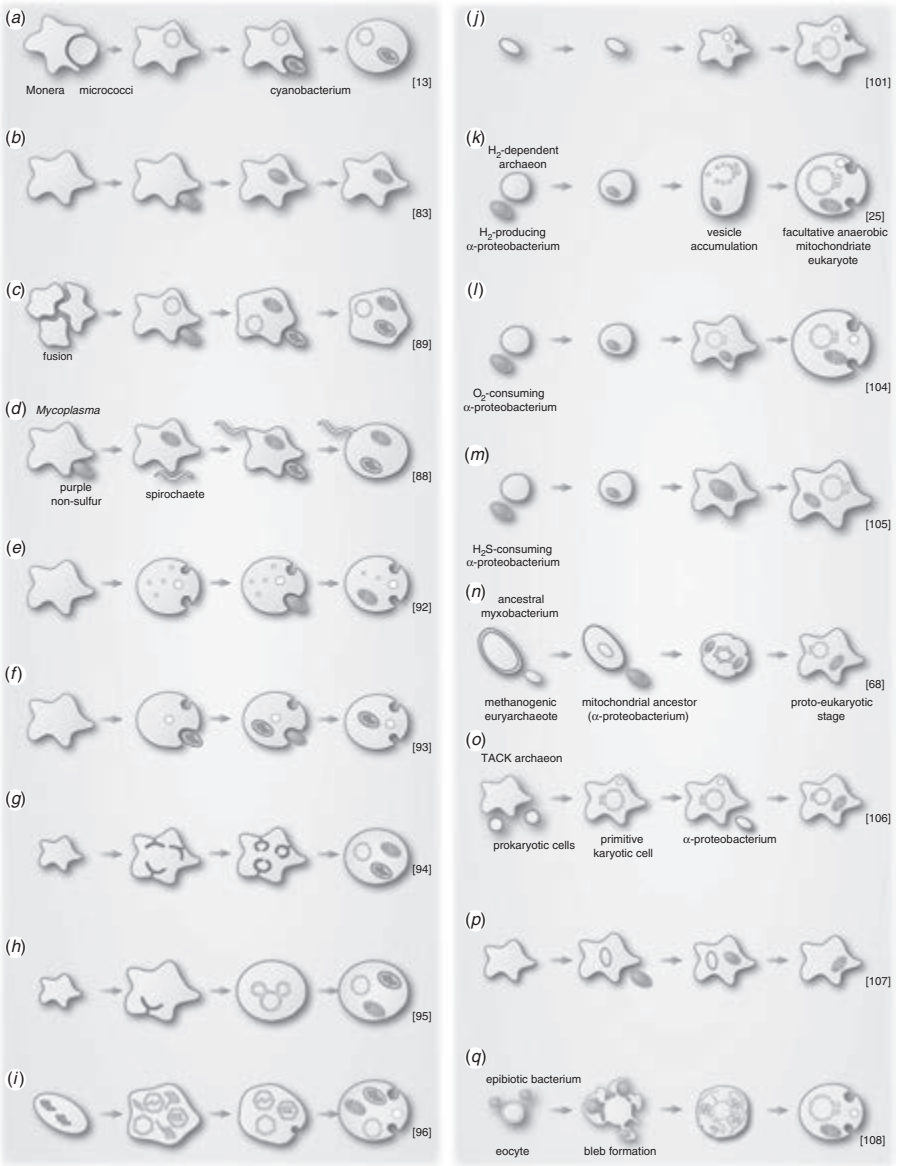


Figure 3.7 Myriad endosymbiotic models postulated for the origin of mitochondria and chloroplasts (sites of photosynthesis in plants). Cells / membranes are cast in different shades and patterns to indicate that they are derived from bacteria, archaea, and cyanobacteria; in some instances, the identity of the underlying cell is unclear. (© The Authors. CC-BY 4.0. *Source:* William F. Martin, Sriram Garg, and Verena Zimorski [2015], Endosymbiotic theories for eukaryote origin, *Philosophical Transactions B* 370[1678]: 20140330, fig. 2.)

Banfield 2018) are the closest known relatives of eukaryotes to date (cf. Da Cunha et al. 2017, 2018). Among the Asgard archaea, a subgroup (phylum) assigned the moniker Lokiarchaeota has attracted much attention, with comparative genomic evidence apparently suggesting that it was dependent on H_2, thus agreeing with the hydrogen hypothesis outlined above. In a similar vein, the phylum Heimdallarchaeota has been gaining increasing prominence because the state-of-the-art phylogenetic analyses indicate that it constitutes the deepest-branching member of the Asgard archaea and is therefore the closest relative of eukaryotes (Bulzu et al. 2019; Williams et al. 2020).

The cultivation of Asgard archaea in vitro has proven to be very challenging, but this barrier was overcome in 2020 when *Prometheoarchaeum syntrophicum* was isolated (Imachi et al. 2020). Prometheoarchaeum derives its energy through the degradation of amino acids by means of syntrophy. As Prometheoarchaeum possesses tentacle-like protrusions, it has been proposed that eukaryogenesis might have occurred through the twin processes of engulfment and endogenization, as explained a few pages hereafter. However, it must be cautioned that phylogenetic analyses are generally subject to much variability, and the current state of affairs will almost certainly undergo subsequent revisions. The reader may consult Eme et al. (2017) and López-García and Moreira (2020) for insightful overviews of eukaryogenesis as well as the attendant questions and uncertainties corresponding to this topic. One of the crucial unresolved issues is the metabolic nature of Asgard archaea: while the superphylum is versatile in this respect, the pathways associated with its respective phyla remain indeterminate (cf. Spang et al. 2019).

Mitochondria-early models, which posit that eukaryogenesis started with, or was closely followed by, the origin of mitochondria, have often invoked bioenergetic arguments in their favor (Lane 2017b). We will delineate the calculations presented in Lane and Martin (2010) and Lane (2015), but it is worth mentioning that the final results were contested by others (Lynch & Marinov 2017; Hampl et al. 2019). We will use the subscripts P and E for prokaryotes and eukaryotes, respectively. Let us denote the mass by m and the mass-specific metabolic rate—the rate of consumption of energy by an organism per unit mass (units of W / kg)—by $\mathcal{R}$. Finally, we denote the ploidy level (the number of copies of each genome in the cell) by $\mathcal{L}$ and the number of haploid genes by $\mathcal{N}$. The quantity of interest is the power per gene ($\mathcal{E}$), which is given by

$$\mathcal{E} = \frac{m\mathcal{R}}{\mathcal{L}\mathcal{N}}. \tag{3.35}$$

The important takeaway is that the power per gene serves as a heuristic measure of the energy available for protein synthesis and does not therefore imply that a higher value of $\mathcal{E}$ automatically translates to an increased genome size.

For an average prokaryote, we have $m_P \sim 2.6 \times 10^{-15}$ kg, $\mathcal{R}_P \sim 1.9 \times 10^2$ W / kg, $\mathcal{L}_P \sim 4$ and $\mathcal{N}_P \sim 5 \times 10^3$. Using this data in (3.35), we obtain $\mathcal{E}_P \sim 2.47 \times 10^{-17}$ W. It turns out that $\mathcal{E}_P$ is nearly constant for prokaryotes regardless of their size. Giant bacteria, such as *Thiomargarita namibiensis* and *Epulopiscium fishelsoni* (two of the largest known bacteria), reach masses of $\sim 10^{-9}$ kg, which is nearly six orders of magnitude higher than the fiducial value chosen earlier for m_P. Hence, based on (3.35), we may be inclined to think that $\mathcal{E}_P$ should increase accordingly. However, it has been shown that such gigantic prokaryotes have an extremely high ploidy number ($\sim 10^4$) that compensates for the increase in mass. Let us now repeat the same calculations for an average eukaryote with $m_E \sim 4 \times 10^{-11}$ kg, $\mathcal{R}_E \sim 60$ W / kg, $\mathcal{L}_E \sim 2$ and $\mathcal{N}_E \sim 2 \times 10^4$. For this choice of parameters, we end up with $\mathcal{E}_E \sim 6 \times 10^{-14}$ W. Thus, it is evident that eukaryotes have $\sim 2.4 \times 10^3$ more power per gene than prokaryotes. Another useful metric that can be calculated is the power per haploid genome, which is defined as $m\mathcal{R}/\mathcal{L}$. It is readily seen that this metric also increases by ~ 4 orders of magnitude as one moves from typical prokaryotes to eukaryotes.

The preceding analysis has not been corrected to account for size. For a true comparison, one must scale up a typical prokaryote to be the same size as that of a eukaryote. For reasons that we shall not get into, ATP synthesis scales with the area whereas protein synthesis scales with the volume. As the surface / volume ratio decreases roughly linearly with increasing size, larger prokaryotic cells will typically be less efficient than their smaller counterparts. If we increase the size by a factor Λ, to leading order, it may be assumed that the numerator of (3.35) scales as Λ^2, whereas the denominator exhibits a scaling of Λ^3. Hence, if one chooses $\Lambda \sim 50$ to reflect the discrepancy between a typical bacterium and protozoan (unicellular eukaryote), it is found that $\mathcal{E}$ for scaled-up prokaryotes might be five orders of magnitude lower than that of eukaryotes.

The importance of endosymbionts (and mitochondria in particular) can be attributed to the fact that they are in competition with each other inside

the relatively stable environment of the host cell. As one mechanism to compete successfully is replication, the loss of "superfluous" genes is advantageous as it facilitates faster replication; see Section 2.4.4 for a description of Spiegelman's paradox, which pertains to the same issue. Hence, endosymbionts probably underwent significant gene losses leading to the formation of organelles. At the same time, many of these genes would have been incorporated into the nucleus, giving rise to genomic asymmetry. The loss of unnecessary genes does not affect the cell from the standpoint of ATP synthesis, but it could provide major benefits vis-à-vis the synthesis of proteins. To see why this is case, the following thought experiment is useful.

We assume that a typical endosymbiont starts off with a typical genome comprising 4000 genes. Of this number, assume that 5 percent of them are lost because of reasons outlined above. In other words, about 200 genes, each encoding for a protein, are lost. Each bacterial protein has about 250 amino acids, and synthesizing one peptide bond (linking two amino acids) requires about five ATPs. Thus, we find that the loss of 5 percent of genes could translate to saving 2.5×10^5 ATPs. Now, we note that each endosymbiont has about 2000 copies of each protein and that there may exist ~ 100 endosymbionts per host cell. Multiplying all these factors, we see that the total energy savings is 5×10^{10} ATPs. Now, this quantity represents the energetic gain accrued during one cell cycle (~ 24 hours), implying that there the power savings are $\sim 5.8 \times 10^5$ ATP molecules per second. In order to synthesize 1 μm of actin filament, given that it forms a vital component of the flexible cytoskeleton in eukaryotes, $\sim 1.3 \times 10^5$ ATP molecules are required. Hence, from these relations, it is evident that the net savings in energy consumption can be repurposed for the production of ~ 4.5 μm of actin filaments per second.

Hitherto, we have focused primarily on the mitochondria-early models, but a multitude of models favor the idea that mitochondria arose late in eukaryotic evolution (mitochondria-late). Many of them revolve around the idea that the host cell had evolved phagocytosis endogenously and was therefore able to engulf the ancestor of mitochondria (Poole & Gribaldo 2014). Its evolution was attributed by some scientists to the ostensible advantages arising from predatory feeding (Cavalier-Smith 2009), while others view bacteriovory (bacteria-bacteria predation) as a causal factor for mitochondrial origin in its own right. Although phagotrophy or bacteriovory are presumed to serve as the central mechanism in a number of mitochondria-late models, the two concepts are not completely synonymous. This class

of models has a long history, since early hypotheses suggested that putative proto-eukaryotes equipped with highly evolved internal cell structures acquired mitochondria through phagocytosis. This notion has, for the most part, been considered untenable as the available data is purportedly far more consistent with the host cell being an archaeon.

However, in more recent developments, much attention has been directed toward the phagocytosing archaeon theory (PhAT), which suggests that the host cell was an archaeon capable of primitive phagocytosis (Martijn & Ettema 2013; Koonin & Yutin 2014). At this stage, it is important to appreciate that no evidence for prokaryotes (archaea and bacteria) capable of phagocytosis sensu stricto exists to date, with the seeming exception of the planctomycete bacterium *Candidatus* Uab amorphum (Shiratori et al. 2019), but the apparent lack of empirical data is not equivalent to reliable evidence for its absence. Moreover, the genomes of Asgard archaea appear to encode many proteins that are homologous (i.e., share common ancestry) with eukaryotes and were possibly responsible for similar functions.

Thus, there is mounting evidence that the archaeal host cell already possessed many essential components of eukaryotic cellular complexity, such as membranous compartments and vesicles, as well as filaments comprising the cytoskeleton (Roger et al. 2017). Hence, these developments led Zaremba-Niedzwiedzka et al. (2017), in their pioneering study of Asgard archaea, to propose that "*these findings may suggest the ability of simple phagocytic capabilities in the archaeal host*" [emphasis added]. This hypothesis is compatible to an extent with the discovery of tentacle-like structures in the Asgard archaeon *Prometheoarchaeum syntrophicum* (Imachi et al. 2020). Lastly, Pittis and Gabaldón (2016) argued that the distance between mitochondrial proteins and the ones found in their closest prokaryotic relatives is relatively short compared to other bacteria-derived components of eukaryotes, implying a late mitochondrial origin, but their analysis has been disputed.

Regardless of which class of endosymbiotic models for the origin of mitochondria is correct, it is mostly agreed that the last universal common ancestor did possess mitochondria. While there is less consensus regarding the significance of this event, the majority of mitochondria-early (mito-early) and mitochondria-late (mito-late) hypotheses accept that the origin of mitochondria was important from the standpoint of eukaryogenesis. Recent developments have provided strong support for the host cell being archaeal in nature, thus favoring only two domains of life (bacteria and archaea) as opposed to the classic three-domain system that included eukaryotes as

a separate branch in the Tree of Life. Moving past these broad statements, evidently a great deal remains unknown regarding eukaryogenesis, especially with regard to the origin and timing of mitochondria. It is instructive to delineate a few examples below to illustrate the schisms that exist.

If the host cell was an archaeon, as indicated by recent developments, one of the prominent mysteries concerns why the cell membranes of eukaryotes resemble those of bacteria instead of archaea (Lombard et al. 2012; López-García & Moreira 2015). Moving on, the expansion in the genome size from prokaryotes to eukaryotes must have been accompanied by a decrease in the mutation rate per base pair, known as the error threshold (defined in Section 2.8.1). If the proposed mechanism for this expansion in gene size occurred prior to the evolution of more refined replication mechanisms, it may have posed difficulties for sustainable reproduction of eukaryotic cells. Next, if we consider mito-early models, there is no direct evidence for prokaryote-prokaryote syntrophy resulting in obligate symbiosis, while the same can be said of phagocytosis in prokaryotes when confronted with mito-late models. It has been argued on energetic grounds that phagocytosis prior to mitochondria was unlikely (W. F. Martin et al. 2017; D. B. Mills 2020), whereas other studies have contested the proposal that mitochondria were necessary for amplifying the bioenergetic capacity of eukaryotes.

3.5.2 Plastids, nucleus, and other organelles

The nucleus is the largest organelle in eukaryotic cells that is characterized by a highly complex structure enclosed within a double-layered membrane. Best known for being the primary repository of genetic material, it plays a central role in controlling the growth, metabolism, and reproduction of the cell. In addition, it serves as the site for assembling ribosomes that are themselves the sites of protein synthesis. There are a number of hypotheses as to how the nucleus may have originated. Broadly speaking, they can be classified into (1) inward folding (invaginations) of plasma membranes in prokaryotes, (2) autogenous origin after the acquisition of mitochondria, and (3) independent endosymbiosis. The reader is referred to Cavalier-Smith (2010) for an in-depth exposition of (1), whereas W. F. Martin et al. (2015) conducted a thorough survey of (2) and (3).

Several hypotheses involving (1) are predicated on the premise that a prokaryote acquired the property of phagocytosis by losing its cell wall.

Ribosomes became internalized and were affixed to a primitive internal membrane that eventually evolved to become the nuclear envelope. The basic idea behind (2) is that mobile ribozymes (RNA enzymes) known as group II introns could have spread to many sites in the host's chromosomes (repositories of the genome) from the mitochondrial endosymbionts. The introns (DNA / RNA segments that do not encode for proteins) that would have inserted themselves into the host genome had to be excised via macromolecular machines known as spliceosomes, which may have evolved from group II introns themselves. However, there is a basic problem that these hypothetical cells would encounter. The rate of removing introns through spliceosomes is several minutes per intron, whereas protein synthesis by ribosomes occurs at a rate of two proteins per minute.

Thus, if a physical barrier can be erected to isolate ribosomes, this setup would enable the spliceosomes to complete their job prior to protein synthesis by ribosomes. It has been proposed that this development initiated the series of steps that led to the origin of the nuclear membrane. The formation of the nuclear membrane would have entailed many failures, but it may have garnered assistance from the transfer of genes encoding for membrane (lipid) synthesis in the bacterial endosymbiont to the host archaeon. The chaotic and enhanced synthesis of lipids within the host cell due to gene transfer could have therefore resulted in the synthesis of an imperfect, permeable, membrane that eventually evolved to become the nuclear membrane. We will not discuss (3) herein as the proposed models are numerous and laden with technical details.

The importance of plastids stems from the fact that they are responsible for the synthesis and storage of organic compounds in autotrophic eukaryotes. The best-known plastids are chloroplasts, which are responsible for carrying out photosynthesis in plants and algae. Unlike the mitochondria, whose origins are mired in controversy, it is almost universally accepted that plastids arose through some variant of endosymbiosis: the host cell was a eukaryote that already had mitochondria, whereas the endosymbiont was a cyanobacterium. This primary endosymbiosis was responsible for the emergence of land plants (*Embryophyta*) as well as green, red, and glaucophyte algae (collectively composing the group *Archaeplastida*). Just as with mitochondria, many genes were transported from plastids (endosymbiont) to the chromosomes of the eukaryote (host). From a metabolic standpoint, it is believed that the cyanobacterial endosymbiont provided either oxygen (from oxygenic photosynthesis), carbohydrates (via carbon

fixation), nitrogen fixation, or some combination thereof. The chief difference between mitochondria and plastids is that the former appears to have been characterized by a single endosymbiotic event (Lane 2017b), whereas the situation with regard to the latter is complicated by the putative existence of multiple endosymbiotic events (primary, secondary, and perhaps even tertiary) that are necessary for explaining the observed diversity of plastids (Archibald 2015; Hadariová et al. 2018).

Endosymbiotic hypotheses have also been delineated for other organelles in eukaryotic cells such as the endoplasmic reticulum (continuous membrane system), peroxisome (contains enzymes for breaking down fatty acids), and the flagellum (whiplike structure for locomotion). However, the evidence in favor of the endosymbiotic origin of these organelles is rather unconvincing. The endomembrane system—the network of cell membranes within the cell interior including the endoplasmic reticulum and the nuclear membrane—was recently proposed by S. B. Gould et al. (2016) to be derived from the formation of vesicles synthesized by the mitochondrial endosymbiont. This hypothesis may also explain why eukaryotic cells have bacterial membranes. On the other hand, the analysis of Asgard archaea does not seem to support this view; instead, the majority of early vesicular machinery was probably archaeal in nature (Dacks & Field 2018).

3.5.3 Endosymbiosis in nature

There are sufficient grounds for supposing that endosymbiosis is not uncommon on Earth, as we shall discuss below. For additional details pertaining to this subject, the reader may consult Estrela et al. (2016) and López-García et al. (2017).

To begin with, there are countless examples of endosymbiosis between prokaryotes and eukaryotes. For example, α-proteobacterial endosymbionts that belong to the genus *Wolbachia* confer advantages such as virus resistance, increased fecundity (more offspring), and metabolic provisions to various insects. The bacteria *Buchnera aphidicola* assists in the synthesis of certain amino acids that are not produced by its host—namely, aphids (insects that suck sap). A number of marine animals are known to have endosymbiotic relationships. For instance, several sponges contain symbionts from the phylum *Poribacteria* that assist in carbon fixation and synthesize a number of unusual proteins. Moving deeper into the sea, γ-proteobacterial symbionts are found in *Riftia pachyptila* (giant tube worms). While the worms provide

H_2S and O_2, the bacteria oxidizes H_2S to gain energy and fix carbon that is used by the worm.

Although we have selectively highlighted instances of endosymbiosis in the animal kingdom, we note that corresponding examples for plants and unicellular eukaryotes also exist. In numerous cases, we observe that the associated endosymbionts are distinguished by tiny genome sizes on account of gene losses. Perhaps the most widely known among them are *Rhizobiales* from the class of α-proteobacteria that fix nitrogen and are found in the roots of the plant family *Fabaceae* (legumes). In the case of unicellular eukaryotes, most of the known endosymbioses involve nitrogen-fixing bacteria, cyanobacteria, and methanogenic archaea. In certain circumstances, it has been shown that endosymbiosis between prokaryotes and eukaryotes is characterized by more than two partners (possibly reaching even five). It is believed that multiple symbioses confer the benefits of synergistic metabolisms in environments lacking in oxygen.

We have already encountered microbial syntrophy in Section 3.5.1. While this phenomenon represents a form of symbiosis, it is not endosymbiosis strictly speaking. Nonetheless, the symbiotic partners are sometimes so tightly integrated that only the pair can be investigated (i.e., the individual partners cannot be studied). The photosynthetic consortium *Chlorochromatium aggregatum* comprises green sulfur bacteria (anoxygenic photosynthetic autotrophs) surrounding a motile central heterotrophic β-proteobacteria. Another famous example in this realm is the syntrophy between δ-proteobacteria capable of anaerobic fermentation and methanogenic archaea. A striking instance of syntrophy entails anaerobic methane-producing and methane-consuming microbes that exchange electrons over long distances (up to 10^{-2} m) via "cables" (Lovley 2017). Due to the diversity of the microbial world, we have only begun to scratch the surface of syntrophic interactions, and countless more discoveries surely await us in the future.

However, we are primarily interested in knowing this: Is endosymbiosis involving two prokaryotes feasible? The question is relevant since this situation corresponds most closely to how eukaryogenesis is anticipated to have taken place. The answer is yes, although such examples are quite rare. The majority of examples we sketch below fall under the somewhat unusual category of nested symbiosis: an endosymbiont within a host that is itself an endosymbiont inside another cell. The mealybug *Planococcus citri* (an insect) contains the γ-proteobacterium *Moranella endobia* that resides within the

cytoplasm of the β-proteobacterium *Tremblaya princeps*. The mitochondria (itself an endosymbiont) of the hard tick *Ixodes ricinus* (an arachnid) contains an unusual bacterium *Midichloria mitochondrii*. A bacterium akin to the parasitic *Bdellovibrio* (belonging to δ-proteobacteria) was found inside cyanobacteria that were endosymbionts of the coral reef sponges *Neofibularia irata* and *Jaspis stellifera*. Finally, another case of intracellular bacteria has been documented within the cyanobacterium *Pleurocapsa minor*.

3.5.4 Sexual reproduction

Sexual reproduction in eukaryotes involves the fusion of two haploid cells (gametes) to form diploid cells, and the latter in turn give rise to haploid cells via a process known as meiosis. Haploid cells contain one set of chromosomes, whereas diploid cells contain two sets; recall that chromosomes contain the genetic material of organisms. Thus, while sexual reproduction is not synonymous with meiosis, it is nonetheless important to understand the origins and significance of the latter in order to comprehend the former. Recent overviews of meiosis and its connections with early eukaryotic evolution are found in Goodenough and Heitman (2014) and Lenormand et al. (2016).

In broad terms, meiosis starts with a diploid cell (two copies of chromosomes), undergoes DNA replication (effectively yielding four copies) and genetic recombination, and is finally subjected to two rounds of cell division to yield four cells (each with one copy). Genetic recombination, one of the most important aspects of meiosis,[11] entails the pairing of homologous (i.e., identical / similar) chromosomes and exchange of genetic material; this shuffling may sometimes involve the splitting and rejoining of DNA strands. A crucial point here is that the capacity for genetic recombination is not unique to eukaryotes. In fact, it has long existed in prokaryotes, where it plays a vital role in repairing harmful breaks and cross-links in double-stranded DNA and incorporating new genetic material received via horizontal gene transfer.

---

11. The transfer of genetic material is, however, not exclusive to meiosis—a series of steps known as the parasexual cycle exhibits properties akin to meiosis and has been documented in certain fungi.

Given this complexity, we can ask ourselves why sexual reproduction is even necessary.[12] The situation is exacerbated by the fact that organisms would only pass on 50 percent of their genes to the next generation via sexual reproduction, as opposed to nearly 100 percent through asexual reproduction. Several other costs are associated with sexual reproduction, especially in more complex organisms (e.g., finding mates). The most common answer offered in this context implies that the variation offered by genetic recombination is propitious from the standpoint of natural selection. This classic idea dates back to August Weismann, one of the pioneers of early evolutionary theory, in the late nineteenth century. There are many underlying subtleties and restrictions relating to this hypothesis, owing to which its validity is not universal (Otto 2009). Yet, under certain circumstances, genetic recombination could propagate advantageous variations by enabling beneficial alleles (variants of a gene) from different individuals to be brought together in the same individual.

In order to appreciate the advantages accorded by genetic recombination in meiosis and sexual reproduction, it is instructive to consider the toy model outlined in the seminal work by Nobel laureate Hermann J. Muller (1964). Let us suppose that we start with a particular mutant individual, who will be represented in the next generation by $(1+\kappa)$ individuals and in the second generation by $(1+\kappa)^2$ individuals and so on. Now, suppose that an average number of $\langle N \rangle$ descendants is necessary to ensure that a new advantageous mutation has transpired. We can calculate the number of generations (η) that must pass in order for $\langle N \rangle$ to be achieved as follows:

$$\langle N \rangle = \sum_{k=0}^{\eta} (1+\kappa)^k = \frac{(1+\kappa)^{\eta+1} - 1}{\kappa}, \tag{3.36}$$

Inverting this equation and solving for η, we obtain

$$\eta = \frac{\log\left(1+\kappa\langle N \rangle\right)}{\log\left(1+\kappa\right)} - 1. \tag{3.37}$$

12. An equally relevant question is why organisms do not always switch between asexual and sexual reproduction, as this strategy is optimal under many circumstances. We will, however, not address this intricate question here.

In the lower limit of $\kappa \to 0$, we end up with $\eta \to \langle N \rangle - 1$, whereas choosing an upper limit of $\kappa \to 1$ yields $\eta \to \log_2 (1 + \langle N \rangle) - 1$. At low values of κ, the number of generations scales linearly with $\langle N \rangle$, whereas it weakens to a logarithmic dependence at high values of κ. It is therefore evident that η will vary by several orders of magnitude for large values of $\langle N \rangle$. Hence, provided that genetic recombination boosts the value of κ, the benefits accruing from sexual reproduction relative to asexual reproduction sans meiosis may prove to be significant.

An equally significant point highlighted by Muller (1964, among other pioneering publications) is that deleterious mutations will tend to accumulate in asexual lineages over the course of many generations. This is because the absence of genetic recombination does not allow for the possibility of lowering the cumulative mutational load beyond a certain value, if one excludes "random" mechanisms such as horizontal gene transfer. This apparently inexorable and irreversible tendency for asexual organisms to amass harmful mutations over time is known as *Muller's ratchet* and has been canonically regarded as another long-term advantage associated with meiosis and sexual reproduction.

Apart from the above advantages linked with meiosis stemming from genetic recombination, several hypotheses suggest that it also played a major role in repairing DNA damage (analogous to prokaryotes) or suppressing its effects. In particular, it has been conjectured that two central characteristics of meiosis, homologous recombination and the reduction in ploidy (diploid to haploid), enabled DNA repair and eliminated dangerous mutations, respectively. Many candidates in this category share close connections with the acquisition of mitochondria via endosymbiosis; the reader can consult Hörandl and Speijer (2018) for a review of this subject. To offer an example, some models are based on the premise that the merger would have allegedly resulted in the contamination of the host cell's DNA by mitochondria, accompanied by a considerable increase in mutation rates through the formation of reactive oxgen species. The emergence of genetic recombination (in meiosis) may have provided an avenue for natural selection to preserve beneficial mutations while purging deleterious ones.

Alternatively, given the higher rate of mutations observed in mitochondrial DNA, it was perhaps necessary for the nucleus to adapt accordingly, as otherwise the mismatch would have grown over time and resulted in the malfunctioning of the cell. In this respect, genetic recombination could have assisted in the generation of sufficient variation, thereby allowing

selection to operate and mitigate this discrepancy in mutation rates. Another possibility is that the acquisition of the mitochondrial endosymbiont presumably resulted in the greatly enhanced production of ROS (adumbrated in Section 3.4.1). As these species are harmful to DNA, it has been proposed that meiosis (and sex) arose to facilitate DNA repair (Mirzaghaderi & Hörandl 2016). The origin of mitochondria, by means of endosymbiosis, and its attendant consequences (e.g., formation of ROS) have even been invoked as a potential explanation for the origin of distinct biological sexes—involving the transition from similar to dissimilar gametes—and why most organelles (including mitochondria) are commonly transmitted via uniparental inheritance (i.e., through only a single parent).

The last point that we need to address is whether sexual reproduction is a necessary requirement for complex multicellular life. For the reasons outlined earlier (e.g., Muller's ratchet) asexual populations are conventionally expected to be short-lived. However, there are fascinating examples of animals and plants on Earth that have purportedly reproduced asexually for a long time, thereby earning the name *ancient asexual scandals*. The best known of them are the bdelloid rotifers (class *Bdelloidea*), microscopic animals of size ~ 0.15–7 mm. These remarkable animals, believed to have originated at least 4×10^7 years in the past, have been demonstrated to possess foreign genes acquired via HGT. There is also evidence that bdelloids exchange genes with each other through the same mechanism.[13] Other ancient asexuals include oribatid mites (a group of arachnids) and possibly darwinulid ostracods (a group of freshwater crustaceans); in particular, asexual species from the former category may have originated several 100 Myr ago. When it comes to plants and fungi, numerous species have been reproducing asexually over long timescales, owing to which we shall not delve into specific cases herein.

3.5.5 When did eukaryotes evolve?

In light of the manifold differences between eukaryotes and prokaryotes, distinguishing the former from the latter in (micro)fossils is relatively feasible in principle. On the other hand, it is plausible that there were many stages

13. However, a recent study indicates that bdelloids may take part on extremely rare occasions in a form of sexual reproduction.

between FECA and LECA. Hence, one is confronted by the question of how eukaryogenesis should be dated, i.e., whether the FECA, LECA, or some intermediate stage should serve as the signpost for identifying the time at which eukaryotes originated. Another complicating factor to take into account is that early eukaryotes may be indistinguishable from their prokaryotic ancestors.

Let us first consider evidence from microfossils reviewed in Javaux and Lepot (2018). Sedimentary rocks from the Changcheng Group in North China have been found to contain well-preserved spheroidal fossils that exhibit fairly unambiguous evidence of morphological complexity, and their large sizes (60–200 μm) are potentially indicative of eukaryotic origin. Moreover, some of them display evidence of multiple-layered cell walls and cysts. These fossils were initially assigned a date of $\sim$ 1.8 Ga, but later studies have downgraded their ages to approximately 1.65 Ga. Other putative microfossils from the Mallapunyah Formation in Australia also date from roughly the same period. While these microfossils are widely accepted as being eukaryotic in nature, older candidates have been identified as eukaryotes, although the available evidence is both scanty and contested. The best-known example from this category is *Grypania spiralis,* discovered in the Negaunee Iron Formation in Michigan, which is dated to $\lesssim$ 2 Ga. These large fossils, characterized by their spiral structure, have been interpreted as photosynthetic algae. Tubular structures from South Africa dating back to 2.7–2.8 Ga are argued to be eukaryotic microfossils on the basis of morphological comparison with living eukaryotes and the organization of carbonaceous material in putative cell walls and interiors, but this study was critiqued and remains unsubstantiated.

Another line of evidence is predicated on the fact that eukaryotes synthesize distinctive organic molecules called sterols, of which the best known is cholesterol. These sterols are preserved in rocks in the form of compounds known as steranes. Thus, in principle, the discovery of such steranes would corroborate the presence of eukaryotes. A famous analysis of sedimentary rocks from the Pilbara Craton in Australia by Brocks et al. (1999) determined that steranes were detected in rocks that were $\sim$ 2.7 Gyr old. However, subsequent studies that relied on ultraclean drilling and sampling have demonstrated that these molecular fossils were probably much younger contaminants (French et al. 2015). Investigation of Transvaal Supergroup sediments dating from $\sim$ 2.5–2.7 Ga yielded steranes that were identified as being eukaryotic in origin, but this claim has attracted skepticism.

Next, we turn our attention to the constraints on eukaryogenesis placed by phylogenetic molecular clocks. There is fairly broad consensus that LECA came into being approximately 1–2 Ga (Eme et al. 2017), with more recent analyses favoring the range of $\sim$1.6–1.9 Ga (Porter 2020). For instance, Betts et al. (2018) concluded that the endosymbiotic origin of mitochondria occurred at some point in the interval $\sim$1.2–2 Ga and that modern eukaryotes originated $<$ 1.84 Ga. However, the error bars associated with molecular clocks generally increase as one travels further back in time, and several older studies inferred that the age of LECA was $>$ 2 Ga. As per the preceding evidence, we will adopt a fiducial value of $\sim$1.8 Ga for the origin of eukaryotes, although it must be recognized that this estimate may be revised by future breakthroughs.

Finally, let us turn our attention to plastids, which were probably acquired after the evolution of fully fledged eukaryotes. The earliest unambiguous fossil evidence for photosynthetic eukaryotes (and therefore plastids) dates from $\sim$1 Ga (Gibson et al. 2018). The fossil in question, from the Hunting Formation in Canada, was christened *Bangiomorpha pubescens* by Nicholas J. Butterfield and belonged to the group *Rhodophyta* (red algae). However, recent fossils dated to $\sim$1.6 Ga from the Lower Vindhyan Group in India have been interpreted as red algae on the basis of analyzing organelle-like structures (Bengtson et al. 2017). If these inferences are proven to be correct, plastids may have been incorporated into photosynthetic eukaryotes by $\sim$1.6 Ga. Molecular clock analyses indicate that the symbiotic origin of plastids may be traced to $\sim$1.1–1.8 Ga, although the common ancestor of *Archaeplastida* (introduced in Section 3.5.2) may have lived as far back as $\sim$1.9 Ga (Sánchez-Baracaldo et al. 2017). Hereafter we choose a fiducial value of $\sim$1.5 Ga for the acquisition of plastids by eukaryotes via endosymbiosis.

3.5.6 Eukaryogenesis beyond Earth

In view of the tremendous difference in complexity between prokaryotes and eukaryotes, at first glimpse the likelihood of this evolutionary leap seems very low. However, given our prior analysis, it appears as though the chief developments that promoted eukaryogenesis could have been the assimilation of mitochondria via endosymbiosis and the potential evolution of certain proto-eukaryotic features in Asgard archaea. To this list, one may potentially add sexual reproduction. If we accept that each of these three

features was essential, the prospects for eukaryogenesis come across as rather dim since all of them were apparently singular events. Moreover, it should be noted that prokaryotic life probably existed by ~ 3.7 Ga, whereas the earliest evidence for eukaryotes is ~ 1.8 Ga: a yawning gap of nearly 2 Gyr that is roughly half the age of the Sun today.

However, it may very well be that our pessimism is unwarranted (Booth & Doolittle 2015). To begin with, our understanding of archaea is still at a nascent stage, and it is therefore conceivable that we will discover other phyla that display similar functional traits akin to those found in Asgard archaea or eukaryotes. Second, while endosymbiosis between two prokaryotes of the kind that led to the acquisition of mitochondria is undoubtedly rare, some examples have already been documented, as noted earlier. This lends credence to the possibility that events of this kind might not be uncommon. Finally, alternatives to sexual reproduction exist in plants, animals, and fungi, and there are tentative grounds to believe that the evolution of meiosis was related in some degree to mitochondrial endosymbiosis; if the latter is not very unlikely, perhaps the same also holds true for eukaryotic sex. To sum up, while there are compelling grounds for supposing that eukaryogenesis is a rare phenomenon, it is nonetheless possible that its probability of occurrence is not very low.

Before wrapping up our analysis, let us conclude with a thought experiment. We have seen that ~ 2 Gyr were potentially necessary to proceed from prokaryotic to eukaryotic life and that it has been ~ 2 Gyr since eukaryogenesis transpired. Hence, if a second instance of eukaryotic evolution was taking place, it could be happening around this time. What would such a cell look like? For starters, we can guess that it would be chimeric in nature, with traits from both prokaryotes and eukaryotes. In light of the presumed centrality of endosymbiosis in the emergence of eukaryotes, it ought to possess endosymbionts. Perhaps it would possess a nucleus-like organelle, but one enclosed by a single membrane instead of the conventional double membrane in eukaryotes.

The serendipitous discovery of the tantalizing microorganism *Parakaryon myojinensis* near a submarine volcano (Yamaguchi et al. 2012) appears to vindicate the above line of reasoning, although our theoretical predictions are admittedly inspired by this finding. Unfortunately, only a single specimen was sampled, so whether the observed cell evinced unique properties because of contamination or instrumentation errors is not known. However, if more organisms that exhibit genuinely chimeric traits are found,

they can pave the way toward gaining further insights into the likelihood of eukaryogenesis.

This leads us to a general point that arguably constitutes one of the central underlying themes of the book. Life on Earth is truly dazzling in its diversity. We have, for instance, probably sampled far less than one percent of all extant microbial species on our planet, to say nothing of extinct organisms that would be more numerous by orders of magnitude taken cumulatively. Hence, the sustained and sustainable exploration of Earth's rich biospheres (both past and current) is necessary, among countless other reasons, from the standpoint of gaining a better understanding of the paths available for biological evolution on other worlds.

3.6 COMPLEX MULTICELLULARITY

The Earth we see today is dominated by animals and plants that are virtually synonymous with complex multicellular organisms. However, as we shall see later, both these species took center stage nearly 1 Gyr after the origin of eukaryotes. This billion-year period, roughly ranging from 1.8 Ga to 0.8 Ga, has been somewhat inaccurately and unfairly termed the "Boring Billion."

We will begin with a brief overview of the many paths that led to simple multicellularity and then explore how complex multicellularity may have arisen. In the process, we will also delineate the events that led to the famous Cambrian explosion and the rapid diversification of plants and animals shortly thereafter.

3.6.1 The many roads to simple multicellularity

Before defining complex multicellularity in Section 3.6.2, it is worth commencing our analysis by designating the advantages associated with simple multicellularity and the paths by which it could have arisen. Multicellularity requires the conglomeration of cells as well as coordinating activity between cells; the latter is facilitated by intercellular signaling.

Before highlighting the positives attributed to multicellularity, it is worth noting that there are inherent disadvantages as well. For instance, extra energy must be spent on activities such as communication and adhesion between cells, due to which multicellularity leads to reduced freedom

of movement per cell. The collective group is more susceptible to a wide array of genetic conflicts and the advent of "cheaters"—individuals that do not partake in cooperation but seek to reap the ensuing rewards nonetheless. Although a number of potential detriments do exist, multicellular organisms have evolved several corrective mechanisms to combat these costs. One of the foremost among them is that the life cycle of these organisms typically begins with a unicellular stage that ensures that subsequent cells share a common descent, thereby enforcing genetic continuity and serving as a bottleneck. Thorough assessments of the manifold costs and benefits stemming from the inception of multicellularity are expounded in Grosberg and Strathmann (2007), Rokas (2008), Niklas and Newman (2013), and Rainey and De Monte (2014).

Let us now turn our attention to the advantages accorded by multicellularity. One of the foremost among them is an enhanced resistance to environmental stresses. It has been shown that multicellularity mitigates stresses arising from temperature, pressure, pH, desiccation, and metal toxicity. Multicellular organisms have also been demonstrated to be more adept at utilizing nutrients compared to their single-celled counterparts. For instance, choanoflagellates (single-celled eukaryotes that are close relatives of animals) are capable of forming multicellular aggregates that cast a wider net to capture their prey. Cable bacteria of the family *Desulfobulbaceae* derive sources and sinks of electrons by connecting together and forming filaments. Multicellular organisms have been shown to possess greater adroitness at collecting external resources that would have otherwise diffused away had the organisms been single-celled.

One of the major consequences of eukaryogenesis was phagotrophy—that is, the ability to engulf food particles. This function enabled the rise of microscopic predators and would have exerted selection pressure on prey to evolve corresponding strategies. The profound insight that species interactions play a crucial role in evolutionary change was articulated by Leigh Van Valen in his well-known paper (1973), where he dubbed this phenomenon the *Red Queen* effect, named after the character from Lewis Carroll's *Through the Looking-Glass, and What Alice Found There*. Multicellularity has been shown in several experiments involving bacteria to confer significant predation resistance. Multicellular organisms also gain advantage from their increased size, including the expansion of feeding opportunities and storing nutrients as reserves for the future. Finally, multicellular organisms regularly take part in the division of labor, thus allowing cells to carry

out complementary tasks (S. A. West & Cooper 2016). One of the best-known examples in this category is *Anabaena,* a genus of cyanobacteria capable of both photosynthesis and nitrogen fixation. However, as these two functions cannot be carried out in unison under normal circumstances, some individuals in aggregate filaments carry out photosynthesis and others perform nitrogen fixation.

Many paths were available for unicellular organisms to become multicellular in nature. One of the most obvious candidates in this regard is that single-celled organisms aggregated together to form colonies and eventually gave rise to multicellularity. As we have seen, one of the advantages associated with multicellularity is enhanced resistance to environmental stress. Hence, it is conceivable that unicellular lifeforms banded together to increase their chances of collective survival. In contrast to colonial aggregation, where multiple cells come together, it has also been proposed that multicellularity may have arisen from sustained divisions of a single cell, followed by either incomplete cell separation or post-adhesion. This idea has a long history and is arguably traceable to influential research by Ernst Haeckel in the 1870s. Studies of the choanoflagellate *Salpingoeca rosetta* appear to indicate that multicellularity arose via cell division and not through cell aggregation, but it is unclear whether the same finding applies to all other species that display facultative (i.e., optional) multicellularity.

Given the many advantages springing from multicellularity, and in view of the various possible routes for its origination, we must now look toward the available evidence to determine the likelihood of its emergence. For starters, laboratory experiments have strikingly demonstrated that the transition to multicellular-type states can occur very rapidly. One of the most famous examples in this category employed *Saccharomyces cerevisiae,* colloquially known as baker's yeast, as the model organism (Ratcliff et al. 2012). In a span of sixty days, these yeast formed clusters that reproduced via multicellular propagules (agents of reproduction such as spores and seeds), exhibited distinction between juvenile and adult phases, and materialized division of labor. All of these traits share deep connections with multicellularity. More recently, the single-celled green alga *Chlamydomonas reinhardtii* gave rise to multicellular structures in response to predation by *Paramecium tetraurelia* in $\sim$300 asexual generations. The eventual form of multicellularity manifested will plausibly depend on both the particulars of the unicellular organism and the specific selection pressures to which it is subjected.

Finally, multicellularity has evolved from unicellular organisms many times. Bacteria, in particular, are distinguished by a rich range of collective behavior that enable them to shuttle back and forth between unicellularity and facultative multicellularity, as further reviewed in Claessen et al. (2014) and N. A. Lyons and Kolter (2015). Myxobacteria (a subset of δ-proteobacteria) are one of the canonical examples of bacterial multicellularity. In environments with scarce nutrients, they swarm together and form aggregates. In due course, they develop into "fruiting bodies" that produce environmentally resistant spores, which are then dispersed into the surroundings. Although bacteria mostly evince facultative multicellularity, unique δ-proteobacteria known as multicellular magnetotactic prokaryotes (MMPs) might be obligately multicellular in nature (Lyons & Kolter 2015). The cells in MMPs are linked tightly through intercellular junctions, and they seem to reproduce through the fission of the whole organism, i.e., in the complete absence of a unicellular phase. These details should be viewed with due caution, since the microbes in question have not been studied in much detail.

Thus, several lines of evidence indicate that the transition from unicellular to multicellular organisms has occurred at least twenty-five times in bacteria and eukaryotes, and the total number may even exceed one hundred. There is evidence that the first appearance of multicellularity in bacteria may date back to ~ 2.5 Ga, although the available data are scarce and ambiguous. Hence, on the basis of the preceding discussion, there are strong grounds for supposing that multicellularity could originate with relative ease on other worlds, given the appropriate conditions. It is credible, therefore, that multicellularity constitutes a "minor" major transition.

3.6.2 The origins of complex multicellularity

The distinction between simple and complex multicellularity is not unambiguous, although it does entail quite a few distinguishing factors. To begin with, simple multicellularity is often manifested in the form of filaments, clusters, and spheres, whereas complex multicellular organisms are characterized by a much wider range of morphologies. Complex multicellularity also displays comparatively more elaborate intercellular communication and adhesion. More importantly, it exhibits sophisticated patterns of cell differentiation, giving rise to specialized cells with distinct functions by virtue

of complex genetic regulatory networks. Finally, virtually all cells in simple multicellular organisms come into direct contact with the external environment at one stage or another, while this is not the case for complex multicellular organisms due to their intricate three-dimensional organization. The reader should consult Knoll (2011), Niklas and Newman (2016), and Sebé-Pedrós et al. (2017) for thorough reviews of complex (especially animal) multicellularity.

The first prerequisite for complex multicellularity is the dynamic cytoskeleton that we encountered when discussing the surmised origin of eukaryotes from Asgard archaea (Stairs & Ettema 2020). Two other important aspects—namely, cell differentiation and death—appear to have arisen as a response to environmental stress and were probably present in early single-celled eukaryotes. Furthermore, as we have seen in Section 3.5.1, the loss of genes from organelles may result in significant energy savings that can be redirected for other purposes. Thus, at least some of the basic requirements for complex multicellular organisms already seem to have been existent in the majority of unicellular eukaryotes.

More specifically, cell adhesion in animals and plants necessitates a wide array of proteins alongside intricate signaling and communication mechanisms. The latter duo have been documented in protists (i.e., eukaryotes that are not plants, animals, or fungi) and are reliant on reactive oxygen species (Taverne et al. 2018). It is worth considering a few examples in detail; we shall focus primarily on animals because the corresponding studies for plants are fewer. The genes that encode for cell-cell adhesion proteins in animals are known to exist in choanoflagellates and early eukaryotes. Some of the signaling proteins necessary for intercellular communication occur in choanoflagellates as well as in the amoeboid protist *Ministeria vibrans,* which is closely related to animals. The gene families that guide the embryonic development of animals have been found in protists, although certain gene families are, not unsurprisingly, unique to the former.

As per the latest genome sequence analyses, the unicellular common ancestor of animals was sophisticated in its own right (Richter & King 2013; Brunet & King 2017; Erwin 2020). It ostensibly possessed a diverse repertoire of genes necessary for cell adhesion and signaling as well as transcriptional regulation (i.e., regulation of gene expression). Yet, at the same time, it is suspected that the materialization of new gene families and genetic regulatory networks was necessary for stimulating early animal evolution. One other major difference between pre-metazoans and animals may have

been that cell differentiation was merely temporal in nature and influenced by environmental cues in the former, whereas spatiotemporal cell differentiation controlled by endogenous factors was reified in the latter.

What does the phylogenetic and fossil evidence tell us about the origins of complex multicellularity? Several molecular clock analyses have converged on the timing of ~800 Ma (Myr ago) for the last common ancestor of animals (Erwin et al. 2011; Sperling & Stockey 2018), although a later origin at ~650 Ma cannot be ruled out (Cunningham et al. 2017). In particular, there are grounds for supposing that the divergence of sponges (members of phylum *Porifera*) could have taken place around this period. However, it must be stressed that many specifics regarding early animal evolution remain contested, including the question of whether *Porifera* or *Ctenophora* (colloquially called comb jellies) was the first phylum to diverge from the base of the metazoan branch in the Tree of Life (Laumer et al. 2019). In the case of land plants, the timing of the divergence from streptophyte algae (a group of green algae) is more variable, but certain studies suggest that it may have occurred as far back as ~0.7–0.9 Ga. From the standpoint of fossils, this period (~800 Ma) is characterized by a putative increase in eukaryotic diversity. However, the oldest known metazoan fossils are much more recent (~560–570 Ma), with one candidate sponge fossil dating from ~600 Ma, while the corresponding value for land plants is even more so (~470 Ma). There also exists indirect evidence for sponges, from the class *Demospongiae,* dating back to > 635 Ma based on the putative identification of sterane biomarkers.

Thus, it seems reasonable to presume, on the basis of the above evidence, that the origins of complex multicellularity might be traced back to ~800 Ma. If we accept this premise, we must ask ourselves why this particular timing was favored. The answer may partly stem from the possibility that the oceans transitioned from a sulfidic (rich in the sulfide anion) state to one that was rich in iron (ferruginous). Given that sulfide is toxic to many eukaryotes, the decline in its concentration could have released the barrier preventing the diversification of eukaryotic cells (Parfrey et al. 2011). Although this hypothesis provides an elegant explanation, it does not seem to be entirely consistent with recent geochemical evidence that favors a more heterogeneous oceanic composition.

There are reasons for believing that the diversification of complex multicellularity may have also been stimulated by abiotic feedback mechanisms. For example, an increase in O_2 levels, in principle, can translate to larger

organism size, either via pure diffusion or respiratory and circulatory systems. In addition, higher oxygen content would lead to oxidative stress, thereby potentially accelerating the formation of reactive oxygen species that play a key role in cell signaling and regulation. Hence, there exist some grounds for presuming that the radiation of complex multicellularity could have been linked to the oxygen levels in the atmosphere and oceans. It must, however, be appreciated that the causal relationship between the rise in oxygen content and rapid diversification of animals is not as simple as once thought (Cole et al. 2020), although the two events appear to have occurred partly in tandem (X. Chen et al. 2015); we will revisit this topic in Section 3.6.4.

Finally, we wish to highlight a crucial point. Even though animals and land plants are the most obvious examples of complex multicellularity, this characteristic has also evolved independently in two groups of fungi and algae (red and brown). Thus, it is conceivable, in light of its multiple origins, that the transition to complex multicellularity is not unlikely.

3.6.3 The significance of complex multicellularity

In order to truly appreciate the significance of complex multicellularity, a brief detour is worth taking. We will focus exclusively on land plants and animals here, although algae are clearly very important from the standpoint of fixing carbon and producing oxygen via photosynthesis; in doing so, they form an integral component of marine food webs.

A general observation is necessary: it is widely, indeed almost universally, accepted that life participates in the alteration of its environment and that the latter serves as a driving force behind natural selection. In the 1970s, James Lovelock proposed the Gaia hypothesis, which could be interpreted as the limiting case of the preceding statement. In a nutshell, as per the Gaia model, the biosphere plays a vital role in maintaining habitable conditions on Earth amenable to its sustained existence (for billions of years) via self-regulating feedback mechanisms (Lovelock 2000). We will not deal with the Gaia hypothesis herein, as its scientific basis is not conclusively established; a thorough, albeit skeptical, overview of this subject can be found in Tyrrell (2013).

The significance of plants hardly calls for much elaboration. The Earth contains around 5.5×10^{14} kg of organic carbon, of which plants comprise $\sim 4.5 \times 10^{14}$ kg, i.e., approximately 80 percent of the total biomass (Bar-On

et al. 2018). The primary production on Earth is also split almost evenly between landmasses and water bodies, with plants being responsible for carbon fixation in the former category. The advent of land plants has also been proposed as the causal mechanism behind the achievement of modern oxygen levels by ~ 400 Ma, although this surmise has not been wholly substantiated yet. Regardless of the actual timing, land plants probably played a key role in ensuring that oxygen levels became high enough to allow for the actuation of fire, as it requires an oxygen content of ~ 70 percent PAL. Among other factors, fire drives the evolution of plant characteristics, enhances biodiversity, may have contributed to the spread of flowering plants, and is potentially capable of regulating our planet's oxygen levels.

Apart from their obvious dual role in the production of oxygen and organic compounds (forming the basis of food webs), plants also affect Earth's modern-day environments in different ways. For example, recent studies indicate that rainfall over Amazonia is initiated by the rain forests themselves. In the broader context, an expansion in vegetation coverage has been linked with increase in the following variables: (1) albedo at near-infrared wavelengths, (2) humidity, (3) minimum temperature, (4) absorption of solar radiation at the surface, and (5) rainfall and cloud formation. On the other hand, increased vegetation coverage may lead to decreases in (1) albedo at visible wavelengths, (2) maximum temperature, (3) continental weathering, and (4) infrared emission. An overview of this subject can be found in McPherson (2007) and Beerling (2007).

Next, let us turn our attention to animals. An eloquent exposition of the manifold ways whereby animals operate as ingenious ecosystem engineers is found in Lewontin (2000). The reader may also consult Odling-Smee et al. (2003) and Butterfield (2011) for comprehensive discourses on this increasingly pertinent subject. First, due to animals being heterotrophs, i.e., reliant on external sources for food—they regulate the quantity of autotrophs (organisms that produce their own food). In doing so, they are responsible for the transformation of plant biomass, thereby accelerating nutrient recycling; the latter is governed by the distribution of animal feces. Moreover, as animals are motile, they play a major role in the dispersal and mixing of nutrients. The capacity for certain animals to feed on other animals (carnivory) shares close connections with changes in primary productivity and nutrient recycling.

Animals alter their environments in a number of ways, of which the best known perhaps is bioturbation. This process refers to the mixing of

sediments and soils due to biological activity, especially that instantiated by animals, since they are both macroscopic and motile. It has been argued that bioturbation regulates the composition of sediments, the topography of landscapes, and biogeochemical cycles. In addition, biomineralization (mechanisms used by organisms to produce minerals) modifies ecosystems through multiple channels (e.g., formation of coral reefs). In aquatic ecosystems, biomineralization also affects the sources and sinks of bioessential elements. Even in environments where the extent of ecosystem engineering by animals is not initially high, it can nonetheless lead to large changes in the biosphere as a consequence of nonlinear dynamical feedback mechanisms. A prospective example involves the transition from the turbid, cyanobacteria-dominated waters of the Proterozoic eon to the clear, algae-dominated waters of the Phanerozoic eon as a result of the perturbation induced by the water-clearing activities of filter-feeding zooplankton (Lenton et al. 2014).

Finally, the capacity for animals to indulge in carnivory ought to have sparked evolutionary arms races between predators and prey by means of the Red Queen effect, thus resulting in greater evolutionary diversity. Greater biological diversity is further connected to a host of crucial factors including ecosystem functioning, stability, and productivity. Thus, it is appropriate to envision animals not only as ecological but also evolutionary engineers. To sum up, it would not be mere hyperbole to suggest that Earth's modern biosphere was "invented" by animals (and plants), at least to some degree.

3.6.4 Factors involved in the rise of animals

Although the majority of laypersons will have heard of the famous *Cambrian explosion,* it is essential to recognize that the Earth's environment was very dynamic in the epochs leading up to the Cambrian. The first period of interest to us is the Cryogenian, which lasted from 720 Ma to 635 Ma. This epoch was characterized by two major global glaciation (Snowball Earth) events: the Sturtian (~ 717–660 Ma) and Marinoan (~ 645–635 Ma) glaciations. It must be observed, at this juncture, that Snowball Earth episodes are neither well understood nor even universally accepted. The epoch that immediately followed the Cryogenian was the Ediacaran (635–541 Ma). Although the significance of the Ediacaran vis-à-vis animal evolution has been appreciated since several decades ago, its importance is being increasingly underscored in recent times (Wood et al. 2019).

One of the striking discoveries from this epoch is the Ediacaran biota: enigmatic, typically soft-bodied organisms that exhibit body plans that are largely distinct from those expressed in modern-day animals, although some fossils display evidence of bilaterian-grade symmetry (approximate left-right mirror symmetry) akin to living animals. The putative animals that evolved during this period probably possessed active motility, complex reproductive strategies, and other ecological innovations evinced by fauna today.[14] The Ediacaran biota became extinct, for the most part (but not wholly), shortly before the Cambrian period (541–485 Ma). While the causes behind this mass extinction remain subject to much debate, it seems plausible that this process occurred through the gradual or pulsed replacement of Ediacaran biota by Cambrian lifeforms (or their precursors) due to the latter outcompeting the former via evolutionary innovations and ecosystem engineering. In toto, a persuasive case could be assembled for the thesis that the Ediacaran biota played an integral role in the subsequent radiation of bilaterians (comprising most animal phyla) and the Cambrian explosion. Reviews of the Ediacaran epoch, which encompass both its fluctuating environments and fascinating biota, can be found in Droser et al. (2017) and Muscente et al. (2018).

The Snowball Earth episodes appear to have been important for the eventual rise of animals for multiple reasons. It has been proposed that the end of the Sturtian and Marinoan glaciations was accompanied by high weathering rates, perhaps linked to melting glaciers, that led to increased nutrient (especially phosphate) influx into the oceans, consequently enabling the rise of algae in the oceans. In turn, the algae would have probably served as an excellent source of food for animals, thus facilitating their diversification. This surmise is borne out partially by analyses of sedimentary rocks that seem to indicate an increase in phosphate availability roughly during this period. This still leaves us with the question of how and why the influx of phosphorus underwent an increase in the late Neoproterozoic. The answer may be connected to the oxygen content and perhaps the global glaciation events (Brocks et al. 2017). The salient details

14. Motility in macroscopic organisms might have arisen much earlier. It has been claimed, for instance, that ~2.1 Ga biota from the Francevillian Basin in Gabon inhabited an oxygenated marine environment and were capable of lateral and vertical migration (El Albani et al. 2019). Further independent evidence is, however, necessary to properly substantiate this assertion.

underpinning this line of reasoning are mapped out in Knoll (2017), and the premise is further refined and expanded on in Laakso et al. (2020).

The potential causal links between global glaciation(s) and the GOE described in Section 3.4.2 are also possibly applicable here. The Snowball Earth episode(s) might have provided the requisite environmental perturbation for the planet to transition from the relatively low atmospheric and oceanic oxygen content in the mid-Proterozoic to a more oxic state during the late Neoproterozoic and Phanerozoic epochs. However, the central point worth appreciating here is not strictly dependent on global glaciations: regardless of the actual cause, the rise in oxygen levels may have regulated the phosphorus inventory within the oceans. This can be understood qualitatively as follows. In a predominantly anoxic environment, the higher abundance of Fe^{2+} (ferrous cation) suppresses the concentration of phosphates in the ocean by way of two distinct avenues. First, the scavenging of phosphorus by adsorption onto ferrous oxides in rivers decreases the influx of phosphorus into the oceans. Second, ferruginous (iron-rich) oceans would act as an efficient trap for phosphates through the formation of ferrous phosphates.

Let us now consider the situation when the oxygen levels rise. This scenario is predicted to oxidize iron, reduce the potency of the "iron trap," and thereby elevate the phosphorus influx and primary productivity of the oceans. The ensuing increase in primary productivity is expected to sustain a higher biomass and eventually result in more organic matter buried per unit time (i.e., higher organic carbon burial). The loss of organic matter through this channel prevents it from being consumed by respiration or oxidative decay, thereby effectively amounting to a net source of O_2, which was generated during the original biosynthesis of this matter. Thus, this series of developments could set up a positive feedback loop favoring a further increase in net oxygen production. Thus, on the basis of the picture we have delineated, oxygen content may have regulated the rise of animals not only on metabolic grounds but also through nutrient availability. The rise in O_2 levels was also capable, in principle, of sustaining increased carnivory that sparked the evolutionary arms races encountered earlier.

In summary, modeling the varying oxygen levels in Earth's oceans and atmosphere from the Cryogenian to Cambrian periods (spanning a total duration of ~ 200 Myr) requires us to meticulously account for the coupled biogeochemical cycles of oxygen, phosphorus, iron, and carbon (along with the other major bioessential elements) as well as the environmental

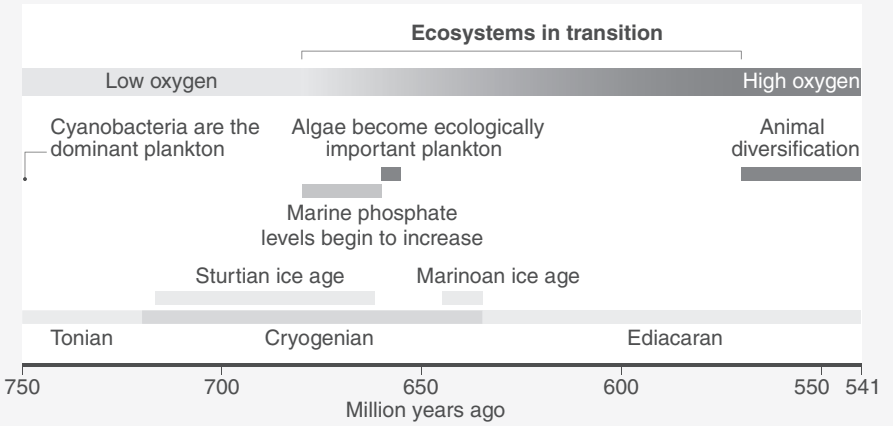


Figure 3.8 The major ecological and evolutionary changes that occurred during the Cryogenian and Ediacaran periods leading up to the Cambrian era. A combination of global glaciation events and enhanced oxygen and phosphorus concentrations may have contributed to the rise in algae and subsequent diversification of animals. (© Macmillan Publishers Limited. *Source:* Andrew H. Knoll [2017], Food for early animal evolution, *Nature* 548(7669): 528–530, fig. 1.)

fluctuations that unfolded over this protean interval (Reinhard et al. 2020). This already intricate picture is presumably further complicated by the necessity of having to integrate novel biological feedback mechanisms—such as predation and bioturbation, which came to fruition by dint of the interplay between biological evolution and elevated nutrient availability (Lenton et al. 2014)—into the global Earth system dynamics.

The transitional period encompassing the Cryogenian and Ediacaran epoch, and the changes that took place, is illustrated in Figure 3.8. In view of oxygen's significance, two immediate questions spring to mind: When did oxygen content in the atmosphere and oceans rise? And what were the causes behind it? The answers to both of these questions remain unresolved to an extent. We begin by noting that this second "rise" in oxygen, christened the Neoproterozoic Oxygenation Event (NOE), was probably protracted and non-monotonic in nature. Hence, evidence from selenium and chromium isotopes (based on changes in their oxidation states) indicates that the NOE might have commenced as early as $\sim$750 Ma. In contrast, the permanent oxygenation of the deep oceans and the achievement of modern

atmospheric oxygen levels may not have taken place until as late as ~ 380 Ma (i.e., well into the Phanerozoic eon), as per analyses of iron- and cerium-based proxies. Other studies that made use of multiple proxies (e.g., sulfur isotopes) have concluded that the Ediacaran was characterized by transient and multiple oceanic oxygenation events (OOEs) in an otherwise anoxic ocean. These OOEs could have stimulated rapid evolutionary innovations and speciation that punctuated longer intervals of stasis in accordance with the famous concept of "punctuated equilibria" explicated by Niles Eldredge and Stephen Jay Gould (1972).

It behooves us now to turn our attention to the second question—namely, the factors responsible for the increase(s) in oxygen levels. As noted above, the Snowball Earth episodes provide one possible and compelling reason. A summary of various biological and geological mechanisms proposed in this context is presented in Catling and Kasting (2017). With regard to biological factors, many are reliant on the enhanced burial of organic carbon, thus contributing to an effective increase in O_2 production. Some of the reasons that have been advanced are (1) the formation of fecal pellets by plankton and (2) increased O_2 production and nutrient supply into the oceans (driving carbon burial) via the greening of the land by plants. The geological proposals also span a wide spectrum, and a few select examples include (1) changes in continental configurations that increased organic carbon burial or release of nutrients to oceans, (2) relative decline in the reduction potential of gases released from the Earth (e.g., through volcanism), (3) subduction of sulfides deposited on the seafloor, and (4) hydrogen escape following the photolysis of atmospheric methane.

The last point that we wish to cover has to do with the renowned Cambrian explosion. This purportedly sui generis period in our planet's history, initiated ~ 540 million years ago (Ma) and unfolding over approximately 20 to 25 Myr, was characterized by the appearance of essentially all major animal phyla in the fossil record (Marshall 2006). Conventional narratives of Earth's history speak of the Cambrian explosion as though it were a singular event. However, our analysis up to this point should give us pause before subscribing to this viewpoint. First, certain components of the genetic tool kit involved in the assembly of animals, and the rudiments of central nervous systems, probably predate the Cambrian explosion to a considerable degree, perhaps on the order of 100 Myr (Erwin et al. 2011).

Second, we have seen how the protean environments of the Cryogenian and Ediacaran periods may have contributed to the rise in oxygen levels

and increased phosphorus availability that were conducive to the radiation of animals. Finally, it is conceivable that many of the requisite ecological interactions and evolutionary innovations arose in the late Ediacaran biota (Darroch et al. 2018). While these arguments seemingly offer strong reasons for envisioning the Cambrian explosion as a more protracted process that built on earlier developments in the late Neoproterozoic, it does not necessarily imply that the Cambrian explosion was wholly gradual either. Instead, the most plausible scenario is that the Cambrian explosion—one among several clustered metazoan radiations (Wood et al. 2019)—was fueled by a combination of new and preexisting eco-evolutionary innovations in labile environments.

3.6.5 A brief evolutionary timeline of the Phanerozoic eon

The Phanerozoic eon (541 Ma-present) gave rise to many important evolutionary innovations (summarized in Figure 3.1). However, in agreement with most analyses, none of them (barring one) arguably merit being classified as MEEs. In spite of this caveat, it is helpful to highlight certain striking evolutionary breakthroughs in this period, although we shall mostly forgo the Cambrian epoch (541–485 Ma) because it was previously explored. Accessible overviews of Earth's evolutionary history are quite prevalent in the literature: Prothero (2007), Benton (2015), Dawkins and Wong (2016), and Schulze-Makuch and Bains (2017) constitute a few representative examples. We note that the atmospheric oxygen content during the Phanerozoic did fluctuate, but the difference between the maximum and minimum levels was less than one order of magnitude.

In the Ordovician period (485-444 Ma), there were several noteworthy developments. Most of the oldest evidence for land plants ($\sim$ 470 Ma), such as deposits of spores, dates from this epoch, although earlier fossils from the Cambrian may come to light someday. More importantly, the Great Ordovician Biodiversification Event (GOBE), also known as the Ordovician radiation, unfolded over a time span of $\sim$ 30 Myr (Servais & Harper 2018). The GOBE is far less appreciated than the Cambrian explosion in certain quarters, but it was nearly as remarkable in its scope and consequences. The Ordovician radiation engendered a significant increase in marine diversity, led to major turnover of the marine biosphere, and was responsible for the genesis of numerous species (and other taxonomical units). It represents one

of the most profound series of biodiversification events in the history of our planet.

Finally, one of the *Big Five* mass extinctions took place at the end of the Ordovician period. The Late Ordovician mass extinctions, also called the Ordovician-Silurian (O-S) extinction events, occurred in discrete pulses and led to the extinction of ~ 85 percent of all marine species. As with most of the major mass extinctions, a number of putative causes have been identified, but there is no clear consensus at this stage. Some of the candidates for the O-S extinctions include cooling induced by glaciation, ocean euxinia (sulfidic and anoxic waters), changes in plate tectonics, and catastrophic astrophysical phenomena such as Gamma Ray Bursts (GRBs). A comprehensive description of the Phanerozoic mass extinctions— including the Big Five—and their proposed drivers can be found in Bambach (2006), D. P. G. Bond and Grasby (2017), and Clapham and Renne (2019).

In the Silurian period (444–419 Ma), unambiguous evidence for vascular land plants appears in the fossil record (Knoll & Nowak 2017). Vascular plants have specialized tissues for transporting water and food throughout the organism. This epoch also contains fossil evidence for bony fish—that is, with skeletons comprising bone tissue. Concrete examples of land animals from the subphyla *Myriapoda* (e.g., centipedes) and *Chelicerata* (e.g., arachnids) also appear in the Ordovician-Silurian record, with possible origins in the Cambrian era. The Silurian was followed by the Devonian (419–359 Ma) era, which has been christened the Age of Fish for self-evident reasons. However, a more notable aspect of this epoch was the radiation of vascular land plants, thus resulting in the formation of extensive forests. It is also believed that the ancestors of four-limbed vertebrates (tetrapodomorphs)—probably of amphibious nature—originated in the Devonian era when lobe-finned fish (e.g., lungfish), which bear the scientific moniker *Sarcopterygii,* evolved proto-limbs (Daeschler et al. 2006; Clack 2012). Toward the end of this period, one of the Big Five extinctions transpired. It is presumed that ~ 70 to 82 percent of species became extinct during this Late Devonian extinction. A number of potential reasons have been advanced, including asteroid impacts, volcanism, and environmental changes triggered by the radiation of vascular plants.

The Carboniferous period (359–299 Ma) followed the Devonian. One of the unique aspects of this era is that the oxygen levels were much higher than today in all likelihood (perhaps the highest in Earth's history), plausibly reaching 35 percent of the atmosphere by volume. The first reptiles,

which would have laid their eggs on land, appear to have originated in the Carboniferous (Benton 2015), with one of the oldest fossils (*Hylonomus*) dating from > 300 Ma. Fossils corresponding to the ancestral group of mammals (synapsids) also date from the Carboniferous period, as do the progenitors (or early variants) of octopodes; one such example that has been debated is *Pohlsepia mazonensis*. The next geological epoch was the Permian (299–252 Ma). This period is noteworthy for having given rise to the cynodonts—possibly warm-blooded animals that subsequently evolved into mammals—as well as the ancestors of dinosaurs (archosauromorphs).

However, the most striking event in the Permian period occurred at its very end ($\sim$ 252 Ma), namely the Permian-Triassic (P-T) extinction event, colloquially known as the *Great Dying*. The P-T transition was characterized by the extinction of $\gtrsim$ 80 percent and $\sim$ 70 percent of marine and terrestrial animals, respectively. While there remains little doubt that a large majority of animals perished in the P-T transition, the evidence for an equivalent extinction of land plants is equivocal (Nowak et al. 2019). Although a number of factors have been propounded for the Late Permian extinction, the primary trigger was probably massive Siberian Trap volcanism. The consequences of massive volcanism are predicted to be severe and varied: release of noxious volatiles, global warming, ocean acidification, and anoxia in surface waters (Knoll et al. 2007). The severity of biodiversity losses during the P-T extinction event was so pronounced that it may have taken $\mathcal{O}(10)$ Myr for terrestrial ecosystems to recover. Many ecological niches are sparsely occupied in the aftermath of mass extinction events, thereby resulting in large-scale radiation of species and colonization of these niches; this trend is generally, but not always, manifested.

The Triassic period (252–201 Ma) is unique in that its beginning and ending were both accompanied by mass extinction events. Due to the preceding P-T extinction event, the recovery of marine and terrestrial ecosystems was slow during the early Triassic. By the late Triassic, close relatives of mammals (e.g., members of the genus *Haramiyavia*) appear in the fossil record (Luo et al. 2015). Fossils of early dinosaurs, such as *Eoraptor* (from $\sim$ 230 Ma), have also been found in the late Triassic. A series of Triassic-Jurassic extinction events brought the chapter of the Triassic period to a close, in the process wiping out $\sim$ 76 percent of marine and terrestrial species. A number of hypotheses have been explored for this extinction, but it seems increasingly likely that it was connected to massive volcanism within the Central Atlantic Magmatic Province.

The Jurassic period lasted from 201 to 145 Ma and is (in)famous in popular culture because of the diversity and dominance of dinosaurs. However, the Jurassic period is also important from the standpoint of the evolution of early mammals. This is because the first eutherians—a group of mammals that includes all living placentals—have been dated to this epoch, such as the shrewlike *Juramaia* (Luo et al. 2011).[15] The ancestors of birds also diverged from dinosaurs during this era. The next period was the Cretaceous, which ranged from 145 to 66 Ma. The Cretaceous represents the longest period of the Phanerozoic, and a number of developments occurred during this period. For instance, phylogenetic and fossil studies indicate, albeit not without controversy, that the common ancestor of birds lived in the late Cretaceous era. Angiosperms (flowering plants) also underwent diversification in the Cretaceous, although their ancestors may have diverged from other seed-producing plants in the Triassic.

The Cretaceous period came to a close with the famous Cretaceous-Paleogene (K-Pg) extinction event. Of the Big Five mass extinctions, the cause behind the K-Pg event is arguably the most unambiguous—it was probably driven by the combined action of a large asteroid (or comet) impact in Mexico and intense volcanism in the Deccan Plateau of India (Burgess 2019), eventually resulting in severe climate disruptions that may have extinguished ~ 75 percent of all species. In the aftermath of the extinction, both birds and mammals underwent a period of rapid speciation and occupied ecological niches that had hitherto been mostly inaccessible to them (Meredith et al. 2011). It is possible that some of the extant placental mammals originated shortly after the K-Pg extinction and diversified soon thereafter. The most recent common ancestor of modern primates, as per some analyses, is traceable to shortly after the K-Pg extinction, but this remains a highly contentious issue. Some studies favor an origination time in the late Cretaceous (~ 85 Ma), whereas the direct evidence from fossils stretches as far back as ~ 57 Ma, if one accepts the premise that the ambiguous *Altiatlasius* was a primate (Tuttle 2014).

Following the Cretaceous, we enter the Paleogene period (66–23 Ma). As noted above, the radiation of mammals and birds occurred in this period.

---

15. *Placentals* refers to mammals that have placentae and give birth to live young that are relatively advanced in terms of development. The placenta is an organ that provides nutrients to the growing fetus.

The adaptive radiation of primates also unfolded during this era. The next epoch was the Neogene, which spanned 23–2.6 Ma. Diverse groups of primates, including the ancestors of modern-day great apes, were living in this period. As per the latest evidence, the last common ancestor of chimpanzees and humans is believed to have existed in the Neogene, although the exact timing remains uncertain; it was, however, probably alive earlier than ~ 4 Ma. The final epoch in this rich and complex tapestry is the Quaternary period (2.6 Ma to present). This era witnessed a series of glaciation events, and the advent of genus *Homo* around 2–3 Ma, although this estimate is subject to variability. The most recent major evolutionary breakthrough, the emergence of anatomically modern humans, may be dated to ~ 300 kyr ago.

In summary, we have seen that the history of life during the Phanerozoic eon witnessed many twists and turns. Despite this inherent complexity, a few general tendencies have been identified by using the fossil record. As a comprehensive treatment of such trends lies well beyond the domain of this book, we shall not cover such aspects in detail but will instead round off our analysis with an illustrative example. To wit, a general tendency has been observed across various lineages—namely, they evolve toward increasingly larger sizes over time; this trend is known as Cope's rule. Let us ask ourselves the question of how the maximum size of terrestrial mammals (one specific animal group) has morphed over time by deploying a mechanistic model.

We denote the maximum mass by $M_{\max}$. The most obvious model that springs to mind is $dM_{\max}/dt \propto M_{\max}$. This does not capture the full picture, however, since an increase in the organism's mass decreases the evolutionary possibilities within the ecological niche that it inhabits. In other words, there exists a theoretical maximum ($K_{\max}$) that cannot be surpassed by terrestrial mammals in a given niche space. Therefore, $dM_{\max}/dt$ has also been predicted to have proportional dependence on the factor $\log(K_{\max}/M_{\max})$, which reflects the difference in logarithmic body mass (F. A. Smith et al. 2010). By combining the results together, we end up with

$$\frac{dM_{\max}}{dt} = \lambda_m M_{\max} \log\left(\frac{K_{\max}}{M_{\max}}\right), \tag{3.38}$$

where λ_m signifies the rate at which the expansion in size occurs. The solution of this differential equation is given by

$$\log\left(\frac{M_{\max}}{K_{\max}}\right) = \log\left(\frac{M_0}{K_{\max}}\right) \exp\left(-\lambda_m t\right), \tag{3.39}$$

with M_0 denoting the maximum mass at $t = 0$. For terrestrial mammals on Earth, $M_0 \approx 7$ kg, $K_{\max} \approx 1.3 \times 10^4$ kg, and $\lambda_m \approx 8 \times 10^{-8}$ yr^{-1}. It is, however, important to appreciate that these values are sensitive to the temperature and the area of the terrestrial niche under consideration. This calculation serves to underscore the fact that the evolution of relatively large intelligent organisms will necessitate a certain amount of time. In other words, it is conceivable that worlds with comparatively short-lived intervals of habitability (for whichever reasons) would not possess sufficient time to permit the evolution of intelligent organisms.

3.6.6 The likelihood of extraterrestrial complex multicellularity

In dealing with the question of whether complex multicellularity is conceivable on other worlds, we are confronted with the same issue as always: the availability of a single data point. With the proviso acknowledgment that life elsewhere need not follow the same paths as on Earth, it would initially seem as though the evolution of complex multicellularity is not very difficult. This is because of a number of reasons. For starters, the routes to simple multicellularity are many, and the timescales over which this phenomenon can arise are apparently quite short. Second, the genetic tool kit that facilitated the emergence of complex multicellularity does not appear to have been wholly novel—that is, many components probably existed in its unicellular ancestors—although there were certainly aspects that were new. Finally, complex multicellularity arose in plants, animals, and two groups each of algae and fungi—a total of six ostensibly independent origins.[16] Even if this number does not come across as being large, one may argue that complex multicellularity has arisen with relative ease on Earth compared to most of the other MEEs encountered earlier.

Thus, all's well that ends well, à la Shakespeare, and we are now free to tackle the next major evolutionary innovation. Or can we? Our prior discussion obscured a rather subtle point—namely, the arguably fundamental difference between animals and the other groups with complex multicellularity. Among all eukaryotes that evolved this characteristic, animals are

16. To this list, one might also add certain prokaryotes such as heterocyst-forming filamentous cyanobacteria because they are characterized by cellular differentiation and functional cooperation (Herrero et al. 2016), although this classification is subject to some subtle caveats.

the only ones to have retained phagocytosis. A number of developmental, structural, and genetic constraints had to be overcome before unicellular heterotrophic organisms endowed with phagotrophy and lacking in cell walls could accomplish the transition to obligate multicellularity (Cavalier-Smith 2017; D. B. Mills et al. 2018). Apart from these developmental constraints on the evolution of animals, their emergence also appears to have been correlated with the rise in atmospheric and oceanic oxygen levels, thus implying that the latter might have been necessary for the former to flourish, and perhaps even originate.

In view of these issues, it is more accurate to differentiate between animal and complex multicellularity. Even if the probability of the latter is not very small, it does not automatically follow that the same holds true for the former. On the basis of the importance of animals as ecosystem and evolutionary engineers in the Phanerozoic eon, which stems from their macroscopic and motile nature, we propose that animal multicellularity should be regarded as a MEE in its own right, distinct from that of complex multicellularity in general. If we adopt this stance, it is plausible that this MEE may have a low likelihood of occurrence on other worlds.

3.7 INTELLIGENCE IN ANIMALS

The notion of *intelligence* encompasses many different concepts, some of which are tangential or antithetical to one another, making this term hard to define and quantify. It is probably safe to say that intelligence in animals shares close connections with cognitive capacity and complexity and is reflected in the myriad avenues whereby organisms store, process, and analyze information as well as in their ability to adapt their actions accordingly. One point worth appreciating here is that, *perhaps* barring humans to some degree, one finds a fairly continuous spectrum of intelligence from unicellular organisms to complex multicellularity on Earth.

For instance, there has been growing awareness and appreciation of microbial intelligence (Baluška & Levin 2016). The formation of complex networks by microbes has been shown to yield emergent characteristics reflective of intelligence such as memory, adaptation, and anticipation. The reader can consult Westerhoff et al. (2014) for a review of this fascinating subject. The slime mold *Physarum polycephalum,* despite being a "simple" protist, has demonstrated a fascinating range of complex behaviors

(Vallverdú et al. 2018). It is particularly adept at finding the shortest path between two (or more) sources of food in a maze. Furthermore, it also exhibits some variant of memory as it anticipated unfavorable conditions that were cyclic and adjusted its locomotion accordingly.

Looking further ahead, the first nervous systems in animals probably appeared before the Cambrian period at some point in the Ediacaran, perhaps around 600 Ma. Nerve cells (neurons) have been hypothesized to be an example of convergent evolution, having possibly evolved in at least two different lineages independently—the *Ctenophora* on the one hand and the rest of the animals with nervous systems on the other. Many of the genetic mechanisms underlying the formation of nervous systems are therefore remarkably conserved across species, especially within mammals, and the same is valid for specialized signaling molecules known as neurotransmitters. The reader may naturally inquire at this stage: What of plants? Although this topic lies beyond the scope of this book, plants are believed to feature sophisticated internal communication and decision-making mechanisms despite the absence of conventional nervous systems, converse with each other for various purposes (e.g., predator warnings), and distinguish between strangers and kin (Trewavas 2014, 2017; Calvo et al. 2020). Notwithstanding promising empirical results in this field, claims pertaining to plant intelligence and consciousness have met with skepticism (Taiz et al. 2019).

A number of misleadingly simple animals, most notably insects, evince impressive cognitive abilities. Insect colonies (e.g., ants and bees) display collective memory when transitioning from one mode of communal behavior to another. Individual insects have also demonstrated the capacity for complex cognition in the realms of feeding, navigation, and social behavior. Moreover, certain insect species seemingly possess cognitive maps that enable them to memorize flight routes and landscapes. Honeybees offer some of the most striking examples of insect cognition: they can count from one to four, add and subtract elements, distinguish between same and different stimuli, and comprehend top-down spatial relations (Chittka et al. 2019). Bees and flies are potentially capable of learning via observation and from other members of the same species (Perry et al. 2017). A number of invertebrates, especially octopodes from the class *Cephalopoda* (which make their appearance in Section 7.7), exhibit a variety of advanced cognitive functions, such as their ability to learn from others (i.e., social learning).

When it comes to vertebrates, the range of high-level cognitive abilities is expanded considerably, as we shall see shortly. At the same time,

we caution that the above lines of evidence should not be construed as grounds for assuming perfect continuity between human and nonhuman intelligence. The debate about which aspects of intelligence are uniquely human has raged on for decades, with no signs of abating soon. We will repeatedly encounter four traits in our subsequent discussion as well as in Sections 3.8 and 7.7.

- **Self-awareness** refers to an individual's ability to perceive oneself as a distinct entity; at the risk of oversimplification, it may be envisioned as possessing an *autobiographical sense*.
- **Tool making and tool use** correspond to the ability of organisms to construct and use tools (namely, devices for implementing certain functions) for simplifying their labor.
- **Culture** is imbued with a wide range of connotations, owing to which a precise definition remains rather elusive. However, for the purposes of our discussion, it embodies a set of distinctive behavioral traits that originate in local populations and are vertically transmitted across generations via social learning. From this standpoint, culture acts as a medium for facilitating the flow of information in a manner that resembles, but is not identical to, genetic information, consequently enabling further innovation and evolution.
- **Language** also admits a very broad spectrum of definitions. Some are based on its structural components (e.g., lexicon, syntax, phonology), whereas others interpret language as a symbolic communication system that enables the sharing of information, thoughts, and experiences. We will implicitly adopt the latter approach henceforth.

Many other characteristics, often perceived as being uniquely human, will be encountered in Section 3.8. However, we shall not tackle protean concepts such as consciousness and morality, for the reason that they are arguably harder to quantify.[17] This is not to say that animals other than humans are

17. In this context, S. Ginsburg and Jablonka (2019) have suggested that the transition to consciousness was principally fueled by learning, which is also one of the chief ingredients of cultural transmission. Other researchers have singled out the propensity to integrate

altogether lacking insofar as these vital features are concerned. Insightful discussions of the evolution of consciousness in animals (humans or otherwise) can be found in D. R. Griffin (2001), Bekoff and Pierce (2009), Damasio (2010), Dennett (2017), and Feinberg and Mallatt (2018), but the path toward achieving an exhaustive understanding of these topics is a long and winding one.

3.7.1 Measures of intelligence

If the definition of *intelligence* is hard to pin down, it is logical to anticipate that any quantitative measures of intelligence could also succumb to the same limitation(s). Indeed, this surmise has turned out to be precisely the case. Nonetheless, it is worth briefly delving into a select few of the commonly used metrics of cognitive ability currently being employed by scientists.

The most simple metric that one can think of is the total mass of the brain. However, the usefulness of this metric is limited because larger animals will typically have larger brains, but it does not inform us a great deal about their relative intelligence. Nonetheless, overall brain size has been argued to constitute a more accurate metric of cognitive ability in nonhuman primates. The next measure that one may consider is the ratio of the brain mass (M_b) to that of the body mass (M_o). However, this factor encounters issues of its own, since it would predict that tree shrews attain a value higher than humans by a factor of a few. Hence, the brain-to-body-mass ratio is not always an accurate indicator of intelligence. On the other hand, it has been shown that this metric is a fairly accurate proxy for the problem-solving abilities of carnivorous mammals (Benson-Amram et al. 2016).

Among the most commonly employed metrics is the encephalization quotient (EQ). One of the chief motivating factors behind EQ is that the brain size does not scale linearly with the body size. When it comes to larger organisms, a disproportionately large brain size could be necessary for ensuring basic survival activities such as thermoregulation and motor skills. The encephalization quotient is calculated by applying the following formula:

---

information as comprising the bedrock of consciousness, and contended that this trait is ubiquitous in the animal kingdom (Tononi and Koch 2015).

$$\mathrm{EQ} = \frac{25}{3}\left(\frac{M_b}{10^{-3}\,\mathrm{kg}}\right)\left(\frac{M_o}{10^{-3}\,\mathrm{kg}}\right)^{-2/3} \tag{3.40}$$

If we assume that the organism area is proportional to $M_o^{2/3}$, we see that the equation can be interpreted as proportional to the ratio of the brain mass to the body area. The numerical constants in (3.40) are empirically determined and do not have a standard theoretical derivation. Let us evaluate the EQ for a typical human. We choose $M_b \sim 1.4$ kg and $M_o \sim 62$ kg, which leads us to $\mathrm{EQ} \approx 7.45$. The northern right whale dolphin (*Lissodelphis borealis*) has one of the highest EQs among nonhuman animals, reaching values as high as 5.55. Chimpanzees have an EQ of approximately 2.5–2.6, whereas elephants have EQs of ~ 1.5–2.

However, even the EQ is far removed from a perfect metric since the values of blue whales (~ 0.4–0.5) and octopodes (~ 0.2) are lower than average despite their high intelligence.[18] In the case of blue whales, the EQ is suppressed due to their high percentage of body fat, which is necessary for thermoregulation, buoyancy, and food storage. When it comes to octopodes, about two-thirds of its neurons are found in its arms. Hence, the delocalization of the neuronal distribution is chiefly responsible for suppressing the EQ of octopodes. Finally, we observe that certain birds manifest modest EQ values in contrast to their proven high intelligence. The former result is attributable to the fact that the neuron-packing density (i.e., the number of neurons per unit of brain mass) of birds like parrots and songbirds is about twice that of nonhuman primates with the same brain mass (Olkowicz et al. 2016).

The preceding discussion reveals that no single quantity to date accurately encapsulates the true intelligence of animals. This limitation also applies to the general intelligence factor (g-factor), a measure of the positive correlations among diverse cognitive abilities; to put it differently, the g-factor reflects the observed tendency that individuals who score highly on a given cognitive task will typically perform well on other tasks (Burkart et al. 2017). Looking beyond the EQ paradigm, Herzing (2014) proposed

18. A rigorous definition of *high intelligence* is not easy. If a species exhibits cognitive abilities that enable it to solve comparatively difficult problems or demonstrates any of the four pivotal traits listed earlier, we interpret it as having high intelligence.

an alternative semi-quantitative framework in which EQ constituted one of five dimensions; the other variables were complexity of communicative signals, individual complexity, social complexity, and interspecies interactions. Efforts along such lines might be fruitful for categorizing either diverse intelligences on our planet or extraterrestrial intelligences, if the latter do exist.

At the minimum, a more nuanced picture of neuronal factors is necessary in order to unravel why certain species are more intelligent than others (Herculano-Houzel 2016). The cerebral cortex is a thin layer that forms the outer surface of the cerebrum, the latter of which is the largest part of the brain. The cerebral cortex is particularly important as it serves as the seat of higher cognitive abilities relating to intelligence. Hence, a more useful diagnostic is the number of neurons (N_c) in the cerebral cortex, expressible as

$$N_c \approx \rho_c A_c L_c, \tag{3.41}$$

where ρ_c is the neuron-packing density encountered previously, while A_c and L_c denote the cortical area and thickness, respectively. Thus, organisms with simultaneously high values of ρ_c, A_c, and L_c will have the highest number of neurons. In the case of humans, we adopt the representative values: $\rho_c \sim 3 \times 10^{13}$ neurons/m^3, $A_c \sim 0.24$ m^2, and $L_c \sim 3 \times 10^{-3}$ m (Dicke & Roth 2016). By plugging these values into (3.41), we find that humans have $\sim 2.2 \times 10^{10}$ cortical neurons. This number is slightly higher than that of elephants and cetaceans (whales, porpoises, and dolphins), despite their much greater values of A_c.

Another factor of importance is the information processing capacity (IPC) of the brain, which depends on a number of factors, including the nerve conduction velocity, distance between neighboring neurons, and possibly the number of synpases (i.e., nerve junctions) per neuron (Dicke & Roth 2016). In the case of humans, it has been estimated that the IPC is about five times higher than cetaceans and elephants, mostly as a result of the higher conduction velocity. Hence, it is plausible that a combination of a high number of cortical neurons and IPC contributes to variations in the cognitive abilities of species, although this parameter is unlikely to constitute the sole differentiating factor. A combination of these two factors, especially the high IPC predicted by theoretical models, may explain why birds display high intelligence despite not scoring high on the aforementioned conventional metrics.

3.7.2 Species with high intelligence on Earth

We will not focus on cetaceans and octopodes at this stage, as they constitute the primary subject of our analysis in Section 7.7. It suffices to say, for the time being, that cetaceans have demonstrated self-awareness, simple tool usage, cultural transmission of information, and maybe even morality. Likewise, there is fairly convincing evidence that octopodes take part in social learning and that they also use simple tools. Here, we shall mostly pay attention to birds, nonhuman primates, and elephants, which have been thoroughly documented to possess high levels of intelligence and emotional sensitivity. We will therefore not tackle domesticated mammals such as pigs, horses, dogs, and cats, which are intelligent to varying degrees; for example, Marino and Colvin (2015) provide a thorough documentation of the impressive cognitive abilities of pigs. The reader is invited to peruse the monographs by Shettleworth (2010), Safina (2015), de Waal (2016, 2019), and Roth & Dicke (2019) for additional details.

3.7.2.1 Birds

Although nonhuman primates are the best-known examples of high intelligence, there is mounting evidence that certain avian taxa—most notably *Corvidae* (e.g., crows) and *Psittaciformes* (e.g., parrots)—are possessed of cognitive abilities that rival those of nonhuman primates (Emery & Clayton 2004; Clayton & Emery 2015; Güntürkün et al. 2017; Aplin 2019).

Members of the crow family (corvids) are sophisticated tool users that either utilize natural tools or modify them for their purposes. New Caledonian crows (*Corvus moneduloides*) are palpably adept at the manufacture, alteration, and deployment of tools using screw pine (genus *Pandanus*) leaves as the raw material; in fact, recent experiments suggest that these birds derive intrinsic enjoyment and motivation from tool use (McCoy et al. 2019). New Caledonian crows are also rare examples of animals that are capable of metatool usage, i.e., enlisting one tool to create / access another. In particular, they were able to harness a short stick to extract a longer stick that was finally employed to obtain food. There is some evidence for tool use in many other birds such as the Hawaiian crow (*Corvus hawaiiensis*) and Goffin cockatoos (*Cacatua goffini*). It has been proposed that the convergent utilization of foraging tools by avians is attributable to the fact that many of them inhabit islands, where there is less competition for embedded prey as well as reduced risk from predators.

An important point worth highlighting in this context is the mounting evidence for cultural transmission in birds. The most widespread and detailed evidence of social learning in birds pertains to migratory routes and song learning. To offer a different example, great tits (*Parus major*) bring back larger prey to their offspring than blue tits (*Cyanistes caeruleus*). It was therefore shown that blue tits raised by great tits tended to bring back larger prey to their offspring than the control group—namely, blue tits raised by their own kind. The opposite behavior was observed in the case of great tits along expected lines. Collectively, these experiments illustrate the existence of social learning, opening the possibility for cultural transmission. In another study, new foraging techniques were taught to only a couple of great tits per subpopulation. It was found that these techniques were rapidly spread to 75 percent of the population (about four hundred individuals) via social networks (Aplin et al. 2015). Eventually, this resulted in the establishment of local traditions that were transmitted across generations. Other experiments involving birds have demonstrated that cultural conformity, stresses, traditions, and adaptive strategies play an important role in their lives, just as they do in humans.

Another important point worth underscoring concerns mental time travel (MTT). It has been described as the mental capacity for reconstructing personal events that occurred in the past (episodic memory) on the one hand, and the formulation of possible future scenarios (foresight) on the other. There has been a long debate about whether MTT is uniquely human or not, with earlier studies favoring the former standpoint while subsequent analyses have been more divided (Corballis 2017). The best-known examples in connection with MTT are scrub jays (genus *Aphelocoma*), which constitute a particular sub-group of corvids. Data from experiments suggests that scrub jays, during the process of caching (storing) and retrieving their food, have knowledge about *what* (type of food stored), *where* (the location of storage), and *when* (times at which food was stored and recovered). Collectively, this has been interpreted as evidence of episodic memory. Similar behavior has been observed in other birds, most notably in the black-capped chickadee (*Poecile atricapillus*). Scrub jays also display signs of anticipating future needs and planning for them accordingly in observations of how they cache food for later usage. Similarly, as per the experiments conducted by Kabadayi and Osvath (2017), ravens (*Corvus corax*) plan for the future (up to seventeen hours in advance) when it comes to tool use and bartering, with their abilities being allegedly on par with nonhuman primates and young human children.

At least one species of birds—the common magpie (*Pica pica*) from the corvid family—has passed the mirror self-recognition test (Prior et al. 2008). This test is meant to assess the capacity for visual self-recognition by determining whether subjects investigate markings (placed on the body) visible only in the mirror and has therefore been regarded as a proxy for self-awareness.[19] Magpies display other interesting characteristics such as recognizing other individuals from the same species and identifying cached objects. A closely related concept is the theory of mind (ToM), which is conventionally defined as the capability to ascribe mental and emotional states to both oneself and other individuals, thereby recognizing that each individual has unique motives and perspectives. There is some tentative, albeit disputed, evidence for a ToM in ravens and scrub jays.

Ravens, in particular, seem to alter their strategies of pilfering caches depending on whether they are confronted with the original storer or not. Similarly, when an experimenter hid food, a chosen raven had the ability to predict the behavior of other bystander ravens by taking into account whether the view of these bystanders was blocked or not. Similarly, scrub jays also alter their caching strategies depending on whether the onlooker had observed them previously. More intriguingly, the tendency to rehide the food or take further protective actions when observed by potential pilferers was much more pronounced in experienced scrub jays that had been pilferers themselves. These findings were argued to imply that scrub jays demonstrate an awareness of the knowledge states of other individuals. Collectively, it has been argued by certain scientists that corvids may be endowed with precursors of the human ToM (Taylor 2014).

The gray parrot Alex was able to count objects up to six, identify ~ 50 different objects, and purportedly comprehended abstract concepts such as bigger and smaller as well as the notion of zero. Alex possessed a vocabulary of around one hundred words and appeared to be cognizant, at least in part, of what he was saying and might have had some understanding of syntax in language. However, most of these claims should be judged with due skepticism, as the possibility that Alex was responding to subconscious cues cannot be eliminated.

---

19. We caution, however, that the methodology and results associated with some of the mirror self-recognition tests have attracted criticism, and the efficacy of this test in terms of measuring self-awareness also remains the subject of debate (Adriaense et al. 2020).

3.7.2.2 Nonhuman primates

The mental feats and high intelligence of nonhuman primates are well appreciated even by laypersons; as a result, we will not delve into this subject in much detail. To begin with, we observe that nonhuman primates have large brains for their size compared to most other mammals. Numerous hypotheses have been advanced to explain this fundamental observation (Dunbar & Shultz 2007, 2017). One prominent class of models suggests that the necessity of foraging for specialized foods consumed by these animals is responsible for greater cognitive demands, thus promoting the expansion in cognitive abilities (Rosati 2017; González-Forero & Gardner 2018). Alternatively, the need for social cooperation in order to solve large-scale ecological and communal problems may have served as the major driving mechanism behind the increase in brain size.

It goes without saying that tool usage is fairly common in nonhuman primates, and not just in the apes (Van Lawick-Goodall 1971; Shumaker et al. 2011). Ring-tailed lemurs (*Lemur catta*) are capable of manipulating objects and have been shown to manipulate food-dispensing toys (puzzle feeders) in the wild. There is some evidence that this behavior is socially learned, and perhaps even culturally transmitted across generations. Capuchin monkeys (of the genus *Cebus*) have been documented to use tools in the wild for cracking open nuts. Chimpanzees (*Pan troglodytes*) are probably the most adept tool users apart from humans. They vary their use of tools depending on the task and are characterized by their diverse tool kits. Chimpanzees have evinced the ability to manufacture multiple tools from the same raw materials or put together different raw materials to construct tools. Gorillas (genus *Gorilla*) and orangutans (genus *Pongo*) have also been documented to use tools in the world, especially the latter.

Apes are particularly adept at quantity representation—namely, they can identify which of two given sets contains a greater number of items, even when this number was high and the items were packed closely together. Chimpanzees, and macaques (genus *Macaca*) to a lesser degree, also have impressive working memories. In other words, when presented with certain stimuli, they could memorize them and either repeat or select them at a later stage. In a striking experiment by Inoue and Matsuzawa (2007), a mixture of adult and young (mother-offspring) chimpanzees were shown the numerals 1 through 9 at different on-screen positions, and had to subsequently touch the correct squares in the same sequence. All chimpanzees

accomplished this task successfully, and the young chimpanzees outperformed human adults (cf. Cook & Wilson 2010). Chimpanzees also have a propensity for problem-solving, as seen from their capacity for finding the way out of mazes.

A number of meticulous studies have documented social learning and cultural transmission in apes as well as monkeys (Cheney & Seyfarth 2007; Van Schaik 2016; Whiten 2017). To begin with, it has been discovered that geographically distinct populations exhibit variations in tool use, foraging, and social behavior. The number of culturally distinct subpopulations ranges between $\gtrsim 20$ in gorillas and orangutans and $\gtrsim 40$ in chimpanzees. There are many documented instances of chimpanzees teaching juveniles, with some of the most widely studied examples coming from termite fishing using tools derived from stems. It has, for instance, been found that female offspring spend more time attending to the fishing practices of their mothers, thus picking up the skill much faster than male offspring. Furthermore, cultural transmission has been shown to occur horizontally (diffusion): the use of moss as a water-sponge spread from an alpha male to others by way of social affiliations. The total number of tools transmitted culturally are > 30 in total for chimpanzees, and about a factor of 2 lower for orangutans. There is some tentative evidence in favor of cumulative culture in chimpanzees (and other nonhuman primates), a trait that had otherwise been regarded as being unique to humans.

Chimpanzees have also been documented to possess the ability to plan for the future. When searching for ephemeral fruits (e.g., figs), chimpanzees demonstrated foresight in orienting their nests toward the food source the previous day, and leaving early in the morning in pursuit of these fruits. Sumatran orangutans (*Pongo abelii*) in the wild emit long calls that serve as signposts of their travel direction. Analyses of these calls indicate that the orangutans make travel plans in advance and utilize the calls for conveying their plans to members of the same species (i.e., conspecifics). Similar behavior indicative of future planning has been documented for chacma baboons (*Papio ursinus*). Laboratory experiments have shown that both orangutans and chimpanzees are willing to forgo immediate gains for the sake of securing a greater reward in the future, and they were able to plan for the future by selecting the appropriate tools. A captive male chimpanzee concealed projectiles at strategic locations, thus enabling him to launch them at visitors before they could back off (Osvath 2009). This set of actions demonstrated sophisticated future planning that took into account spatial

positioning as well as the ability to recognize the areas where the visitors would tend to cluster. Other experiments have demonstrated that chimpanzees and orangutans display signs of autobiographical memory and the ability to recall events on the basis of specific cues, both of which mirror traits that exist in humans. The aforementioned instances have been interpreted by some scientists as evidence of mental time travel, but we caution that this issue has not been decisively settled yet.

Monkeys and apes are also capable of imitation, which is considered a higher cognitive ability in certain quarters. Imitative behavior, especially emulation, facilitates the rapid (and intergenerational) acquisition of knowledge by an individual watching the actions of another conspecific. Imitation has been documented in many monkeys, orangutans, and especially chimpanzees. In the Furuvik Zoo of Sweden, three out of five chimpanzees were able to imitate humans at rates almost comparable to those of humans imitating chimpanzees—to wit, the converse scenario (Persson et al. 2018). The chimpanzees seemed to recognize the signs of being imitated by humans (i.e., imitation recognition) in 36 percent of the cases to boot. The rationale behind such cross-imitation games was argued to spring from social and communicative motives. Although humans do exhibit ostensibly more sophisticated imitative abilities than the great apes, it may very well be that the gap between the two is not as profound as originally thought.

Both chimpanzees and orangutans have passed the mirror self-recognition test, although the fraction of individuals that pass the test is typically < 50 percent. The issue of whether great apes possess a ToM is a classic question that has attracted polemics from opposing camps over decades. However, the tide appears to be swaying, if not changing, with respect to this long-standing debate. An innovative series of experiments conducted by Krupenye et al. (2016) may have established that three species of apes—chimpanzees, bonobos (*Pan paniscus*), and orangutans—comprehend false beliefs—namely, they recognize that the actions of other individuals are dictated by beliefs about reality, even when said beliefs may be false. In the experiments, a human agent was led to believe that the desired object was present in one of two possible locations, although it was situated in neither of them (as known to the apes via video). When the agent returned to retrieve the object, in most of the cases, the apes correctly anticipated the location that the human would visit, even though they already knew that the object did not exist there anymore.

Subsequent experiments have lent further credibility to the notion that apes can distinguish between the mental states (false versus true beliefs) of conspecifics and act accordingly. In addition, chimpanzees display the capacity to warn other chimpanzees of incoming danger by monitoring the information that is accessible by the rest of the group. Moreover, orangutans vary their gestures depending on the extent to which these signals are understood and grasped by their interlocutors. The fundamental question of whether great apes are endowed with a ToM is by no means conclusively settled by the preceding results, but they offer arguably compelling grounds in favor of this perspective. Evidently, further research along similar lines is imperative to make substantive headway in unraveling this knotty issue.

Finally, we note that several apes have been taught sign languages to varying degrees of success, with notable examples being Koko, Nim Chimpsky, and Kanzi. Although some of them were allegedly capable of comprehending over 1000 signs and complex questions, the tests were not conducted in rigorous environments. Furthermore, most of these apes did not evince an understanding of grammar, nor did they ask questions themselves for the most part. Setting aside the thorny issue of whether apes can learn human languages, it is worth noting that many of them use complex nonverbal gestures, of which some appear to be similar across species. Ongoing research indicates that (1) these gestures are acquired through social interactions, (2) they are intentionally directed toward specific individuals with the expectation of eliciting certain behavioral responses, and (3) they vary depending on the recipient and context (i.e., are multimodal in nature) as well as the gender of the signaler (Byrne et al. 2017; Fröhlich et al. 2019). To summarize, great ape gestures are drawn from a wide repertoire with the purpose of intentional communication. Hence, while they are very distinct from human language, perhaps they might provide clues about how human languages originated.

3.7.2.3 Elephants

Elephants do not perform as well as expected when it comes to aspects like learning via visual discrimination. Instead, they excel in tasks that involve the use of long-term memory as well as their excellent spatiotemporal orientation. In addition, they have been associated with traits such as altruism and close-knit societies, which are less commonly linked with other animals.

The reasons behind these differences are believed to partly stem from the lower neuron density in the cortex as well as the slower speed of information processing relative to primates.

To begin with many instances of tool use in elephants have been documented. For instance, elephants use and modify branches in several ways for the purpose of repelling flies. Asian elephants (*Elephas maximus*) have been observed to move plastic objects to a particular location and stand on them to acquire food that would otherwise be beyond their reach. This case illustrates not only tool use but spontaneous problem-solving ability as well. Interestingly, this solution was not obtained through trial and error, implying that it may constitute an example of insight-based problem-solving. In other experiments, they have pulled trays for accessing food, thus evincing goal-directed behavior. Elephants also display evidence of quantitative judgment: they were able to distinguish between two sets of food and identify the larger portion.

Elephants, as remarked earlier, are well-known for their remarkable long-term memories, which enable them to recognize individuals from long ago, recall the best possible routes and locations when searching for food and water, and undertake tasks with high accuracy that they had been taught several years in the past. As a result, one could argue that elephants have episodic memory in that they can recall what-when-where, although more rigorous tests in controlled environments are necessary. In spite of their highly cooperative societies, documented cases of social learning are somewhat rare. However, in captive African bush elephants (*Loxodonta africana*), indirect evidence for social learning appears to have been gleaned because conspecifics actively watching a selected individual interact with the feeding apparatus were observed to spend more time with the latter afterward.

Elephants are fairly capable of imitative behavior. An Asian elephant named Koshik was able to reproduce five Korean words, which were understood by $\sim$ 15 to 56 percent of Korean native speakers (Stoeger et al. 2012). Koshik's ability to imperfectly mimic human speech originated from placing his trunk inside the mouth and modulating vocal production. Elephants have also manifested signatures of vocal learning in other milieu—an African elephant was able to imitate the sounds made by trucks. Asian elephants are among the few animals that have passed the mirror self-recognition test (Plotnik et al. 2006), indicating that they are conceivably self-aware. Inasmuch as the premise that elephants possess a ToM is concerned, the evidence in its favor is indirect. Elephants are known to reassure conspecifics

in distress, thereby potentially indicating they are aware of, and pay attention to, the mental states of others.

We will round off our discussion by highlighting other select examples of elephant cognition. To begin with, elephants display the capacity for sophisticated communication. For instance, they emit different alarm calls in response to threats by humans and bees and adapt their behavior accordingly. African elephants comprehend human gestures and can interpret them to locate food. They are adept at cooperative tasks and function at a level comparable to chimpanzees. This was demonstrated in an example where two elephants were able to simultaneously pull on two ends of a rope to access a reward and understood that the positive outcome was attainable only via joint action. Elephants have the ability to distinguish between different humans through an array of cues ranging from odor to color of body garments. They are also capable of distinguishing the gender, age, and ethnicity of humans on the basis of acoustic signals alone. Last, but certainly not least, elephants are one of the prime examples for the actuality of animal empathy (de Waal 2008), as there are good reasons for believing that they apprehend the physical abilities, emotional states, and goals of conspecifics, even when they differ from their own.

3.7.2.4 Fish, amphibians, and reptiles

Cichlids, a group of fish, are regarded by some scientists as evincing considerable cognitive abilities. They are capable of recognizing other individuals on the basis of both visual and auditory cues. Cichlids have also demonstrated the capacity for remembering the past behavior of other individuals and responding accordingly—for example, by excluding cheaters. There are reported examples of fish (e.g., guppies) taking part in social learning through observing conspecifics. Social learning in fishes encompasses antipredator strategies, food sources, mate choices, and cooperative hunting, to name a few. As fish are quite removed from mammals and birds in terms of their habitat and physiology, this has engendered an unfortunately widespread tendency toward underestimating their cognitive abilities (C. Brown 2015). In 2019, researchers presented experiments claiming that bluestreak cleaner wrasse (*Labroides dimidiatus*) passed the mirror self-recognition test (Kohda et al. 2019), but the results remain strongly disputed.

The evidence for high intelligence in both amphibians and reptiles is comparatively limited in terms of reliable empirical studies. It has, however, been argued that the lizard *Anolis evermanni* is capable of solving complex tasks that require high cognitive abilities and the utilization of multiple strategies. Similarly, some lizards such as *Eulamprus quoyii* and *Uta stansburiana* display tentative evidence of spatial memory. Finally, younger *Eulamprus quoyii* individuals appear to have the potential to partake in social learning, while older individuals do not.

3.7.3 The likelihood of high extraterrestrial intelligence

What are the prospects for high, but nontechnological, extraterrestrial intelligence? Before answering this question, a recap of some pertinent points is necessary. First, we have seen that a number of routes are accessible for attaining high intelligence on Earth. Octopodes have distributed nervous systems with relatively small brains, birds also have small brains but are characterized by high neuron-packing densities in the cortex, whereas mammals (especially nonhuman primates) have large brains. Second, we find high intelligence in a number of groups: birds (especially corvids and parrots), nonhuman primates (particularly the great apes), cetaceans (whales and dolphins), cephalopods (octopodes to be more specific), and elephants and other mammals (especially pigs). Of these groups, the first octopodes probably arose several 100 Myr ago in the ocean, the rapid diversification of extant cetaceans took place approximately 20–30 Ma, whereas the ancestors of present-day great apes began diverging around 10 Ma. Thus, we see that high intelligence has arisen repeatedly in very different epochs, habitats, and animal groups. A similar conclusion may also apply to neurons and other components of neuronal systems, which are presumed to have evolved independently more than once in animals.

The final question that we must ask is, What are the possible conditions under which high intelligence is beneficial and therefore likely to confer a selective advantage? Naturally, the answers to this question depend on the choice of hypothesis to explain the genesis of intelligence; synopses of the potential pathways to intelligence have been set forth in Bogonovich (2011) and Rospars (2013). Social behavior can promote intelligence because the former requires characteristics such as sophisticated communication, cooperation, and memory. Alternatively, organisms with predatory lifestyles or

specialized foraging patterns also require many of the same traits; examples include communication among wolves and memorization of water holes by elephants. From a physiological standpoint, organisms that manifest dexterous motor control (e.g., limb movement in octopodes and humans), complex sensory perception (e.g., echolocation in cetaceans), and combinations thereof would probably correlate with high intelligence (or the cruder rubric of large brains) for a multitude of reasons primarily attributable to expeditious and structured information processing and transference (Chittka & Niven 2009; Stout & Hecht 2017).

Hence, on the basis of these arguments, extraterrestrial biospheres with diverse and intricate ranges of habitats, organisms, and food webs may have a reasonable likelihood of giving rise to highly intelligent (but nontechnological) species.

3.8 TECHNOLOGICAL INTELLIGENCE

> Nevertheless the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind.
>
> —Charles Darwin, *The Descent of Man, and Selection in Relation to Sex*

Was Darwin correct as regards the above quotation? If not, what are the key discontinuities that demarcate humans from nonhumans? These are fundamental questions that have been debated in evolutionary biology since the nineteenth century, and we shall refrain from wading too deeply into this particular morass. One must, however, take care to avoid false dichotomies. To offer an inexact analogy, one regularly encounters phase transitions in statistical physics wherein the first derivative of the thermodynamic free energy is continuous and the second derivative is discontinuous (e.g., the transition from paramagnetism to ferromagnetism). Hence, restricting oneself to scrutinizing only the first derivative would yield incomplete information about the underlying phase transition behavior. In the same spirit, it appears appropriate to dispense with diametrically opposite concepts as they are ultimately limiting in their scope.

Instead, it is our objective herein to delineate a few of the most prominent characteristics that are often perceived as being distinctly "human" as well as to sketch their emergence from nonhuman antecedents. We will focus on (1) language, (2) cumulative cultural evolution, and (3) sophisticated tool construction and deployment; it must, however, be realized that

the trio are not wholly independent of each other. The three examples that we highlight are both subjective and limited to a considerable degree, partly because a more comprehensive analysis in this vein lies far beyond the scope of this book; we defer to Suddendorf (2013), Roth and Dicke (2019), and K. Andrews (2020) for overviews of this subject. Before delving into the aforementioned trio, a brief introduction to miscellaneous attributes that are regarded as possibly unique to humans is worthwhile.

3.8.1 Potentially unique human traits and their significance

One of the classic long-standing debates has revolved around the question of whether animals have a theory of mind (ToM). The importance of ToM is partly ascribable to its close connections with language and culture. Species with a ToM possess, in principle, the capacity for inferring the inner thoughts and feelings of others, which is undoubtedly an essential factor in the initiation and maintenance of social relations within large groups. One of the best-known candidates for explaining the evolution of human brains—the social brain hypothesis—suggests that the cognitive demands that arose from living in large and complex societies were responsible for playing a key role in hominin evolution.[20] Therefore, in this scenario, the significance of having a full-fledged ToM, a bona fide prerequisite for social cognition, is patently evident (Dunbar 2003).

On the basis of our preceding discussion, there are compelling grounds for supposing that, at the very least, the precursors of ToM are found in nonhuman animals, even if not a full-fledged ToM (Krupenye & Call 2019). The most compelling evidence for a ToM in animals comes from nonhuman primates, especially chimpanzees. Experiments involving pigs, parrots, and corvids also indicate that these species are endowed with something akin to a ToM. A detailed exposition of the viewpoint that animals do not possess a ToM can be found in D. C. Penn et al. (2008). However, taking into account recent developments, it appears more plausible that nonhuman primates have a ToM of sorts, albeit one that is not commensurate with adult humans vis-à-vis representational abilities; the same could apply to certain avian species (e.g., ravens). Instead, ToM in animals may be on par with that

20. Hominins comprise modern humans along with extinct ancestral species that belong exclusively to the human lineages.

of human infants—who appear to manifest implicit mind-reading mechanisms and an understanding of false beliefs (Heyes & Frith 2014)—in some respects.

The next point concerns mental time travel (MTT). A number of advantages are associated with the capacity for MTT. First, the ability to simulate events that can occur in the future enables a species to take action and prevent or mitigate negative consequences and increase the likelihood of positive outcomes. In short, it may confer an evolutionary advantage to such species (e.g., by increasing reproductive success). Second, it has been argued that language shares close links with MTT. Recounting MTT requires sophisticated communication methods (language), owing to which the faculty of language might have evolved in conjunction with increasingly complex MTT abilities. Moreover, it was postulated that the formal structures underpinning MTT and language are similar insofar as both attributes are characterized by the generation of longer entities from strings of shorter entities (Corballis 2017). The natural question, then, is whether MTT is unique to humans.

This postulate was, indeed, the crux of the classic hypothesis espoused by Thomas Suddendorf and Michael Corballis, *inter alia,* in the 1990s (Suddendorf & Corballis 2007). However, this notion has been increasingly challenged during the past couple of decades by two different lines of inquiry. The first is behavioral: we reviewed several studies in Section 3.7.2 that imply the existence of MTT in nonhuman animals. The purported examples in the literature range from chimpanzees and other nonhuman primates to rats and birds. One could, however, argue that these publications were affected by implicit biases or that alternative interpretations of the same evidence exist. The second avenue is therefore more promising, originating in neuroscience. The hippocampus, a crucial component of the brain, is known to serve as the seat of episodic memory and mental time travel. The hippocampal system is not unique to humans; it has been identified in other animals. Most notably, the hippocampus of rats enables them to construct mental spatial maps that account for location and orientation. Experiments centered on measurements of rats' hippocampi have demonstrated that they are capable of envisioning alternative routes or events and replaying previous ones in their minds. Hence, it is conceivable that the rudiments of MTT can be traced to nonhuman animals.

Another aspect that merits recognition is the propensity for humans to form large social groups characterized by a high degree of cooperation

(Bowles & Gintis 2011; Tomasello 2014). We will return to this theme, but a particular variant of social behavior is worth highlighting at this stage. The insect order *Hymenoptera,* comprising ants and bees, among others, is famously *eusocial*. Traditionally, eusocial species are distinguishable by the collective rearing of progeny, presence of multiple adult generations in a colony, and reproductive division of labor (i.e., some individuals reproduce while others do not). Eusociality has been documented not only in insects but also in at least two species of mammals—for instance, the naked mole rat (*Heterocephalus glaber*)—and three species of crustaceans from the genus *Synalpheus*.

Let us now turn our attention to *Homo sapiens*. First off, it is self-evident that multiple human generations coexist in the same social group. Second, one may propose that humans or their ancestors took part in the collective rearing of young (Hrdy 2009) and that a subgroup of humans who care for offspring in most societies are typically infertile (grandmothers). In this respect, humans would appear to fulfill the conventional criteria for eusociality. Needless to say, this thorny issue has been the subject of intense (even vitriolic) debate, with many, perhaps the majority, favoring the viewpoint that humans ought not be classified as a eusocial species sensu stricto; for a contrasting position, the reader should refer to Wilson (2012). Irrespective of whether humans can be regarded as eusocial or not, there is little doubt that (1) human societies are far more complex and widespread in comparison to virtually all other mammals and (2) the capacity for prosocial behavior had a major impact on human language and culture (Hare 2017; Wrangham 2019), as we shall discuss shortly.

In closing, we observe that our list of hypothesized traits unique to humans is not exhaustive. This field amounts to a veritable ocean of models, thereby leaving us with the unenviable task of sampling only a few "species" to deduce the immense biodiversity inherent in this figurative "ocean." One conjecture worth mentioning before moving on posits that the difference between humans and nonhumans directly stems from the former's ability to track and reason about higher-order structural relations by starting from first-order perceptual relations (D. C. Penn et al. 2008).

3.8.2 Language

The origin and evolution of language is replete with multitudinous hypotheses, due to which the ensuing discussion will be not only synoptic in nature but also highly selective. The reader is invited to peruse Pinker

(2007), Tomasello (2003, 2008), Fitch (2010), Arbib (2012), Scott-Phillips (2014), V. Evans (2014), and Berwick and Chomsky (2016) as well as the classic and eminently readable analyses by Hockett (1960) and Lenneberg (1967) for reviews of this Brobdingnagian subject.

To begin with, we can ask, In what respects does the faculty of language differ between humans and nonhumans? In a seminal review, Noam Chomsky and his collaborators asserted that the unique component of human language is attributable to recursion—namely, the ability to generate an essentially infinite number of expressions by starting from a finite set of building blocks (Hauser et al. 2002); even at the time this hypotheses was propounded, it elicited highly polarized responses and thus remains the subject of intense debate to this day. A substantial number of elements underlying the faculty of language are shared across species and include both physiological and cognitive characteristics (Jarvis 2019). Basic components such as speech perception and auditory and visual sensing in humans are similar to those in other animals. As for cognitive capacities, we have already encountered the likes of MTT and ToM and pointed out that precursors of both these traits exist in nonhuman animals.

With regard to the divergences, complex vocal control (and perhaps anatomy) distinguishes human speech from the communication systems of other primates, but it is important to recognize that speech is not a prerequisite for language, as seen from the existence of sign languages. A second aspect of language that is extensively regarded as being uniquely human concerns our capability to combine elements (e.g., words) in concordance with certain rules to generate more complex structures (e.g., sentences) that convey complex semantic messages; this alleged facet has historically attracted much fame as well as opprobrium. Finally, there appear to be elusive, but potentially distinctive, differences in the capacities of humans and nonhumans to comprehend and interpret the underlying meanings of signals depending on context.

A number of models have been proposed to explain the mechanisms by which human language originated. However, Számadó and Szathmáry (2006) have argued that most of them are unable to account for all of the following criteria: (1) shared interests among communicating parties, (2) underlying concepts being grounded in reality, (3) the capacity to convey generalizations, and (4) why only humans evolved the faculty of language. Some of the central hypotheses for the origin of language are delineated below:

- **Grooming:** As the size of hominin groups increased, the ability to virtually "groom" more than one individual simultaneously became a substitute for physical grooming.
- **Hunting:** The necessity of taking part in group hunting spurred the development of language as a means of coordinating hunting efforts and communicating information about prey and environment among members of the hunting party.
- **Mental tool:** Language evolved primarily as a mode of thought and was subsequently co-opted for communication purposes, thereby functioning as a prominent exaptation.[21]
- **Motherese:** The need for communication between mothers and their children, especially during the periods the latter had to be separated from the former during food collection, led to the development of language.
- **Sexual selection:** Language served as an ornament of sorts that enabled females to assess the fitness of males.
- **Coevolution:** Language ought not be viewed in isolation, given that it could have coevolved with other faculties that are either unique or highly pronounced in humans, such as tool making, prosocial behavior, and MTT.

It is also possible to classify models depending on what was the essential feature embodied in hypothetical "protolanguages" that eventually facilitated the emergence of languages (Bickerton 2014). Perhaps the best known among them are the lexical protolanguages comprising a sizable vocabulary (lexicon) but lacking in complex syntax. A second set of models draws inspiration from the gestures used by great apes and thereby proposes that protolanguages were gestural. A third class of models, with origins traceable to Darwin, suggests that protolanguages were musical (Mithen 2006). To wit, the capacity for vocal learning executed other roles, such as demarcating territory or facilitating sexual selection (analogous to birds), before being exapted into the precursors of language. A combination of vocalizations and gestures may have constituted the components of mimetic protolanguages

21. The issue of whether language's primary function is communication or internal thought may prove to be a false dichotomy for the simple reason that language is arguably needful in both contexts.

that resemble pantomime in humans and apes. Fitch (2017) has proposed that the evolution of language unfolded across four stages: (1) vocal learning ($\lesssim 4$ Ma), (2) mimesis (< 1.8 Ma), (3) propositional statements and semantics ($\lesssim 0.5$ Ma), and (4) syntactic structure ($\lesssim 0.2$ Ma).

We will not delve into a potential timeline for the evolution of language in humans because tracing many of the intermediate steps remains very difficult at this juncture. Nonetheless, a couple of points do merit a mention. The FOXP2 gene in chimpanzees and humans differs only in the coding positions of two amino acids. This has been argued to be important because of the putative importance of FOXP2 in the development of oral motor functions and speech. Yet, it is equally important to recognize that the FOXP2 gene constitutes only one of several language-related genes. From a physiological standpoint, a region known as the inferior frontal gyrus (including the famous Broca's area) in the brain, associated with language processing functions, is much more expanded in humans relative to other primates. Collectively, these reasons might explain why chimpanzees do not have the same degree of vocal control and linguistic abilities evinced by humans.

The FOXP2 gene in Neanderthals (*Homo neanderthalensis*) is essentially identical to that of modern *Homo sapiens,* which has been construed by some scientists as signifying the presence of language in the former species; it should, however, be realized that this hypothesis remains unproven as of now. Finally, the earliest evidence of symbolic art, perceived as a proxy for language in certain circles, is manifested in the archaeological record around 0.1 Ma. If we accept the premise that symbolic art was reflective of a society with full-fledged language, the former would effectively date the emergence of the latter to ~ 0.1 Ma at the minimum.

3.8.3 Cumulative cultural evolution

Cultural evolution, whose twin foundations are arguably innovation and selective vertical transmission of information, complements genetic evolution. There is growing consensus—principally dating from the latter half of the twentieth century, albeit with deep roots wending back to the writings of Charles Darwin—that culture has profoundly sculpted the trajectory of hominin evolution. Notable twenty-first century overviews of this subject include Richerson and Boyd (2005), Mesoudi (2011), Henrich (2016), Laland (2017), and Heyes (2018). Human culture has not evolved in strict

isolation, of course, since it has been indelibly modulated by natural and sexual selection (Barkow et al. 1992; G. Miller 2000; Buss 2019) and has shaped them in turn. In recent times, mechanistic formulations that seek to establish connections between cognition, culture, information theory, and thermodynamics have also sprung up (Veissière et al. 2020).

A brief recapitulation of the major consequences of cultural evolution is in order. First, we observe that social learning and cultural transmission function akin to that of natural selection, thereby resulting in the generation of distinctive cultural adaptations that are dependent on the environment, transmittance of successful variants over generations, and accumulation of cultural traits over time. Second, the analogy between cultural and genetic evolution is not exact because the former can either be horizontal (transfer to unrelated individuals) or oblique (transfer to relatives) as opposed to just vertical (transfer from parent to offspring).[22] Furthermore, cultural transmission is distinguished by rapid and flexible forms of adaptation that could occur throughout the entire lifetime of individuals. Finally, the genetic and cultural modes of evolution do not unfold in isolation. Instead, there is mounting evidence for gene-culture coevolution, especially in birds, primates, cetaceans, and obviously humans (Whitehead et al. 2019). The archetypal example for this process is lactose tolerance in humans. Milk did not constitute a significant component of the diet after infancy for the majority of human history, and adults did not typically produce the enzyme lactase for digesting milk. However, with the advent of cattle domestication and milk consumption (a pivotal milestone in human culture), individuals with the requisite genetic mutation for producing lactase in adulthood were consistently favored by natural selection.

Previously, we have seen that social learning exists in mammals, birds, and fishes. It is therefore essential to ask ourselves: In what respects is human culture unique in comparison to (nonhuman) animal cultures? At first glimpse, it does seem accurate to contend that no other animal culture—at least on the basis of our current understanding—approaches the complexity, diversity, and scale of human culture. But this bromide ought not be mistaken for an authentic insight, as it only reflects the external characteristics

22. This statement is evidently an idealization since it neglects the phenomenon of horizontal gene transfer, which is quite common in prokaryotes and relatively less so in eukaryotes.

and not the actual mechanism by which these traits arose. A question of this scope does not have a single answer, in all likelihood. Several scientists have articulated the hypothesis that humans are endowed with an innate capacity for cumulative culture. However, this stance is not universally accepted as others have invoked individual cognitive abilities for explaining the prima facie uniqueness of human culture; the reader may consult Pinker (2010) for more details. The analogy of a ratchet has been invoked to explain how cumulative culture functions: high-fidelity information transfer is facilitated by social learning mechanisms with low error rates, thus opening avenues for successive modifications of cultural traits and thereby ratcheting up their complexity and efficiency.

What is indisputable, though, is the corpus of evidence showing that human culture is cumulative. From a historical standpoint, many purported "revolutions" actually built on prior contributions. The cotton gin patented by Eli Whitney in 1794, for example, conceivably drew inspiration from the charkha machines that were being used in India several centuries ago. Laboratory experiments involving both children and adults have illustrated how simple technologies or strategies are refined, enhanced (in terms of complexity and efficiency), and transmitted by humans across laboratory generations. In contrast, when we contemplate animals, there is scarce evidence that animals display cumulative culture. There are reasons to believe that social learning across generations does happen in certain primates and birds (Whiten 2019), but it remains an open question whether innovations introduced by individuals are genuinely assimilated into groups and cultural innovations in one domain are modified and adopted in a different domain with some degree of regularity. Thus, in toto, the evidence for cumulative culture in nonhuman animals is rather weak at the moment, although we wish to emphasize that considerably more empirical evidence is necessary in order to definitively resolve this vital issue (Mesoudi & Thornton 2018).

A number of reasons have been proposed to explain why the propensity toward cumulative culture is ostensibly unique to humans. The first class of explanations could be termed *cognitive*. Some of them postulate that humans have a higher degree of creativity and innovation in comparison to other species, with the latter being characterized as more conservative or limited in terms of behavioral flexibility. Given how important accurate social learning mechanisms are for cumulative culture, the ability to imitate (beyond naive copying) or teach others is clearly significant; it has therefore been argued that humans score higher than other animals in this arena. It is

evident that imitation or teaching must be accompanied by filtering strategies to ensure the promulgation of maladaptive traits is minimized. From a physiological perspective, notable distinctions exist in the neural circuits of nonhuman primates and humans—which contribute to the sophisticated object-manipulation, sensory-perception, and information-parsing capabilities of the latter—that we shall not address here; this burgeoning subject is reviewed in Stout and Hecht (2017).

The final trio of models within the first category are intimately linked to phenomena encountered previously: complex symbolic communication (language), theory of mind, and prosocial behavior. All of these crucial features are theoretically capable of amplifying, either directly or indirectly, expedient information transmission across generations with concomitantly low error rates. In this context, recent phylogenetic studies indicate that a combination of large brains, long reproductive life spans, and prosociality promoted cultural evolution in primates. In turn, the latter is presumed to have exerted a positive feedback on cognitive abilities, life span, and brain volume, consequently driving the coevolution of all these characteristics.

The second category is oriented toward classifying and analyzing the social learning strategies that dictate when, what, and to whom information is transmitted. The first factor that has been identified is conformity. A high degree of conformity can hinder novel innovations, while it is conceivable that a high degree of nonconformity would lead to chaotic social learning. Hence, there might exist a Goldilocks zone of weak conformity that favors the transmission of beneficial variations.[23] Likewise, copying either successful strategies or individuals in a selective manner (selective copying) has proven to be more effective at information transmission in comparison to indiscriminate copying (Kendal et al. 2018).

The third category of hypotheses are predicated on differences in social structure (Derex & Mesoudi 2020). Primates are often characterized by modestly sized groups comprising one dominant individual or a few of them. This tendency toward monopolization could prevent others from accessing resources or instigating inventive behavior, thus contributing to the suppression of innovation. It should also be noted that demographics, geography, and ecology play important roles in cultural transmission.

23. The motif of the Goldilocks zone is one that will recur throughout this book in a multitude of settings and forms.

Theoretical and empirical studies have found that an increase in group size is positively correlated with higher cultural complexity, indicating that larger population sizes may have stimulated the acceleration of cultural transmission (cf. Vaesen et al. 2016). In the same spirit, both models and data apparently concur that the number of individuals that a given animal can copy regulates the persistence of stable cultures with high-fidelity transmission. Therefore, in conjunction with sizable populations, a high degree of social connections also appears to be either indispensable or desirable for cultural complexity.

3.8.4 Advanced tool construction and use

One of the defining traits of humans that immediately springs to mind is our complex, large-scale technology, to the extent that the phrase "Man the Tool-Maker" was adopted as the title of a fairly influential book by the English anthropologist Kenneth Oakley (1968). The cognate phrase *Homo faber* (i.e., human as *maker*), also widely employed in common parlance, has an even longer history, including contributions from distinguished philosophers such as Henri Bergson and Hannah Arendt. A number of perspectives ranging from the physiological to the cultural have been brought to bear on the question of why humans evolved highly sophisticated tool-making skills. The reader should refer to Ambrose (2001), Johnson-Frey (2004), Vaesen (2012), and Barham (2013) for in-depth analyses of this subject.

An appropriate starting point is the differences in hand-eye coordination between humans and chimpanzees. There are reasons to believe that humans score higher on this metric. Although chimpanzee muscle outperforms human muscle by a factor of ~ 1.35 in terms of maximum power output, humans appear to possess relatively intricate motor control. In proportional terms, more neural tissue is dedicated to controlling hand movements in humans compared to chimpanzees. Moreover, an area of the human brain located in a region known as the left anterior supramarginal gyrus (aSMG) appears to play a key role in observing and reasoning about tool use and in its actual execution (Orban & Caruana 2014). Interestingly, this area does not appear to be activated in macaque monkeys, thus potentially signifying the importance of this area in humans.

Humans are seemingly more adept at reasoning about the causes behind external phenomena compared to nonhuman primates. This is manifested in two different forms. The first entails the ability to perceive correlations

between causes and effects and thereby arrive at inferences about the underlying causal relationships (inferential reasoning). The second involves the ability to recognize that the same causes may underpin two outwardly distinct events (analogous reasoning). This feature is important vis-à-vis tool use because possessing a deeper understanding of causal mechanisms enables a given species to construct more effective tools and harness natural energy sources more efficiently. There is also mounting evidence that humans evince a greater propensity for conceptualizing and internalizing the specific functions associated with a particular tool. This facility enables us to utilize tools depending on the context and paves the way for the development of more complex tools. It is worth remarking that the left inferior parietal cortex (including the left aSMG) in the human brain has been recently linked with reasoning about the properties of physical objects (Reynaud et al. 2016), thereby playing a key role in mechanical knowledge and tool use.

Humans are also known for their greater executive control, which comprises several distinct components, including self-control, triggering select behaviors in the absence of external stimuli, foresight, and monitoring action directed toward some long-term goal (M. N. Muller et al. 2017). All of these components are advantageous in tool construction and use. For instance, it has been suggested that monitoring ongoing action, and perhaps foresight, was necessary in the construction of certain stone tools (e.g., hand axes) as the final structure had to be envisioned, and progress with regard to its construction had to be closely monitored throughout to avoid mistakes. Similarly, foresight would have been necessary in constructing animal traps since it necessitates planning ahead and entails knowledge of animal routes. While the relevance of executive control is plausible, many of the individual features—foresight, self-control, and long-term monitoring—are documented to an extent in nonhuman primates. However, some of these traits (self-control specifically) display positive correlations with brain volume, implying that they could have been pronounced in humans and their ancestors.

Moving ahead, there exist several cognitive capacities linked to tool use that we have already encountered. One of the most prominent is social learning. As we saw earlier, social learning in humans is possibly unusual in that it has been associated with selectivity, progression beyond rote copying and learning, and cumulative transmission of information. A closely related ability is teaching, and it seems reasonable to presume that active pedagogy is more prominent in humans relative to other primates. Social learning and

teaching are significant because they can facilitate high-fidelity transfer of knowledge about tool use and construction. The next broad class that we wish to highlight is loosely termed "social intelligence" by some authors and includes ToM, reciprocal altruism, goal-sharing, and prosocial behavior (Sterelny 2012).

ToM has been invoked to explain the divorce of tool production and usage, i.e., individuals could opt to construct a particular tool even if they do not need it themselves because they recognize that others require it. Reciprocal altruism, roughly understood as tit for tat, is considered a plausible candidate for explaining the antecedents of cooperative behavior (Trivers 1971; Kurzban et al. 2015).[24] In turn, cooperation may engender specialization and division of labor when it comes to tool manufacture and gathering raw materials. Finally, goal sharing is valuable in mitigating the costs arising from the production of complex technologies by distributing them across groups. As far as this book is concerned, there exists fairly compelling evidence for social intelligence in nonhuman primates (and possibly some birds), but ostensibly not to the same degree as that of humans. Finally, some hypotheses contend that language facilitated the construction of sophisticated tools by improving communication channels and information transfer, whereas others propose the converse or suggest that they coevolved.

3.8.5 A brief timeline of hominin evolution

At the outset, we emphasize that explicating the history of hominin evolution is a tremendously complicated endeavor owing to the relative paucity of data, plenitude of interpretations, and the whirlwind pace of progress in paleoanthropology and paleogenomics. In light of these major hindrances, the ensuing account should be viewed in the spirit of a plausible narrative rather than a compendium of definitive truths. Bearing the selfsame caveats in mind, Klein (2009), Reich (2018), Coolidge and Wynn (2018), and Manning (2020) each provide a cogent tour d'horizon of the emergence and evolution of hominins.

---

24. It was conventionally held that chimpanzees did not exhibit signatures of this trait, but recent experiments arguably indicate that they do return favors to conspecifics (Schmelz et al. 2017).

It is natural to commence our story with the last common ancestor of chimpanzees and humans, which immediately confronts us with an obstacle. The process of speciation to yield the ancestors of humans and chimpanzees was potentially complex and long-drawn. It is therefore not surprising that the proposed timing of this split ranges widely (between ~ 4 and 13 Ma), with current studies favoring a divergence time of $\gtrsim 7$ Ma (Muller et al. 2017; Chintalapati & Moorjani 2020); for example, the analysis of primate genomes and molecular clock rates by Besenbacher et al. (2019) suggested that the split in chimpanzees and humans occurred around 6.6 Ma. Species such as *Sahelanthropus tchadensis* (~ 6 to 7 Myr old) and *Orrorin tugenensis* (~ 5 to 6 Myr old) are viewed in some quarters as candidates for the last common ancestor of chimpanzees and humans, or close relatives thereof.

The next major signpost in our chronology is the genus *Ardipithecus,* which was extant by 4.5–5.5 Ma. Many aspects of *Ardipithecus* remain mysterious and unresolved, but they deserve to be mentioned nonetheless on account of their prospective importance. First, *Ardipithecus* is conceivably one of the earliest hominins documented hitherto in the fossil record after the human-chimpanzee divergence (White et al. 2015). Second, a number of studies indicate that *Ardipithecus ramidus* was one of the first, if not the very first, facultative bipeds on the basis of the anatomical structure of its lower back, pelvis, and foot. Bipedalism has been associated with a number of advantages ranging from improved field of vision and lower energy expenditure during locomotion to freeing up the hands, thus enabling them to be utilized for other activities (e.g., carrying food). More controversially, *Ardipithecus* is believed by some to have possessed rudiments of prosocial behavior and singing abilities (Clark & Henneberg 2017).

The genus *Australopithecus* constitutes one of the earliest definitive hominins. The exact timing of its origin remains unclear, but there is fossil evidence indicating its existence at least 4.2 Ma. *Australopithecus* comprises several species, some of which may eventually be classified under a different genus, and persisted for roughly 2.5 Myr. It is widely believed that *Australopithecus* was bipedal on the basis of footprint morphology, but its brain size was comparable to that of chimpanzees. The diet of *Australopithecus* involved food derived from grasses and other plants in savannas, which would have differed from the diet of apes. One of the most striking discoveries from this period concerns the evidence for stone tools. Bones recovered from Dikika, Ethiopia, dating back to ~ 3.4 Ma allegedly reveal marks of stone tools utilized in removing flesh and extracting bone marrow (McPherron

et al. 2010), although this evidence is not universally accepted. Perhaps more strikingly, stone tools assigned a date of ~3.3 Ma were discovered at Lomekwi, an archaeological site in West Turkana, Kenya (Harmand et al. 2015). It is suspected that the Lomekwi tools were manufactured by a hominin species called *Kenyanthropus platyops,* whose exact relationship with the contemporaneous *Australopithecus* genus remains riddled with uncertainty.

The origin and timing of the genus *Homo,* which includes humans (*Homo sapiens*), is still extensively debated. A partial mandible (jawbone) christened LD 350-1 dating from ~2.8 Ma from the Afar Regional State, Ethiopia, appears to exhibit characteristics of both *Australopithecus* and *Homo* but is perceived in some quarters as an early specimen of the latter on morphological grounds. LD 350-1 illustrates the difficulties associated with defining the genus *Homo* in an unambiguous manner and underscores the necessity for more rigorous analyses (Kimbel & Villmoare 2016). *Homo* comprises several early species, whose relationships with each other, and even with the genus itself, are decidedly unclear. Of them, the best known is *Homo erectus,* which originated approximately 2 Ma in Africa (Herries et al. 2020), but its extinction date has proven harder to pin down. Other early variants of *Homo* have been distinguished from *Homo erectus* utilizing criteria such as dental and facial structure as well as brain and body size.

Some of the key aspects that distinguished early non-*erectus Homo* (henceforth *Homo*) and *H. erectus* from *Australopithecus* are worth highlighting (Antón et al. 2014). The mean brain size of the species *Australopithecus afarensis* (from the genus *Australopithecus*) was ~478 cm^3, whereas that of early *Homo* and *H. erectus* were ~629 cm^3 and ~863 cm^3, respectively. However, not all early *Homo* fossils have the same brain size; one specimen of *Homo habilis* (OH 7) appears to have a cranial volume nearly equal to that of *H. erectus*. Second, the degree of sexual dimorphism in terms of body size decreased in early *Homo* relative to *A. afarensis,* which has been interpreted as evidence for relatively prosocial behavior and high cooperation in the former. Third, in broad terms, the diet of early *Homo* was possibly more diverse than that of *A. afarensis* and may have involved a higher proportion of meat. Finally, partly on account of larger body and brain sizes, the ontogenetic (developmental) periods for early *Homo* or *H. erectus* were probably higher than that of *A. afarensis* on average.

One of the most profound innovations introduced by genus *Homo,* which is potentially attributable to *H. erectus,* was the control of fire (Gowlett

2016; Wrangham 2017). The exact timing of when humans discovered the ability to control fire is unknown—there is increasingly solid evidence that it might have occurred as early as ~ 1.5 Ma, although the earliest unambiguous evidence for fire only emerges nearly 1 Myr afterward. The strongest support for this claim is arguably from the site FxJj20 AB in Koobi Fora, Kenya. Analysis of the soil, bone fragments, and spatial patterns indicates that the heated archaeological material is plausibly anthropogenic in origin (Hlubik et al. 2019). A number of advantages stem from the ability to control fire (Wrangham 2009). For starters, it provided a ready means of combating the cold, thereby permitting the geographical expansion of hominins to settle new habitats, in principle.

Of greater relevance, however, is the proposal that cooked food yields more energy per unit of mass, regardless of whether the food is derived from plants or animals (Carmody & Wrangham 2009); further research is necessary in order to conclusively validate or refute this hypothesis (Dufour & Piperata 2018). In addition, fire would have assisted in the sterilization of harmful bacteria (A. R. Smith et al. 2015). Collectively, a richer diet—whether cooked or not—would have enabled the sustenance of large brain and body sizes that are automatically accompanied by increased energy requirements (Zink & Lieberman 2016). From a cultural standpoint, the hearth may have brought together different members of hominin societies and facilitated greater social interactions and cooperation. Finally, jumping far ahead in time, harnessing fire played a vital role in the large-scale expansion of agriculture and mechanization. In closing, we caution that the aforementioned advantages arising from fire control should be balanced against the drawbacks, such as the costs of collecting fuel (A. G. Henry et al. 2018).

Let us now jump ahead in time to ~ 0.55-0.77 Ma (Stringer 2016), while being cognizant of the fact that the brain size of hominins expanded in the interim period, reaching a volume of ~ 1250 cm^3, which was nearly comparable to that of modern humans (~ 1350 cm^3). The reason behind choosing this temporal interval is because the last common ancestor of *Homo sapiens* (modern humans) and *Homo neanderthalensis* (Neanderthals) may have lived in this period, on the basis of the analysis of nuclear DNA from fossils of hominins recovered from Sima de los Huesos in Spain (Meyer et al. 2016). It is not yet clear what species the last common ancestor belonged to, but two of the primary contenders appear to be *Homo heidelbergensis* and *Homo antecessor*. Neanderthals have attracted a great deal of interest ever since the nineteenth century, and one of their defining features is the large

brain size (~ 1350 cm^3). It is known that interbreeding between Neanderthals and modern humans took place and that extinction of this species was mostly complete by ~ 40 kyr ago, apart from isolated pockets. A number of prospective causes have been advanced for their demise, ranging from climate change to being outcompeted by *H. sapiens* (Stewart & Stringer 2012).

Next, we turn our attention to anatomically modern humans, who are conventionally distinguished from other hominin species on the basis of their mandibular, dental, and facial morphology. The oldest robust evidence to date for *Homo sapiens* in this category is from ~ 300 kyr ago and comprises fossils from Jebel Irhoud in Morocco (Richter et al. 2017; Galway-Witham et al. 2019). This timing is apparently consistent with analyses of ancient South African genomes, which suggest that the earliest divergence in modern human populations occurred around 260 to 350 kyr ago (Schlebusch et al. 2017). At this stage, it should be appreciated that anatomical modernity is *not* the same as behavioral modernity. While early *H. sapiens* achieved present-day brain size by ~ 300 kyr ago, the shape of the human brain acquired its current structure merely 35 to 100 kyr ago (Neubauer et al. 2018).

Brain shape, besides the standard rubric of overall size, matters a great deal when it comes to cognitive functions because modern humans are characterized by their globular brains with bulging parietal and cerebellar areas. The parietal regions are vital for sensory perception, self-awareness, long-term memory, and tool use, to name a few. Likewise, the cerebellum plays a crucial role in motor control, spatial orientation, social cognition, and possibly language. Hence, it is reasonable to suppose that progressive physiological changes in the brain went hand in hand with the gradual development of cognitive abilities in humans, eventually culminating in the *Upper Paleolithic revolution* approximately 40 to 50 kyr ago, characterized by the seemingly rapid appearance of behavioral modernity marked by symbolic cognition and technological complexity (Bar-Yosef 2002). The premise of a bona fide, dramatic, and discontinuous "revolution" has, however, attracted spirited arguments *pro et contra* this hypothesis: the reader may compare Klein (2000, 2008) on the one hand with McBrearty and Brooks (2000) and Kissel and Fuentes (2018) on the other for comprehensive, but opposing, analyses of this paradigm.

The last triad of developments that we wish to highlight are very recent in comparison to the timescales encountered earlier. The first two, the

systematic domestication of plants and animals, arose roughly at the same time (~ 12 kyr ago) and constitute the backbone of the Neolithic revolution (Zeder 2011); strictly speaking, a few species (e.g., dogs) were brought into the fold earlier. The domestication of animals not only yielded produce such as milk and meat but also offered the chance to exploit their labor power (e.g., plowing); other benefits to humanity included utilization of the keen olfactory sense of dogs during hunting. Agriculture had a number of positive effects primarily stemming from the fact that the systematic cultivation of plants drove an increase in food productivity and permitted a stable and sedentary lifestyle, thus enabling rapid and sustained population growth. Additionally, it spawned several key innovations (e.g., agricultural implements) and had major impacts on social structure and the division of labor. Yet, one must not fall into the trap of viewing agriculture as an unblemished agent of positive change, given that it led to nutritional deficiencies, concomitant diseases, and possibly socioeconomic inequality in tandem.

The final stage of this triad that brings our winding journey to a close is the industrial revolution that originated a few centuries ago and remains underway. The exponential growth in technology has transformed human societies as well as the Earth's climate and biosphere, to the point where a new geological epoch, the Anthropocene, has been coined (Ellis 2018). As with agriculture, we observe that the industrial revolution has stimulated much progress but has simultaneously engendered grave challenges such as anthropogenic climate change.

3.8.6 Extraterrestrial technological intelligence

Our prior discussion has illustrated the fact that humans are characterized by an unusual, if not unique, combination of features such as language, high cooperation, cumulative cultural transmission, and sophisticated tool construction and use. On the surface, the gap between humans and nonhumans comes across as being considerable, thereby suggesting that the transition to technological intelligence is relatively unlikely. However, certain counteracting forces ought to be taken into account before rushing to judgment.

First, many of the aforementioned traits do not exist in nonhuman primates sensu stricto, but there is fairly compelling evidence for their precursors. Second, the above characteristics ought not be perceived as

wholly independent, seeing as how many of them probably coevolved due to pleiotropic effects. In this spirit, it is more compelling to interpret the advent of *Homo sapiens* vis-à-vis the emergence of a distinctive "socio-cognitive niche" (Whiten & Erdal 2012) that encompasses language, prosocial behavior, theory of mind, and cultural transmission, among others. Finally, there exist sufficient, although not definitive, grounds for presuming that the temporal course of hominin evolution was neither wholly discontinuous (characterized by dramatic jumps) nor completely gradual (total absence of transitional phases). Instead, the evolution of modern humans arguably exhibits an inherently *mosaic* character originating from time-varying evolutionary rates and the emergence of new adaptations in different time periods (R. A. Foley 2016).

Hitherto, we have implicitly avoided a vital question. Why did most, if not all, of the major breakthroughs in hominin evolution unfold in Africa over the past ~ 10 Myr? To put it differently, why did none of the primates in other parts of the world evolve into hominins? For instance, the split in Old World monkeys (*Catarrhini*) and New World monkeys (*Platyrrhini*) is presumed to have occurred ~ 40 Ma, and yet none of the latter evolved into species with functional capabilities equivalent to the hominin lineage. The answer might stem, at least in part, from the strong selection pressures arising from the unusual environmental variations in Africa. In particular, the climate of East Africa was characterized by transient periods of extreme precipitation and aridity that punctuated the overall long-timescale drying trend. These rapid shifts in climate were accompanied by changes in vegetation and fauna, consequently serving as potential drivers of hominin evolution (Maslin et al. 2015). At the same time, however, climate change in eastern (or southern) Africa cannot be viewed in isolation, as it was intimately linked to fluctuations in global factors including stellar insolation, extent of polar ice, and tropical ocean temperature gradients (Levin 2015).

Thus, in seeking to generalize the sole example on Earth, it would appear at first sight that a minimum of two criteria might be necessary for technological intelligence: the emergence of a socio-cognitive niche and rapidly (but not overly) fluctuating environmental conditions that apply selection pressure. Even though the number of criteria is small, we should not jump to the conclusion that technological intelligence has a high probability of occurrence because each of these conditions involves many implicit constraints. In contrast to the high, albeit nontechnological, intelligence that arose several times on our planet, technological intelligence has ostensibly

evolved only a single time. Strictly speaking, we lack sufficient information to assess the likelihood of extraterrestrial technological intelligence. Notwithstanding this fact, on the basis of what we have delineated thus far, it seems conceivable that technological intelligence could be quite uncommon on other worlds.

3.9 PARADIGMS FOR MAJOR EVOLUTIONARY EVENTS

Up to this stage, we have chronicled the major evolutionary events (MEEs) in the history of our planet and explored their likelihood of occurrence on other worlds. However, it must be recognized that there is no canonical list of MEEs, although certain evolutionary breakthroughs are common to all studies. Hence, it is worth highlighting a few select studies that have attempted to delineate the MEEs on our planet using various methodological approaches.

3.9.1 Major evolutionary transitions

The most famous classification of MEEs was arguably proposed by John Maynard Smith and Eörs Szathmáry in their seminal work *The Major Transitions in Evolution* (1995), which was updated two decades later (Szathmáry 2015); for related perspectives, see S. A. West et al. (2015) and van Gestel and Tarnita (2017). We will refer to the breakthroughs in this list as major evolutionary transitions (METs).

A number of theoretical concepts are held to underpin the METs. For starters, METs reflect the tendency of biological entities to have undergone complexification over time. At the same time, it has been claimed that this statement should not be misconstrued as being synonymous with orthogenesis (progressive evolution)—that is, the notion that biological evolution is characterized by an intrinsic directionality.[25] The observed increase in complexity does not change the fact that microbes have, and still continue to, dominate Earth's ecosystems. Next, the METs are characterized by the fusion of lower-level evolutionary units to yield higher-level units. A classic example in this regard is eukaryogenesis, in which two (or more)

25. This point bears repeating since there are no reasons to suppose that there exists an inevitable path from microbes to technological intelligence.

prokaryotes underwent endosymbiosis to yield eukaryotes. The division of labor also plays a key role in METs, given that the combination of specialized functions performed by lower-level units can translate to an increased fitness advantage for the higher-level units. Furthermore, novel inheritance mechanisms arise during METs. One of the best examples we have encountered earlier is culture, which facilitates the transmission of information through a new channel that parallels and supplements genetic inheritance.

The list of METs has undergone some subtle changes from its inception. To begin with, the original METs proposed in Smith and Szathmáry (1995) are as follows:

- From replicating molecules to the enclosure of populations of these molecules within compartments.
- From individual replicators to the origin of chromosomes (i.e., linked molecules that replicate together).
- From the RNA world, in which RNA served as both the repository of genetic information and enzymatic activity, to the dual world of DNA and proteins that fulfill these respective functions.
- From prokaryotes to eukaryotes, which is expected to have involved endosymbiosis.
- From asexual (clonal) reproduction to sexual populations.
- From single-celled protists to complex multicellularity (i.e., animals, plants, and fungi).
- From single individuals to the formation of colonies (i.e., eusocial societies).
- From nonhuman primates to humans.

As remarked earlier, this list has been subject to revisions by many authors. It is instructive to consider the updated list proposed by Szathmáry (2015):

- Origin of protocells. This stage can be broken down in several stages such as the emergence of autocatalytic networks, robust replicators, and lipid vesicles.
- Origin of the genetic code and prokaryotic cells. This transition entails the formation of the dual DNA-protein world that we find in virtually all organisms.

- Origin of eukaryotic cells. As noted in Section 3.5, a number of distinct components demarcate eukaryotes from prokaryotes including organelles, phagocytosis, and sexual reproduction.
- Origin of plastids. It has already been pointed out that this endosymbiotic event took place after eukaryotes had acquired a more or less modern form in other respects.
- Origin of complex multicellularity in plants, animals, and fungi.
- Origin of eusocial animal societies, which can be roughly envisioned as the emergence of superorganisms.
- Origin of societies with natural language. An important point is that the emergence of language probably took place in conjunction with the evolution of other traits.

In addition to these innovations, some authors have proposed including an ongoing recent transition entailing the origin of electronic cultural transmission.

Despite the undeniable influence of METs in shaping the way we look at macroevolution, a number of criticisms have been leveled against it; a cogent summary is furnished by O'Malley and Powell (2016). One of the most well-known among them pertains to the perceived lack of unity among these transitions. The transition to human societies has been viewed as particularly problematic because of issues associated with viewing human societies as units of selection. Moreover, some of the METs arose more than once (e.g., complex multicellularity), whereas others did not. Second, METs have been criticized for implicitly justifying orthogenesis, although this stance was disavowed by the original authors themselves. Finally, the list of METs does not account for metabolic innovations, which have undoubtedly played a huge role in shaping Earth's biosphere.

3.9.2 Megatrajectories

The concept of megatrajectories was introduced by Knoll and Bambach (2000) to explore the issue of directionality in evolution. The simplest paradigm, one that has been invoked by the likes of Stephen Jay Gould, to explain a rise in complexity is diffusion: it is self-evident that the variance increases linearly with time. There is, however, a core puzzle that presents itself. Can a process of simple diffusion capture the entire essence of evolutionary change and explain the observed increase in variance completely?

Knoll and Bambach suggest that diffusion does not suffice to explain all of the observed trends. Each megatrajectory introduces new evolutionary entities that utilize novel functional traits to acquire new resources, thus leading to an expansion in the spatial extent and diversity of ecosystems (ecospace). However, the maximum amount of diversity achievable within a given megatrajectory is bounded from both above and below, and surmounting the upper bound requires the emergence of a new megatrajectory via the accumulation of traits.

A total of six megatrajectories were proposed, each of which corresponded to an increase in ecological complexity, outlined below.

- The origin of life culminating in the Last Universal Common Ancestor (LUCA) of extant life.
- The diversification of metabolic pathways in prokaryotes and the concomitant ecological changes (e.g., effects on biogeochemical cycles). In view of the fact that metabolic innovations arose at different points in time, specifying a precise timing for this megatrajectory is difficult. However, geological proxies indicate that several metabolic pathways of biogeochemical import had evolved by approximately 3.4 Ga.
- The evolution of the eukaryotic cell.
- The origin of complex multicellularity, which was characterized by cell differentiation enabling the division of labor.
- The takeover of the land by embryophytes (land plants), which also facilitated the diversification of animals.
- The emergence of technological intelligence, which facilitated the dramatic increase in the global exploitation of ecosystems.

3.9.3 Energy expansions

Judson (2017) proposes a schema in which the history of the Earth was divided into different energetic epochs that entailed the evolution of organisms capable of exploiting a new source of energy. Although each energy expansion led to an increasing inventory of energy sources, they did not eliminate any of the previously existing sources. Each energy expansion also led to major transformations of the environment, which in turn had significant evolutionary consequences. We will briefly describe the energy

expansions, but the reader may consult Judson for the rationale behind their selection.

- Geochemical energy (e.g., at the hydrothermal vents encountered in Chapter 2) could have served as the first energy source for life.
- Sunlight is a plentiful source of energy, owing to which the evolution of photosynthesis (first anoxygenic and then oxygenic) enabled microbes to tap stellar energy.
- Oxygen, formed as a byproduct of oxygenic photosynthesis, provided a rich energy source for organisms via aerobic respiration.
- The ability to acquire energy by actively hunting and consuming other organisms proved to be a rich source of energy to animals; this led to the energy epoch of flesh eating.
- Fire provided a potent and controllable source of energy with manifold advantages, as outlined in Section 3.8.5.

3.9.4 MEEs on other worlds: A synthesis

Until now, we have focused on individual MEEs that unfolded on Earth and discussed their likelihood of occurrence elsewhere. However, what we really need to ask ourselves is the following question: What are the essential steps leading from the origin of life to technological intelligence on extrasolar worlds? It goes without saying that a definitive answer is impossible, but it is worth presenting our personal take on a potential solution. It is our thesis that, including both abiogenesis and technological intelligence, a total of five critical steps must be surmounted for technological intelligence to emerge. Our position is that these five MEEs might possess a certain degree of universality, but a couple of essential clarifications are necessary. This statement does *not* imply that (1) the realization of these five MEEs is easy or (2) the putative technological species will physiologically resemble humans (although they may share common functional traits).

The first step is self-evident: abiogenesis. It is a tautological statement that life cannot get started without the origin of life. We have no way of knowing the actual evolutionary innovations that led up to the emergence of the first lifeforms. Unless the number of available theoretical routes to abiogenesis are very few, it seems highly unlikely that the exact counterparts of the RNA world, and the transition thenceforth to DNA (genetic information) and proteins (metabolism), would exist elsewhere. Once the

first microbes have originated, it is reasonable to expect that they would undergo metabolic diversification.

We are now confronted with the second critical step. In our view, photosynthesis constitutes a natural means of bypassing the energetic bottleneck that would otherwise exist if all microbes had to depend only on geochemical energy. In contrast, stellar energy is much more plentiful and any microbes that evolved mechanisms to harness stellar radiation could gain a significant advantage. Earlier, we saw that at least two different biological pigments arose on Earth for capturing and converting light energy. However, recall that photosynthesis also requires a suitable electron donor. Ideally, this electron donor must be available in plentiful supply to ensure that the ecosystem is not constrained by the abundance of raw materials.

In this regard, on planets loosely analogous to early Earth, with liquid water and atmospheres with sizable CO_2 inventories, it seems reasonable to surmise that H_2O would comprise a plentiful electron donor. The ensuing result, oxygenic photosynthesis, has the advantage of generating O_2 as an energy-rich product that permits aerobic respiration and macroscopic organisms. Finally, oxygenic photosynthesis appears to have originated only once on Earth as a number of distinct criteria had to be fulfilled.[26] Hence, we identify oxygenic photosynthesis as the second critical step and conjecture that its putative generality stems from the plentiful availability of energy (light) and raw materials (water). On the other hand, we caution that oxygenic photosynthesis is *not* a prerequisite for raising atmospheric and oceanic O_2 content. The splitting of water by energy derived from charged particles (Chapter 7) or electromagnetic radiation (Section 4.3.1) can drive the plentiful production of abiotic oxygen in certain scenarios instead, thus potentially creating aerobic environments even sans oxygenic photosynthesis.

The third critical step that naturally springs to mind is eukaryogenesis. Why do we suspect this step is potentially generic to some degree? The reason stems from our current understanding of eukaryogenesis on Earth. There are intriguing, but not conclusively proven, bioenergetic grounds for supposing that the acquisition of mitochondria via endosymbiosis paved

26. The origin of oxygenic photosynthesis is, of course, intimately linked to the evolution of the photosynthetic reaction center, which is itself suspected to have arisen a single time in Earth's history by ~ 3 Ga at the minimum (Orf et al. 2018).

the way for an increase in biological complexity. In other words, one may be inclined to agree with other studies that posit eukaryogenesis—by way of endosymbiosis and phagocytosis (with the precise order being indeterminate)—as a means of bypassing the bottleneck on organismal size and complexity. In the absence of endosymbiosis, it is unclear what other mechanism could yield the requisite complexity. Hence, *if* technological intelligence is the desired final outcome, we suggest that eukaryogenesis potentially constitutes a necessary critical step. Needless to say, our claim does not imply that the endosymbiont and host cell must always correspond to mitochondria and archaea respectively; in actuality, what matters is the existence of their functional analogs. In light of the associated jump in complexity, the probability of this transition might be quite low for reasons documented earlier.

When it comes to identifying our fourth critical step, complex multicellularity is an obvious candidate prima facie. Complex multicellularity is characterized by cell differentiation that facilitates the efficient division of labor and transport of nutrients, among other factors, thereby resulting in enhanced organismal complexity. To put it differently, we deem complex multicellularity as an essential step for bypassing yet another bottleneck on organism complexity and thereby empowering the expansion of life into new ecological niches. While these reasons may seem compelling enough for postulating that this step possesses the desired degree of generality and novelty, we notice that it stands out from the first three in one crucial respect: the trio were ostensibly singular events on Earth, whereas a minimum of six instances of complex multicellularity are known.

However, there is a subtle issue at play, which was already pointed out in Section 3.6.6—namely, distinguishing between animal and complex multicellularity. We note that the former emerged only once, thus ensuring animal multicellularity resembles the three critical steps proposed earlier in this respect. Moreover, a unique combination of properties (macroscopic size, phagocytosis, and motility) enabled animals to operate as effective ecosystem engineers. Lastly, technological intelligence arose within the animal kingdom. In other words, we hypothesize that organisms endowed with the same functionality as animals are essential from the standpoint of potentially evolving into technological species. Before proceeding further, it should be recognized that we are not disregarding the importance of plants (or equivalent autotrophs); the diversification of animals was clearly regulated by the access to land plants on Earth. We are, instead, conjecturing

that the evolution of complex multicellularity in plantlike lifeforms may not prove to be exceedingly difficult; on Earth, the emergence of complex multicellularity occurred in the *Archaeplastida* at least three times.

The final critical step, by tautology, is technological intelligence because we were interested in the question of critical steps leading up to the evolution of technological species. Again, while it would be foolish to presume that technological species would physiologically resemble humans, it is plausible that they share some characteristics in common with humans: the socio-cognitive niche introduced in Section 3.8.6 and dexterous appendages analogous to limbs for tool construction. We also point out that technological intelligence has arisen only once on our planet, thereby preserving commonality with the first four critical steps.

In summary, we have delineated five MEEs that might be generic, to a certain degree, on other worlds that had initial environmental conditions akin to that of early Earth. The five critical steps are (1) abiogenesis, (2) oxygenic photosynthesis, (3) eukaryogenesis, (4) animal multicellularity, and (5) technological intelligence. Unlike the METs, it should be acknowledged that these five MEEs are not unified by underlying theoretical principles. Nevertheless, they do exhibit commonality in multiple respects since each of them (in the context of our planet) (1) was a singular event, (2) drove major ecological reorganization due to induced environmental changes, and (3) facilitated the circumvention of preexisting bottlenecks on complexity. Furthermore, in case either the endosymbiont or host cell during eukaryogenesis was aerobic, it would likely give rise to an appealing logical order wherein the inception of step $X+1$ was contingent on step X, at least insofar as Earth's evolutionary history is concerned.

We find it sufficiently important to belabor the central point that these five steps, even if they are universal to an extent as conjectured, are not guaranteed to emerge in either a deterministic or highly probable manner. For instance, the attainment of sufficiently high oxygen levels to promote the rise of animals may require multi-Gyr timescales, in which case high-mass stars would be ruled out. In the next two chapters, we will delve into select planetary and stellar constraints that can affect the evolution of life on extrasolar worlds. We round off our analysis by cautioning the reader that the proposed five critical steps are merely hypothetical and are not definitively corroborated by empirical evidence. Nonetheless, it is our belief and expectation that they constitute useful benchmarks for envisioning how evolutionary trajectories might unfold on some other worlds.

3.10 THE CRITICAL STEPS MODEL

We will describe the critical steps model introduced by B. Carter (1983) and analyze the results. Before embarking on the discussion of the formalism, it is helpful to outline the basic premises of this mathematical model.

The underlying idea is that there exist n critical (hard) evolutionary steps that must be realized in order to result in a particular outcome, with n being an unknown quantity a priori. It is also assumed that these steps are stochastic and essential in nature. The critical step j is associated with an occurrence rate Λ_j, and the fact that it is intrinsically unlikely is quantified via the criterion $\Lambda_j t_H \ll 1$, where t_H denotes the total habitability interval of a given planet and $j \in \{1, 2, \ldots, n\}$. Finally, the critical steps model is implicitly based on the notion that the evolutionary process can be demarcated into easy and hard steps, with the latter being distinguishable from the former. Naturally, this approach constitutes an oversimplified account of how evolution operates and should therefore be viewed in the spirit of a useful mathematical abstraction. Yet, it is worth appreciating at the same time that several frameworks have identified a discrete number of MEEs; we encountered some of them in Section 3.9.

3.10.1 Mathematical preliminaries

The total probability $\mathcal{P}_{n,n}(t)$ that all of the n steps occur in the interval $(0, t)$ is given by

$$\mathcal{P}_{n,n}(t) = (\Lambda_1 t) \ldots (\Lambda_n t) = t^n \prod_{j=1}^{n} \Lambda_j = \mathcal{C}_n t^n, \tag{3.42}$$

where $\mathcal{C}_n$ is a constant whose value we shall determine shortly. In the above equation, we have assumed that all of the critical steps can be treated as random variables. Now, we impose the constraint that all of these n steps unfold within $(0, t_H)$ on a given planet. In this event, the criterion $\mathcal{P}_{n,n}(t_H) = 1$ must be satisfied, which yields $\mathcal{C}_n = t_H^{-n}$.

The probability distribution function (PDF) for the n-th step in the sequence of n steps, denoted by $P_{n,n}(t)$, is found by differentiating (3.42) and using the above value of $\mathcal{C}_n$. Upon carrying out this calculation, we end up with

$$P_{n,n}(t) = \frac{n t^{n-1}}{t_H^n}. \tag{3.43}$$

Let us now derive the PDF when the r-th step occurs at time t, such that the remaining $n-r$ steps take place after t; they must, however, occur before the end of the habitability interval (t_H). This PDF, expressed as $P_{r,n}(t)$, is computed as follows:

$$P_{r,n}(t) = P_{r,r}(t) \int_t^{t_H} P_{n-r,n-r}(t'-t)\, dt' \tag{3.44}$$

From (3.43), we see that $P_{r,r}(t) \propto t^{r-1}$ and $P_{n-r,n-r}(t'-t) \propto \left(t'-t\right)^{n-r-1}$. Substituting these expressions into the above equation, we obtain

$$P_{r,n}(t) = \mathcal{C}_r t^{r-1} \left(t_H - t\right)^{n-r}, \tag{3.45}$$

with $\mathcal{C}_r$ representing a new constant that must be calculated. This is done by integrating $P_{r,n}(t)$ from 0 to t_H and requiring that the result should equal unity. After simplifying the result, we find that the PDF is given by

$$P_{r,n}(t) = \frac{n!}{(n-r)!(r-1)!} \frac{t^{r-1} \left(t_H - t\right)^{n-r}}{t_H^n}. \tag{3.46}$$

We can now use (3.46) to compute the mean time $\bar{t}_{r,n}$ for the r-th step to occur as shown below.

$$\bar{t}_{r,n} = \int_0^{t_H} t P_{r,n}(t)\, dt = \left(\frac{r}{n+1}\right) t_H \tag{3.47}$$

An inspection of this formula reveals that the average spacing (Δt_n) between two consecutive steps is approximately equal, with

$$\Delta t_n = \frac{t_H}{n+1}. \tag{3.48}$$

The cumulative probability $\mathcal{P}_{r,n}(t)$ for the r-th step to occur in the interval $(0, t)$ is found by integrating (3.46) between 0 and t, which leads us to

$$\mathcal{P}_{r,n}(t) = \frac{n!}{(n-r)!(r-1)!} B\left(t/t_H\,;\, r\,,\, n-r+1\right), \tag{3.49}$$

where B is the incomplete beta function. In the case of $r = n$, we see that (3.49) reduces to $\mathcal{P}_{n,n}(t) = (t/t_H)^n$, consequently leading to our initial result.

3.10.2 Critical steps model(s) for Earth

In Section 3.1, we noted that $t_{H,\oplus} \sim 5.5$–6.5 Gyr; we will restrict ourselves to the more conservative limit of ~ 5.5 Gyr henceforth. Let us posit that the evolution of technological intelligence (i.e., *Homo sapiens*) constitutes the n-th step, with the mean timescale for our emergence being $\bar{t}_{n,n} \approx 4.5$ Gyr. By making use of (3.47) in conjunction with the above values, it can be verified that $n \approx 4.5$. Thus, it would seem reasonable that either a four- or five-step model may constitute the best fit. Several studies have indeed converged on the conclusion that a five-step model is quite plausible (A. J. Watson 2008; B. Carter 2008; McCabe & Lucas 2010).[27]

However, even if we consider a five-step model, many candidates present themselves, some of which we encountered in Section 3.9. We will not explore all possibilities herein but will restrict ourselves to two different candidates that are described in the next paragraph. The putative steps and their timing are presented, but the justification behind the latter will not be elaborated here, as it is outlined earlier in the chapter and in Lingam and Loeb (2019f). In order to assess the plausibility of these models, we introduce the parameter

$$\delta_F = \frac{1}{n}\left[\sum_{r=1}^{n}\left(\mathcal{P}_{r,n} - 0.5\right)^2\right]^{1/2}. \tag{3.50}$$

This makeshift rubric was introduced by McCabe and Lucas (2010). A higher value of δ_F implies that the model is relatively inaccurate for the following reason. In the event that the cumulative probabilities $\left(\mathcal{P}_{r,n}\right)$ approached either 0 or 1, this would tend to increase δ_F. However, at the same time, it would imply that the events are clustered toward the beginning and end, respectively. In both these limits, the model would constitute a comparatively poor fit, since the expectation is that the spacing between events should be roughly even.

---

27. A quantitative treatment of this topic was also undertaken by Robin Hanson, which is accessible through this link: http://mason.gmu.edu/~rhanson/hardstep.pdf.

The first model is based on the METs delineated in Section 3.9.1 but was reduced to five steps by A. J. Watson (2008) and Lingam and Loeb (2019f). The proposed critical steps are (1A) abiogenesis (~ 3.7 Ga), (2A) eukaryogenesis (~ 1.8 Ga), (3A) plastids (~ 1.5 Ga), (4A) complex multicellularity (~ 0.8 Ga), and (5A) technological intelligence (~ 0 Ga). For this five-step model, we obtain $\delta_F \approx 0.068$ after making use of (3.49) and (3.50). The second model we consider is identical to that presented in Section 3.9.4. To recapitulate, the potential critical steps involved are (1B) abiogenesis (~ 3.7 Ga), (2B) oxygenic photosynthesis (~ 2.7 Ga), (3B) eukaryogenesis (~ 1.8 Ga), (4B) animal multicellularity (~ 0.8 Ga), and (5B) technological intelligence (~ 0 Ga). When we calculate δ_F for this five-step model, we end up with $\delta_F \approx 0.029$.

Alternatively, we can choose a six-step model for the sake of comparison. Let us work with the megatrajectories specified in Section 3.9.2. The six megatrajectories are (1) origin of life (~ 3.7 Ga), (2) metabolic diversification of prokaryotes (~ 3.4 Ga), (3) eukaryogenesis (~ 1.8 Ga), (4) complex multicellularity (~ 0.8 Ga), (5) advent of land plants (~ 0.5 Ga), and (6) technological intelligence (~ 0 Ga). For this six-step model, we find that $\delta_F \approx 0.069$, which is virtually identical to the five-step model (1A)–(5A) from the previous paragraph, although both of them are higher than the five-step model (1B)–(5B) by a factor of 2.4. Hence, on the basis of our analysis, there are tentative grounds to suppose that the five-step model propounded in Section 3.9.4 represents the most plausible choice. However, this result should be taken with more than the proverbial pinch of salt for two reasons. First, the timing of most MEEs remains very uncertain, and changing these values will duly impact δ_F and the ensuing conclusions. Second, the mathematical model constitutes an idealized rendition of the evolutionary process (e.g., the demarcation of hard and easy steps), and may therefore fail to account for inherent subtleties.

Finally, our analysis has been predicated on the assumption that the habitability window on Earth will close ~ 1 Gyr in the future, at least inasmuch as complex multicellular organisms are concerned (de Sousa Mello & Friaça 2020). Instead, if we suppose that the habitability interval extends ~ 2 Gyr in the future, our results are altered accordingly. We will not delve into this possibility here, since it has already been explored in Lingam and Loeb (2019f). It was proposed that a six-step model might be more plausible, with humans constituting the fifth critical step. In other words, one critical step may await the Earth in the future: the emergence

of superintelligence—namely, artificial (machine) intelligence that greatly surpasses human intellect (Bostrom 2014)—is a potential candidate.

It is tempting to construe this outcome as signifying that the most abundant group of objects that give rise to signals of life might entail technological equipment in interstellar space, and not biological chemistry on rocky / icy planets and moons. If so, one would also be tempted to conclude that the search for nontechnological (e.g., microbial) life on habitable worlds is misguided. However, we strongly caution that the above line of reasoning is markedly speculative, and detecting signals from putative technological infrastructure interspersed throughout the Galaxy could prove to be immensely challenging, owing to their presumably weak and uncharted signatures.

3.11 CONCLUSION

> In the rank grass crickets are chirping,
> And parasol-trees, startled, let fall their leaves.
> Clouds for stairs, the moon for floor,
> To Heaven the way is blocked by a thousand barriers,
> And floating rafts ply to and fro
> To no avail.
>
> On this night magpies form a star bridge to span the Milky Way,
> Where Cowboy [*Altair*] and Weaving Maid [*Vega*] keep their yearly tryst.
>
> —Li Qingzhao, *The Seventh Day of the Seventh Lunar Month*

In our seeking knowledge of life beyond Earth, two stratagems immediately present themselves: the search for "simple" life (biosignatures) and the search for technological life (technosignatures). As noted at the beginning of this chapter, our likelihood of success in the search for biosignatures and technosignatures is sensitive to two terms in the Drake equation—f_l and f_i—whose significance was explicated earlier. Thus, investigating the feasibility of technological intelligence on other worlds is not only important as an intellectual exercise but also from the standpoint of answering the question, "Are we alone?"

In view of this fact, we have journeyed through the daedal evolutionary history of the only planet that is unequivocally confirmed to host life hitherto, the Earth. Life on our planet has been characterized by the abiding intricate interplay of contingency and convergence. It is thus no exaggeration to contend that the Earth, during its ~ 4 Gyr of hosting life,

has undergone such dramatic changes that it can be envisioned as a succession of *distinct* habitable planets. One of the classic examples in this regard is the atmospheric oxygen inventory: it was virtually nonexistent until ~ 2.4 Ga, after which it persisted at low levels until $\sim 0.6 \pm 0.2$ Ga, when it underwent a further increase to reach modern-day levels; this approximate trend was punctuated by episodic fluctuations during various epochs.

After tracing the course of life on Earth, there are arguably compelling grounds for supposing that evolution transcended a series of bottlenecks on ecosystems through novel breakthroughs. The exact nature, and number, of these major evolutionary events is subject to much debate, but they are widely considered to be less than ten in total. Our own hypothesis is that a total of five major breakthroughs are distinguishable: the emergence of life, oxygenic photosynthesis, eukaryotes, animals, and technological intelligence. Each of these major evolutionary events happened only once on our planet, thus potentially implying that they may have a low likelihood of occurrence on other worlds. At the same time, these innovations enabled the novel utilization and expansion of ecospace, suggesting that they might be endowed with some degree of universality. It bears repeating at this juncture that humans do *not* represent either the end or the zenith of the evolutionary process. Hence, our focus on technological intelligence as the "end point" was primarily governed by our intention to analyze the feasibility of this event occurring on other worlds, not by the expectation that technological species embody the apex of the "Great Chain of Being" (*scala naturae*).

Consider the following simple thought experiment. On the basis of our proposal that the aforementioned five critical steps are uncommon, we assign an equal probability of ~ 0.01 per step on a habitable planet with the requisite environmental conditions for supporting life. In this event, we find that $f_l \sim 10^{-2}$, but $f_l \cdot f_i \sim \left(10^{-2}\right)^5 \sim 10^{-10}$. In other words, while planets with microbial life would be somewhat rare, this expression implies that planets with technological intelligence are extremely so. As such, this may appear to sound a death knell in the search for technosignatures, but this analysis misses an important point. Despite our theorizing in this chapter, it should be humbly accepted that we know very little about how evolution has unfolded on Earth, let alone on other worlds. Hence, it is theoretically conceivable that there are many avenues to technological intelligence, thus raising its likelihood of occurrence on other worlds. In other words, the best policy entails maintaining a healthy balance of skepticism and

open-mindedness and allowing the data from the searches for biosignatures and technosignatures to serve as the final judge.

In conclusion, we wish to return to a point of philosophical import. There has been much hand-wringing about whether the difference between humans and nonhumans is "one of degree and not of kind", as proposed by Darwin. On a different front, there has been speculation about humanity's place in the Cosmos: on the one hand, we have the Copernican Principle, which strongly disfavors the notion that we are "special" in certain respects and on the other hand there is the Anthropic Principle, which is summarized by Brandon Carter, one of its chief architects, as follows: "Although our situation is not necessarily central, it is necessarily privileged to some extent" (1974, p. 291). It seems to us that these stances are not mutually exclusive and therefore represent false dichotomies of sorts. Humans are unique in their own way, just as every other species on Earth and beyond is, regardless of whether they are conscious and intelligent. While this statement is essentially a mere tautology, it nevertheless serves to remind us that expending much energy on debating notions such as *special* and *commonplace* might be counterproductive.

Of more importance, due to its time-critical nature, is the fact that *Homo sapiens* possesses an unparalleled capacity for shaping future trajectories of life on our planet, as exemplified by the following sentiments from Knoll (2015, p. 247):

> On this planet, at this moment in time, human beings reign. Regardless of who or what penned earlier chapters in the history of life, we will write the next one. Through our actions or inaction, we decide the world that our grandchildren and great grandchildren will know. Let us have the grace and humility to choose well.

Several scientists have argued that unwise actions by technological species may comprise one of the reasons underlying the *Great Silence,* i.e., the apparent absence of any signatures of technological intelligence despite decades of searching for them. There are many other solutions, discussed in Chapter 8, but one of the most sobering among them is the notion that the Great Filter lies ahead of us.[28] This ought to encourage us in the search for

---

28. The "Great Filter," coined by Robin Hanson (see Section 8.2.4), refers to one or more bottlenecks that prevent the expansion of technological intelligence in the Universe.

relics of dead technological species in the form of megastructures, space debris, and industrial pollution, to name a few. In the course of searching for such technosignatures, perhaps we shall come across the relics of dead species who became extinct because of their overweening hubris. A discovery along these lines might serve as a timely reminder of our responsibility to ourselves and other denizens of the planet and underscore the possibility that technological intelligence could be a transient phase in the Cosmos.

PART 2

ASPECTS OF EXTRATERRESTRIAL BIOSPHERES

Chapter 4

HABITABILITY: STELLAR FACTORS

The bright sun was extinguish'd, and the stars
Did wander darkling in the eternal space,
Rayless, and pathless, and the icy earth
Swung blind and blackening in the moonless air;
. . . . . .
The world was void,
The populous and the powerful was a lump,
Seasonless, herbless, treeless, manless, lifeless—
A lump of death-a chaos of hard clay.
. . . . . .
The winds were wither'd in the stagnant air,
And the clouds perish'd; Darkness had no need
Of aid from them—She was the Universe.

—Lord Byron (George Gordon), *Darkness*

A number of ancient civilizations speculated that the Universe was populated by worlds (planets and moons) beyond the Solar system. For millennia, this conjecture remained purely theoretical as no extrasolar planets (exoplanets) had been conclusively identified. The primary difficulty associated with finding extrasolar worlds stems from the fact that these objects are several orders of magnitude fainter than their host stars, owing to which their detection was rendered very challenging. However, with the advent of increasingly sophisticated telescopes, the late twentieth century witnessed the discovery of exoplanets. One of the most notable developments in this realm was the detection of 51 Pegasi b, the first exoplanet orbiting a solar-type star, by Michel Mayor and Didier Queloz in 1995 (Mayor & Queloz 1995). These two scientists were subsequently awarded the Nobel Prize in Physics in 2019 for this momentous breakthrough.[1] A meticulous account

1. *The Nobel Prize in Physics 2019,* Nobel Media AB 2020, Oct. 14, 2020, https://www.nobelprize.org/prizes/physics/2019/summary/

of the major landmarks in exoplanetary science can be found in Perryman (2018).

From 1989 to 2009, exoplanets were detected at a steady rate and numbered in the hundreds by the time NASA's *Kepler* spacecraft was launched (2009). The *Kepler* mission was designed with the chief purpose of detecting exoplanets, thereby facilitating the rapid advent of this field. In conjunction with the *Kepler* mission, a flotilla of ground-based telescopes have driven the discovery of some of the most famous exoplanets, which we shall encounter shortly hereafter. The total number of exoplanets detected as of 2021 is well over four thousand, and this number continues to increase.[2] Hence, it is no exaggeration to contend that exoplanetary science constitutes one of the most exciting frontiers in astrophysics. We will delve into some of the most common methodologies used for discovering exoplanets in Chapter 6; further details are furnished in Winn and Fabrycky (2015) and Perryman (2018).

It may be argued that one of the most significant reasons for studying exoplanets is to open up the possibility of detecting extraterrestrial life outside the Solar system. At this stage, it must be noted that extrasolar moons (exomoons) also represent plausible and compelling abodes for extraterrestrial life. Another advantage attributed to exomoons is that they are likely to be more numerous than exoplanets in our Galaxy. Nonetheless, we shall mostly restrict ourselves to exoplanets since they have been the subject of a larger number of theoretical and observational studies. The evidence for exomoons has been slim thus far; one of the more promising candidates in this respect was a potential Neptune-sized object orbiting the Jupiter-sized planet Kepler-1625b (Teachey & Kipping 2018), but the existence of this candidate has been thrown into question by subsequent studies.

Given the importance of water as a solvent for life-as-we-know-it,[3] most studies in astrobiology tend to adopt a "follow the water" approach. In turn, this has motivated the development of the concept of the circumstellar habitable zone (HZ)—namely, the annular region surrounding the host star in which a planet is theoretically capable of hosting liquid water on its surface. We will not discuss the HZ further at this juncture, because we shall address it in more detail later. In connection with the HZ, a couple

---

2. See http://exoplanet.eu/catalog/

3. In our discussion henceforth, whenever we make use of the word *life,* it must be understood to implicitly signify life-as-we-know-it, unless stated otherwise.

of noteworthy discoveries of exoplanetary systems over the past few years merit a special mention.

A rocky exoplanet was discovered by Anglada-Escudé et al. (2016) in the HZ of Proxima Centauri, the star nearest to the Earth at a distance of 1.3 pc. Proxima Centauri has a mass of $0.12\,M_\odot$, where $M_\odot$ denotes the mass of the Sun. It has a rotational timescale of 82.6 days, and the age of this star (~ 4.85 Gyr) is slightly higher than that of the Sun. This exoplanet, termed Proxima Centauri b (or Proxima b for short), has a minimum mass of $1.3\,M_\oplus$, where $M_\oplus$ is the mass of the Earth. The second major breakthrough involved the discovery of seven roughly Earth-sized planets orbiting the star TRAPPIST-1 at a distance of 12.1 pc in 2016–2017 (Gillon et al. 2017). The mass of the star is $0.09\,M_\odot$, while its rotation rate and age are 3.3 days and 7.6 ± 2.2 Gyr, respectively. Of the seven planets, it is believed that at least three of them reside within the HZ. The masses of these seven planets range between $\sim 0.3\,M_\oplus$ and $\sim 1.16\,M_\oplus$, and the radii fall within $\sim 0.77\,R_\oplus$ and $\sim 1.15\,R_\oplus$ (Grimm et al. 2018), where $R_\oplus$ is the radius of the Earth. Another discovery worth highlighting is the planet LHS 1140b, with radius and mass of $1.4\,R_\oplus$ and $6.6\,M_\oplus$, respectively, in the HZ of a $0.15\,M_\odot$ star (LHS 1140) at a distance of 12 pc (Dittmann et al. 2017). Lastly, the planet Ross 128b (with $M \geq 1.35\,M_\oplus$) orbits the nearby M-dwarf Ross 128 (with $M_\star \approx 0.17\,M_\odot$, where $M_\star$ stands for "mass of the star") at a distance of roughly 3.4 pc (Bonfils et al. 2018), but it appears to be situated within the inner edge of the HZ, thus making it susceptible to rapid water loss via a greenhouse effect.

As noted earlier, one of the primary reasons that motivates the study of exoplanets is their potential for hosting extraterrestrial life. In order to quantify the propensity of a given planet to host life, it is important to determine whether it is habitable. The word *habitable* has been often misused and overused, owing to which we shall adopt the following definition espoused in the 2015 version of the NASA Astrobiology Strategy:[4]

> Habitability has been defined as the potential of an environment (past or present) to support life of any kind. . . . Habitability is a function of a multitude of environmental parameters whose study is biased by the effects that biology has on these parameters.

4. Available at https://astrobiology.nasa.gov/nai/media/medialibrary/2016/04/NASA_Astrobiology_Strategy_2015_FINAL_041216.pdf

As per this definition, it is apparent that habitability applies to either planets or moons, because they are conventionally perceived as the archetypal abodes for life. Hence, one might be tempted to believe that only planetary factors—such as its surface gravity, atmospheric composition, and temperature, among others—govern habitability. It is, however, essential to recognize that planets and moons in stellar systems ought *not* be viewed in isolation. In other words, many of their properties are dictated, either directly or indirectly, by the host star(s) around which they orbit. The same principle applies to habitability, given that there exist a number of factors influenced by stellar parameters.

Hence, we will address the stellar aspects of habitability herein. This chapter expands on our review paper (Lingam & Loeb 2019d), and readers are invited to consult this work for a more comprehensive list of references. We shall concern ourselves with elucidating how myriad characteristics of putative extraterrestrial biospheres could be influenced by their host stars. Single main-sequence stars like the Sun comprise the bedrock of this chapter, although ~ 40 percent of the entire Milky Way population might possess one or more stellar companions. Second, we shall focus on the mass of the star ($M_\star$) as the chief variable; to put it differently, $M_\star$ is treated as a proxy for all other stellar parameters. In actuality, variables such as the age, magnetic fields, and rotation rate will play a major role in regulating habitability, but determining these parameters to a high degree of accuracy is challenging; the mass has the additional benefit of being a physically transparent variable.

Lastly, much of our attention will be devoted to exoplanets orbiting M-dwarfs. Broadly speaking, M-dwarfs are stars whose masses fall typically in the range of $0.075 < M_\star/M_\odot < 0.6$. One of the most important features of M-dwarfs is that they can be classified into two distinct "flavors" (Chabrier & Baraffe 2000). M-dwarfs that obey $M_\star \gtrsim 0.35\, M_\odot$ are distinguished by stellar interiors composed of an inner radiative zone and an outer convective envelope. In the radiation zone, energy is transmitted as electromagnetic radiation via photon diffusion. On the other hand, in the convection zone, the transport of energy is via the convection (i.e., sinking and rising) of plasma. In contrast, when $M_\star \lesssim 0.35\, M_\odot$, these M-dwarfs become *fully* convective—that is, they lack the radiative zone altogether.

Apart from this distinction, M-dwarfs also manifest significant variation in stellar radii, effective temperatures, surface magnetic fields, and flaring activity. Not only are M-dwarfs the most common (> 70 percent of all stars in the Solar neighborhood belong to this category; T. J. Henry

et al. 2006, 2018) and long-lived in the Universe, about 20 percent of them are known to possess roughly Earth-sized planets in their HZs (Dressing & Charbonneau 2015). These planets are also comparatively accessible to detailed observations, for reasons that shall be explained in Chapter 6. Finally, the nearest temperate planets discovered as of 2019–2020—Proxima b, the TRAPPIST-1 system, and LHS 1140b, to name a few—are located around M-dwarfs. The reader may consult Tarter et al. (2007), Scalo et al. (2007), and Shields et al. (2016) for in-depth overviews of the habitability of M-dwarf exoplanets.

4.1 THE HABITABLE ZONE AND ITS EXTENSIONS

As remarked previously, the HZ is the region around the host star(s) where standing bodies of liquid water can exist on the surface. The HZ has a long and interesting history, dating back to at least the nineteenth century, and includes seminal contributions from the likes of Edward Maunder, Hubertus Strughold, Harlow Shapley, Su-Shu Huang, and Stephen H. Dole in the early- and mid-twentieth century. Historical accounts of the development of the HZ concept are expounded in Gonzalez (2005) and Lorenz (2019). The reader may consult Ramirez (2018) for a comprehensive technical review of habitable zones.

The modern formulation of the HZ was first worked out in the seminal paper by Kasting et al. (1993), wherein it was assumed that the exoplanets under consideration comprise atmospheres akin to that of the Earth, with the primary greenhouse gases being carbon dioxide (CO_2) and water vapor (H_2O). The inner edge of the HZ was computed by determining the point at which the rapid loss of water occurred by means of photolysis (i.e., splitting molecules via stellar irradiation) as a result of enhanced radiation fluxes from the star. As one moves toward the outer edge of the HZ, subsequent studies have shown that the onset of cooling occurs via two factors: (1) an increase in the albedo due to high atmospheric CO_2 levels and (2) condensation of CO_2 initiated at sufficiently high atmospheric CO_2 pressures, thus mitigating the greenhouse effect.

The power input (P_{in}) to the planet from the host star is given by

$$P_{\text{in}} = \left(\frac{L_\star}{4\pi a^2} \right) \times \left(\pi R^2 \right) \times \left(1 - A_p \right), \tag{4.1}$$

where the first factor on the right-hand side represents the stellar flux, the second factor quantifies the area intercepted by the planet, and the third factor denotes the amount of radiation absorbed by the planet. $L_\star$ denotes the stellar luminosity, a is the star-planet distance (which equals the semimajor axis for a circular orbit), and R and A_p are the radius and albedo of the planet. The stellar luminosity is further expressible as

$$L_\star = \left(4\pi R_\star^2\right) \times \left(\sigma T_\star^4\right) \tag{4.2}$$

after using the Stefan-Boltzmann law; σ denotes the Stefan-Boltzmann constant. Here, we have assumed that the star is a black body, with $R_\star$ and $T_\star$ denoting the radius and effective temperature of the star. Similarly, if we treat the planet as a black body with temperature $T_{\rm eq}$, the emitted power ($P_{\rm out}$) is

$$P_{\rm out} = \left(4\pi R^2\right) \times \left(\sigma T_{\rm eq}^4\right). \tag{4.3}$$

By demanding that $P_{\rm in} = P_{\rm out}$, we obtain

$$\sigma T_{\rm eq}^4 = \frac{L_\star \left(1 - A_p\right)}{16\pi a^2}. \tag{4.4}$$

From this expression, it can be seen that a planet with the same value of $T_{\rm eq}$ and A_p as that of the Earth is characterized by an orbital radius $a_\star$, given by

$$a_\star \equiv 1\ {\rm AU} \left(\frac{L_\star}{L_\odot}\right)^{1/2} = 1\ {\rm AU} \left(\frac{M_\star}{M_\odot}\right)^{3/2}, \tag{4.5}$$

where $1\ {\rm AU} \approx 1.5 \times 10^{11}$ m, and the last equality follows if one invokes the mass-luminosity relationship $L_\star \propto M_\star^3$, which is partially justified on both empirical and theoretical grounds (Böhm-Vitense 1992). Henceforth, we will refer to planets with basic parameters (e.g., size, effective temperature, atmospheric composition, and pressure) and properties—especially in the geological, chemical, and physical domains—similar to the Earth as *Earth-analogs*. In doing so, we emphasize that these worlds are not true analogs since they are not guaranteed to be inhabited and sustain Earthlike biospheres.

From the above formulae, we clearly see that low-mass stars emit less radiation, implying that their HZs are situated closer with respect to the

Sun. A second point is that stellar luminosity increases over time. If we hold all other factors fixed for the sake of simplicity, it can be seen from (4.5) that the HZ will move outward over time. This qualitative feature has been confirmed and quantified by several numerical studies that employed detailed climate models. Rushby et al. (2013) developed a detailed model to estimate the HZ lifetime of Earth-analogs (t_{HZ}) as a function of the stellar mass. Naively, we may expect that the HZ lifetime would be a constant fraction of the overall stellar lifetime, but the corresponding scaling relations are more complicated. By adopting the prescription from Lingam and Loeb (2019f), we have

$$\begin{aligned} t_{HZ} &\sim 0.55\, t_{\odot} \left(\frac{M_{\star}}{M_{\odot}}\right)^{-2} \qquad M_{\star} > M_{\odot}, \\ t_{HZ} &\sim 0.55\, t_{\odot} \left(\frac{M_{\star}}{M_{\odot}}\right)^{-1} \qquad 0.5M_{\odot} < M_{\star} < M_{\odot}, \\ t_{HZ} &\sim 0.46\, t_{\odot} \left(\frac{M_{\star}}{M_{\odot}}\right)^{-1.25} \qquad M_{\star} < 0.5M_{\odot}, \end{aligned} \tag{4.6}$$

with $t_{\odot} \approx 10^{10}$ yr representing the main-sequence lifetime of the Sun.

The inner and outer edges of the HZ for atmospheres with Earthlike greenhouse gases vary from study to study, but the most well-known limits in the 2010s were worked out by Kopparapu et al. (2013). The effective stellar flux (S_{eff}) is defined as the ratio of the stellar flux S incident on the planet to the solar flux received at the Earth ($S_0 \sim 1360$ W/m^2). The corresponding distance d can be computed once S_{eff} is known via the following formula:

$$a = 1\ \text{AU} \sqrt{\frac{L_{\star}/L_{\odot}}{S_{\text{eff}}}} \tag{4.7}$$

The variable S_{eff} is calibrated in terms of $T_{\star}$ (effective stellar temperature) after carrying out numerical simulations and is expressible as

$$S_{\text{eff}} = S_{\text{eff},\odot} + \mathcal{A}\tilde{T}_{\star} + \mathcal{B}\tilde{T}_{\star}^2 + \mathcal{C}\tilde{T}_{\star}^3 + \mathcal{D}\tilde{T}_{\star}^4, \tag{4.8}$$

with $\tilde{T}_{\star} = T_{\star} - 5780$ K. For the inner edge of the HZ, the parameters are given by $S_{\text{eff},\odot} = 1.0140$, $\mathcal{A} = 8.1774 \times 10^{-5}$, $\mathcal{B} = 1.7063 \times 10^{-9}$, $\mathcal{C} = -4.3241 \times 10^{-12}$, and $\mathcal{D} = -6.6462 \times 10^{-16}$. In order to calculate

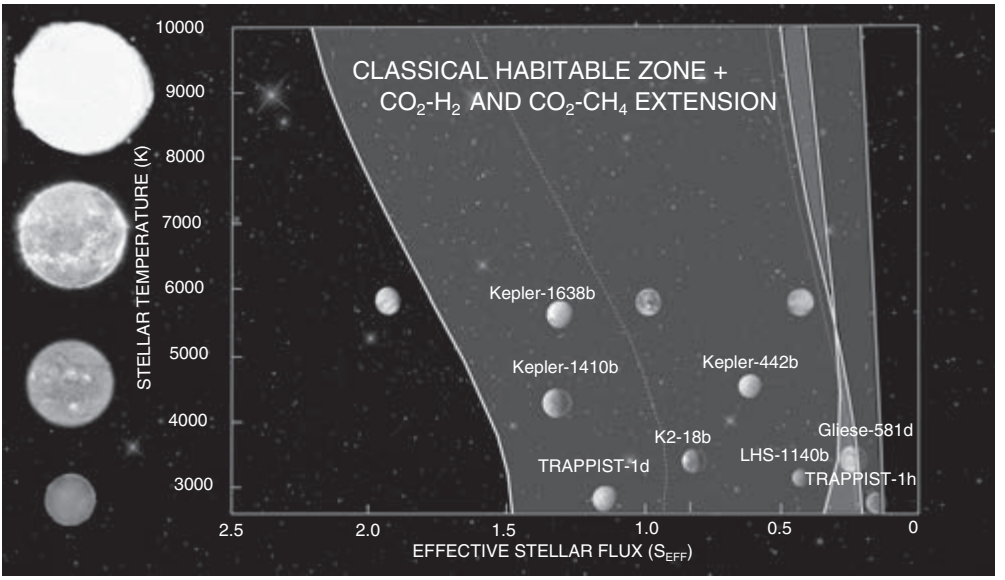


Figure 4.1 The solid lines demarcate the limits of the optimistic HZ derived from empirical considerations, while the dashed lines constitute the conservative limits of the HZ; the latter range is typically invoked more often than the former. The aforementioned curves correspond to the classical HZ computed using CO_2 and H_2O as the greenhouses gases. The light-shaded region at the right end represents the extension of the HZ for CO_2-CH_4 atmospheres. The dark-shaded region at the extreme right end delineates the HZ limits for CO_2-H_2 atmospheres. The *y*-axis depicts the effective temperature of the star, whereas the *x*-axis is the stellar flux incident on the planet normalized by the solar flux at Earth. (CC-BY. *Source:* Ramses M. Ramirez [2018], A more comprehensive habitable zone for finding life on other planets, *Geosciences* 8[8]: 280, fig. 8.)

the outer edge of the HZ, the constants to be used are $S_{\text{eff},\odot} = 0.3438$, $\mathcal{A} = 5.8942 \times 10^{-5}$, $\mathcal{B} = 1.6558 \times 10^{-9}$, $\mathcal{C} = -3.0045 \times 10^{-12}$, and $\mathcal{D} = -5.2983 \times 10^{-16}$.

Hitherto, we have only concerned ourselves with the possibility of CO_2 and H_2O serving as the greenhouse gases. In actuality, there are a number of other candidates, including molecular hydrogen (H_2), methane (CH_4), and nitrous oxide (N_2O). If these gases are included in the atmospheric composition, the outer edge of the HZ is extended further outward in some cases, as seen in Figure 4.1. One of the key points discernible from this plot is that TRAPPIST-1h, the outermost planet of the TRAPPIST-1 system discovered to date, falls outside the conventional HZ. On the other hand,

this planet lies within the domain of the modified HZ in the event that H_2 contributes non-negligible greenhouse warming.

Thus, the inner and outer limits (and therefore the width) of the HZ evidently depend on a number of factors beyond just the stellar temperature or, equivalently, its mass. To start with, as we witnessed in the preceding paragraph, the composition of the greenhouse gases in the planetary atmosphere clearly plays a vital role. In addition, a number of other factors must be taken into consideration in drawing the boundaries of the HZ, such as the planet's mass, rotation rate, and the fraction of its surface occupied by land and water. Moreover, the dynamical evolution of the HZ is also complex, as its variation is dependent on the spectral type of the star in concurrence with several planetary characteristics. Hence, we shall not address this multifaceted issue further; the interested reader should consult Ramirez (2018) for the salient details.

In closing, a couple of important points must be appreciated regarding the HZ, as this term has been loosely used by the media and, at times, in the scientific literature. First off, not every planet in the HZ is guaranteed to possess standing bodies of liquid water on its surface. Next, planets and moons outside the HZ are capable of sustaining liquid water (and perhaps even life) by virtue of subsurface oceans beneath the surface; this point is one that we shall explore in Chapter 7. Furthermore, the existence of liquid water serves as a necessary, but not sufficient, condition for the existence of life. As noted in Section 2.7, other necessary requirements include sufficient abundances of bioessential elements (e.g., nitrogen and phosphorus) and free energy flows. For these reasons, it is crucial to avoid conflating the intricate notions of the HZ and habitability.

On account of the limitations associated with the HZ, in conjunction with the tendency to interpret it as instantiating a different concept (habitability), some scientists have advocated the utilization of alternative terminology (Moore et al. 2017; Tasker et al. 2017) or even dispensing with this phrase altogether. Yet, it is equally important to appreciate the benefits accruing from deploying the HZ paradigm. For starters, a diverse array of planets and stars are encompassed within the modern formulations of the HZ as a result of the new developments in this field. Consequently, these formulations of the HZ include planets that are not Earthlike and open up the possibility of habitable worlds orbiting pre- and post-main-sequence stars or stellar binaries. Of more importance, perhaps, is the fact that we are limited by finite observational resources, because of which the selection of

the optimal target stars and planets becomes very important. In this respect, the HZ offers a plausible filter for identifying and selecting high-priority exoplanets that merit in-depth observations.

Before bringing our analysis to a close, it must be recognized that our focus has been on identifying the HZs for single, main-sequence stars. As remarked previously, the HZ has been applied to a wide range of stellar systems. Moreover, several formulations of the HZ have been developed for exomoons—to wit, the so-called circumplanetary habitable zone—that take stellar illumination, tidal heating, and dynamics arising from gravitational interactions with their host planets into account (Heller et al. 2014). Lastly, on a broader scale, several authors have attempted to delineate the limits of the Galactic Habitable Zone (GHZ)—that is, the region within our Galaxy (and others) where the development of complex life is favored (Lineweaver et al. 2004; Gonzalez 2005; Prantzos 2008). While these concepts possess intrinsic value in their own right, we shall not tackle them as they are presumably subject to greater systemic uncertainties, especially the GHZ.

4.2 STELLAR WINDS

In qualitative terms, stellar winds refer to streams of energetic, charged particles (plasma) that propagate from the outer regions of stellar coronae into interplanetary space. There are several avenues by which stellar winds regulate planetary habitability, as we shall describe in more detail below.

4.2.1 Planetary magnetospheres

It is a well-known fact that, in the presence of magnetic fields, charged particles are deflected. Hence, when a given planet has an intrinsic magnetic field, the stellar wind plasma in its vicinity will be deflected. The "cavity" that is formed as a result of this deflection by the planetary magnetic field constitutes the planetary magnetosphere. A succinct overview of magnetospheres can be found in Gombosi (1998). In this context, estimating the magnetopause distance (R_{mp}), which represents the outer boundary of the magnetosphere, is useful since it serves as a proxy of sorts for the size of the magnetosphere.

Let us begin by supposing that the planet has a pure dipole magnetic field. If the planet has a magnetic moment of magnitude $\mathcal{M}$ and aligned along the z-axis, the magnetic field at distance r is given by

$$\mathbf{B} = \frac{\mu_0}{4\pi}\frac{\mathcal{M}}{r^3}\left(2\cos\theta\,\hat{r} + \sin\theta\,\hat{\theta}\right). \tag{4.9}$$

We are primarily interested in the radial dependence on r, owing to which we shall ignore the term inside the brackets on the right-hand side; this is equivalent to choosing $\theta = \pi/2$. Let us denote the magnitude of the magnetic field at the surface of the planet by B_s. At the magnetopause distance, the corresponding magnetic field strength B_{mp} is given by

$$B_{mp} = B_s\left(\frac{R}{R_{mp}}\right)^3, \tag{4.10}$$

where R represents the radius of the planet; it is straightforward to verify that $\mathcal{M} \approx B_s R^3$. The magnetopause distance is calculated by demanding that the magnetic pressure arising from the planetary magnetic field approximately equal the total stellar wind pressure (P_{sw}). The total stellar wind pressure may be further expressed as the sum of the dynamic (P_{dyn}), magnetic (P_{mag}), and thermal (P_{th}) pressures of the stellar wind. Of this trio, the first term is typically the most important, followed by the second term. They can be expressed as

$$P_{dyn} = \rho_{sw} v_{sw}^2 \tag{4.11}$$

and

$$P_{mag} = \frac{B_{sw}^2}{2\mu_0}, \tag{4.12}$$

with ρ_{sw}, B_{sw}, and v_{sw} denoting the mass density, magnetic field strength, and velocity of the stellar wind, respectively; here, μ_0 is the vacuum permeability.

As per the preceding discussion, we can use the condition $B_{mp}^2/(2\mu_0) = P_{sw}$ to solve for R_{mp}, which leads us to

$$R_{mp} = R\left(\frac{B_s^2}{2\mu_0 P_{sw}}\right)^{1/6}. \tag{4.13}$$

A more accurate estimate of R_{mp} necessitates the inclusion of an extra factor that is very close to unity, because the assumption of an ideal dipole planetary magnetic field is not realistic. We note that the characteristic value of the magnetopause distance for present-day Earth is $R_{mp} \approx 10\, R_{\oplus}$. By inspecting the above formula, we observe that $R_{mp} \propto P_{sw}^{-1/6}$. As per this relation, it is apparent that a higher wind pressure would automatically yield a smaller magnetosphere, ceteris paribus.

To understand how P_{sw} varies as a function of stellar properties, it is instructive to contemplate a couple of toy models, with the proviso understanding that these analytical approaches do not capture the full complexity of the issue. To begin with, let us assume that $P_{sw} \approx P_{dyn}$. Since the star is losing mass via the stellar wind, we shall treat the latter as being spherically symmetric. The stellar mass-loss rate $\dot{M}_{\star}$ is thus expressible as

$$\dot{M}_{\star} = 4\pi a^2 \rho_{sw} v_{sw}, \tag{4.14}$$

where a is the orbital radius introduced earlier. For the stellar mass-loss rate, we shall adopt the prescription provided in Johnstone et al. (2015), based on comprehensive numerical simulations. A prominent caveat is that this scaling relation is more accurate for $0.4 < M_{\star}/M_{\odot} < 1.1$. The stellar mass-loss rate is

$$\frac{\dot{M}_{\star}}{\dot{M}_{\odot}} \approx \left(\frac{R_{\star}}{R_{\odot}}\right)^2 \left(\frac{\Omega_{\star}}{\Omega_{\odot}}\right)^{1.33} \left(\frac{M_{\star}}{M_{\odot}}\right)^{-3.36}, \tag{4.15}$$

where $\Omega_{\star}$ represents the rotation rate of the star and $\dot{M}_{\odot} \sim 2 \times 10^{-14}\, M_{\odot}\, \mathrm{yr}^{-1}$ is the present-day solar mass-loss rate. This can be further simplified by using the mass-radius relationship $R_{\star} \propto M_{\star}^{0.8}$ (Johnstone et al. 2015) to yield

$$\frac{\dot{M}_{\star}}{\dot{M}_{\odot}} \approx \left(\frac{\Omega_{\star}}{\Omega_{\odot}}\right)^{1.33} \left(\frac{M_{\star}}{M_{\odot}}\right)^{-1.76}. \tag{4.16}$$

This expression is valid provided that the star is *not* a rapid rotator—to wit, when its rotation rate is below the saturation value:

$$\Omega_{\mathrm{sat}} = 15\Omega_{\odot}\, (M_{\star}/M_{\odot})^{2.3} \tag{4.17}$$

Note that $\Omega_{\odot}$ is the rotation rate of the Sun, which has an approximate period of twenty-four days at the equator, although the exact magnitude

depends on the latitude. In the event that the star is a rapid rotator, its mass-loss rate must be replaced with

$$\frac{\dot{M}_\star}{\dot{M}_\odot} \approx 37 \left(\frac{M_\star}{M_\odot}\right)^{1.3}. \tag{4.18}$$

As a consistency check, let us consider Proxima Centauri despite the fact that it does not fall within the optimal mass range for this model. We make use of $M_\star/M_\odot \approx 0.12$, $R_\star/R_\odot \approx 0.15$, and $\Omega_\star/\Omega_\odot = 0.29$ in (4.15). The substitution yields $\dot{M}_\star \approx 5\,\dot{M}_\odot$, which is within an order of magnitude of the estimate $\dot{M}_\star \approx 0.8\dot{M}_\odot$ obtained via numerical modeling (Garraffo et al. 2016) or the observational constraint $\dot{M}_\star < 14\dot{M}_\odot$ derived from X-ray measurements (Wargelin & Drake 2002).

Before assembling our relations together, it should be appreciated that v_{sw} is typically not expected to exhibit significant variations beyond the immediate vicinity of the star, based on theoretical profiles of the stellar wind. Hence, we may treat it as being roughly constant for the purpose of our analysis. Upon substituting (4.5) and (4.16) into (4.14) and (4.11), we obtain

$$P_{dyn} \propto \Omega_\star^{1.33} M_\star^{-4.76} \tag{4.19}$$

for Earth-analogs. Thus, as per this expression, it is evident that Earth-analogs around low-mass stars experience much higher stellar wind pressures. Let us make use of the above scaling for an Earth-analog orbiting Proxima Centauri. From the expressions in the previous paragraph, we end up with a value of P_{sw} that is $\sim 4.3 \times 10^3$ higher than the solar wind pressure at Earth. This estimate is surprisingly close to the factor of $\sim 2 \times 10^3$ obtained from detailed simulations by Garraffo et al. (2016). In reality, during the orbit of Proxima b, P_{sw} changes by a couple of orders of magnitude, thus modifying R_{mp} by a factor of 2–5.

Next, let us consider the scenario where only the magnetic pressure component of the stellar wind, namely (4.12), is considered. To a reasonable degree, one may suppose that magnetic flux Φ_B is considered. Since the latter is given by $\Phi_B \sim BR^2 = \text{const}$, we have

$$B_{sw} \sim B_\star \left(\frac{R_\star}{a}\right)^2, \tag{4.20}$$

where $B_\star$ is the magnetic field at the surface of the star. This expression is not accurate for distances very close to the host star but is reasonably valid otherwise (Vidotto et al. 2013). The last piece of our puzzle, the stellar magnetic field, is supplied from the dynamo scaling law of Reiners and Christensen (2010) derived from simulations:

$$B_\star = 0.48\,\mathrm{T}\left(\frac{M_\star}{M_\odot}\right)^{1/6}\left(\frac{L_\star}{L_\odot}\right)^{1/3}\left(\frac{R_\star}{R_\odot}\right)^{-7/6}. \tag{4.21}$$

As this scaling law applies only to brown dwarfs and M-dwarfs, its validity is compromised when $M_\star > 0.5\,M_\odot$. In the case of Proxima Centauri, employing $M_\star/M_\odot \approx 0.12$, $R_\star/R_\odot \approx 0.15$, and $L_\star/L_\odot \approx 0.0017$ yields $B_\star \sim 0.36$ T, which is roughly six times higher than the actual value deduced via observations. By using the approximate mass-luminosity and mass-radius relationships, the above expression simplifies to

$$B_\star = 0.48\,\mathrm{T}\left(\frac{M_\star}{M_\odot}\right)^{0.23}. \tag{4.22}$$

Finally, we substitute (4.5), (4.20), and (4.22) into (4.12), thus obtaining

$$P_{mag} \propto M_\star^{-7/3} \tag{4.23}$$

for Earth-analogs. Now, let us suppose that we wish the magnetopause distance to be comparable to that of the Earth. From (4.13), it can be verified that this condition amounts to requiring that B_s^2/P_{mag} is roughly constant since we have assumed $P_{sw} \sim P_{mag}$. Thus, by invoking (4.23), we see that $B_s \propto M_\star^{-7/6}$ must be valid for an Earth-sized magnetosphere to exist. In turn, this implies the planetary magnetic field ought to be at least an order of magnitude higher than that of the Earth (whose magnetic field is $\sim 10^{-4}$ T) if the planet orbits a star of mass $\sim 0.1 M_\odot$.

Let us turn our attention to (4.13) once again. Hitherto, we have seen that P_{sw} is likely to be orders of magnitude higher than the solar wind pressure at Earth for temperate planets around low-mass stars. One could rightly argue that a sufficiently high value of B_s will compensate for this increase in P_{sw}, thus ensuring that sizable magnetospheres are still feasible. However, dynamo theory suggests that planets orbiting low-mass stars may have weak magnetic fields. A number of dynamo scaling laws have been derived, and

a summary of these relations can be found in Christensen (2010). Classical models imply that the planetary magnetic field B_s obeys $B_s \propto \Omega_c^{\alpha}$, where Ω_c is the rotation rate of the planetary core and $\alpha \sim 0.5$ in several instances; however, more recent theories indicate that $\alpha = 0$ might also occur. For example, a simple dynamo scaling law proposed by D. J. Stevenson (1979) originated from the premise that the Elsasser number in the core was on the order of unity. The Elsasser number Λ is conventionally interpreted as the ratio of the Lorentz force and the Coriolis force and is defined as

$$\Lambda = \frac{\sigma_c B^2}{2\rho\Omega}, \tag{4.24}$$

where σ_c, ρ, and Ω denote the conductivity, density, and rotation rate, respectively. Thus, by imposing the condition $\Lambda \sim 1$ in the core, we see that $B_s \propto \Omega_c^{1/2}$. Now, let us suppose that the planet rotates akin to a rigid body, which implies that Ω_c can be replaced by the planet's rotation rate. Theoretical and numerical studies both suggest that temperate planets around low-mass stars are possibly synchronous rotators,[5] with their rotation periods equal to their orbital periods, due to the tidal gravitational force exerted by the host star via tidal locking (see Section 5.3). In this event, the rotation rate may be reduced considerably, thereby leading to a corresponding decrease of the planetary magnetic field. For example, consider the planet TRAPPIST-1g, which lies in the HZ of TRAPPIST-1. It has a mass and radius nearly equal to the Earth and an orbital period of ~ 12.4 days. Under the assumption of synchronous rotation and holding all other factors fixed, we end up with $B_s \sim 0.3 B_{\oplus}$, where $B_{\oplus} \sim 10^{-4}$ T is the Earth's magnetic field strength.

Hence, to summarize our analysis to date, Earth-analogs around low-mass stars are typically (but not always) expected to have smaller magnetospheres due to the combination of higher stellar wind pressures and potentially weaker magnetic moments relative to Earth. Until recently, it was assumed that weaker planetary magnetic fields and smaller magnetospheres conferred reduced protection against the stellar wind, thereby amplifying atmospheric escape rates; we will tackle this subject in Section 4.2.2. For

5. Not *all* planets around low-mass stars are guaranteed to exist in a state of synchronous rotation. The mechanisms that can result in asynchronous rotation include the likes of orbital dynamics, sufficiently massive atmospheres, triaxial deformation, and semiliquid interiors.

instance, numerical simulations carried out for Proxima b, based on the presumption of CO_2-dominated atmospheres, concluded that the average atmospheric escape rate in the unmagnetized case ($\mathcal{M} \approx 0$) was higher than the magnetized scenario with $\mathcal{M} \approx 0.3\,\mathcal{M}_\oplus$ (where $\mathcal{M}_\oplus$ is the Earth's magnetic dipole moment) by a factor of approximately 2 (Dong, Lingam, et al. 2017). A similar result, involving a decrease of the escape rate by ~ 10 after switching on the magnetic field, was obtained from numerical simulations of the temperate exoplanet TOI-700 d discovered by *TESS* (Transiting Exoplanet Survey Satellite) after specifying an Earthlike atmosphere (Dong et al. 2020). Yet, as we shall sketch in Section 4.2.2, a higher magnetic field does not automatically translate to lower atmospheric escape rates in all regimes.

Another significant effect worth highlighting at this stage concerns the role of planetary magnetic fields in regulating the flux of cosmic rays that reach the Earth. Cosmic rays are composed of high-energy particles and contribute to the radiolysis (decomposition of chemical species via radiation) of complex biomolecules. In addition, cosmic rays also react with N_2-O_2 atmospheres and stimulate the formation of nitrogen oxides (NO_x) and drive the depletion of ozone (O_3). Recent numerical studies indicate that the planet's magnetic moment and atmospheric column density collectively influence the amount of cosmic rays that reach the surface, since the magnetosphere and the atmosphere provide two layers of shielding. A summary of how these quantities influence the radiation dose rate at the surface is provided in Table 4.1. It can be seen that modifying the magnetic moment from $\mathcal{M} = 0$ to $\mathcal{M} = 10\mathcal{M}_\oplus$ for Earthlike atmospheres leads to a decrease in the radiation dose rate by a factor of about 6. In contrast, if we lower the atmospheric column density to ~ 10 percent of the Earth's value, the radiation dose rate declines by a factor of 240 as one transitions from $\mathcal{M} = 0$ to $\mathcal{M} = 10\mathcal{M}_\oplus$. Hence, it is apparent that the sensitivity of the biological dose rate to the magnetic moment of the planet also requires us to take the atmospheric column density into consideration.

4.2.2 Atmospheric escape

Broadly speaking, atmospheric escape encompasses a diverse array of mechanisms that are capable of imparting sufficient energy to particles, consequently enabling them to attain speeds higher than the escape velocity, i.e., the threshold to escape the gravitational pull of the planet. The reader may

Table 4.1 Total biological dose rate (DR) as a function of the atmospheric column density and magnetic moment

Magnetic moment (in $\mathcal{M}_\oplus$)	$\mathrm{DR}^{(4)}$ (in Sv/yr)	$\mathrm{DR}^{(3)}$ (in Sv/yr)
0.0	6.5×10^{-4}	0.553
0.1	4.8×10^{-4}	0.527
0.15	4.6×10^{-4}	0.510
0.25	4.4×10^{-4}	0.405
0.5	4.2×10^{-4}	0.257
0.75	3.9×10^{-4}	0.216
1.0	3.4×10^{-4}	0.172
2.0	2.8×10^{-4}	0.053
3.0	2.3×10^{-4}	0.015
6.0	1.9×10^{-4}	5.7×10^{-3}
10.0	10^{-4}	2.3×10^{-3}

Notes: $\mathrm{DR}^{(4)}$ and $\mathrm{DR}^{(3)}$ denote the total biological dose rates for atmospheric column densities of 1.036×10^4 kg/m^2 (Earthlike) and 10^3 kg/m^2, respectively. The planet is taken to be analogous to Earth in terms of size and mass, while the corresponding stellar parameters are $M_\star = 0.45 M_\odot$ and $a = 0.153$ AU.

The sievert (Sv) is the SI unit for dose equivalent, and quantifies the equivalent biological effect arising from the deposition of 1 J of energy in 1 kg of human tissue. The average biological dose rate for humans on Earth due to natural background radiation is $\sim 2.4 \times 10^{-3}$ Sv/yr, while the corresponding value in the International Space Station is ~ 0.16 Sv/yr. This table is based on the data provided in Grießmeier et al. (2016). (*Data source:* J.-M. Griessmeier, F. Tabataba-Vakili, A. Stadelmann, J. L. Grenfell, and D. Atri, Galactic cosmic rays on extrasolar Earth-like planets. II. Atmospheric implications [2016], *Astron. Astrophysics* 587: A159.)

consult Lammer (2013) and Gronoff et al. (2020) for in-depth expositions of the various atmospheric escape mechanisms. Most studies in this area have tended to focus on the escape of neutral particles, but it is vital to recognize that ions also escape the planet via multifarious channels. The simplest mechanism that comes to mind is collisions. Particles that derive excess energy from collisions may acquire a higher velocity than the escape velocity and therefore end up being lost to space.

One of the most intuitive mechanisms for atmospheric loss is known as Jeans escape, which operates along the following lines. In the presence of atmospheric collisions, the Maxwell-Boltzmann distribution offers a fairly accurate description of the distribution function $\mathfrak{F}_{\mathrm{MB}}(v)$,

$$\mathfrak{F}_{\mathrm{MB}}(v) = \left(\frac{m_X}{2\pi k_B T_a} \right)^{3/2} \exp\left(-\frac{m_X v^2}{2 k_B T_a} \right), \tag{4.25}$$

where m_X is the particle mass of the chemical species X undergoing escape and T_a represents the temperature of the atmosphere. Next, the mass flux (units of kg m^{-2} s^{-1}) of the escaping particles ($\mathcal{F}_J$) is expressible as:

$$\mathcal{F}_J \approx \frac{1}{4} m_X \, n_a \, \langle v \rangle_e, \tag{4.26}$$

where n_a denotes the number density of species X in the atmosphere. The factor of 1/4 has been introduced in (4.26) because of geometric considerations similar to those encountered in Section 4.1. In the above equation, $\langle v \rangle_e$ refers to the average speed of the particles that are capable of escaping the planet's atmosphere. We can estimate $\langle v \rangle_e$ from

$$\langle v \rangle_e = \int_{v_{\rm esc}}^{\infty} v \, \mathfrak{F}_{\rm MB}(v) \, d^3 v, \tag{4.27}$$

with the planet's escape velocity ($v_{\rm esc}$) defined as

$$v_{\rm esc} = \sqrt{\frac{2GM}{R}} = 11 \, {\rm km/s} \sqrt{\frac{M}{M_\oplus}} \left(\frac{R}{R_\oplus} \right)^{-1/2} = 11 \, {\rm km/s} \left(\frac{R}{R_\oplus} \right)^{1.35}, \tag{4.28}$$

where the last equality is obtained by invoking the mass-radius relationship $M/M_\oplus \approx (R/R_\oplus)^{3.7}$ for rocky planets (Zeng et al. 2016). Hence, it is left as an exercise for the reader to verify that (4.27) transforms into

$$\langle v \rangle_e = 2 \sqrt{\frac{2 k_B T_a}{\pi \, m_X}} \left(1 + \frac{m_X v_{\rm esc}^2}{k_B T_a} \right) \exp \left(- \frac{m_X v_{\rm esc}^2}{k_B T_a} \right). \tag{4.29}$$

By combining (4.29) with (4.26), the Jeans escape flux can be calculated. In general, both n_a and T_a are functions of the atmospheric height ($v_{\rm esc}$ also exhibits a weak dependence in reality). The particles are conventionally treated as being capable of escaping the atmosphere once they cross the exobase, the boundary beyond which collisions are negligible. The exobase temperature and density vary from planet to planet; in the case of Earth, $T_a \sim 800$–1250 K and $n_a \sim 10^{11}$–10^{12} m^{-3} for hydrogen. By defining $\gamma_{\rm esc} = m_X v_{\rm esc}^2 / (k_B T_a)$ and plotting (4.26) as a function of $\gamma_{\rm esc}$, it is easy to verify that this function declines sharply with $\gamma_{\rm esc}$. Hence, provided that all other factors are held equal, the loss of lighter species (e.g., hydrogen) is highly

favored by Jeans escape. Observations have revealed that ∼ 30 percent of the total hydrogen lost from Earth is via Jeans escape. In contrast, this mechanism is negligible insofar as the escape of heavier species such as nitrogen and oxygen is concerned.

When dealing with charged particles, it must be borne in mind that they can also be accelerated by electric fields. The electric field **E** for a plasma is given by

$$\mathbf{E} = -\mathbf{V} \times \mathbf{B} + \eta \mathbf{J} + \frac{\mathbf{J} \times \mathbf{B} - \nabla p_e}{n_e e} + \ldots, \tag{4.30}$$

where **V** denotes the plasma bulk velocity, **J** represents the current, **B** is the magnetic field, η denotes the electrical resistivity, n_e refers to the electron number density, and p_e is the electron pressure. The above expression, known as Ohm's Law, is obtained by taking the difference of the ion and electron momentum equations expressed in terms of fluid variables (akin to the Euler equation from hydrodynamics). The complete version of Ohm's Law comprises many additional terms that have been omitted here (Lingam et al. 2017). The first and second terms on the right-hand side are the most widely encountered, since they are manifested in the simplest plasma fluid model known as resistive magnetohydrodynamics (MHD). The last two terms on the right-hand side are often collectively referred to as the Hall effect. All of the aforementioned terms contribute to the acceleration of charged particles, thus enabling them to exit the planetary atmosphere (Brain et al. 2016).

For example, the first term on the right-hand side, sometimes called the motional electric field ($\mathbf{V} \times \mathbf{B}$) drives atmospheric escape via the *ion pickup* mechanism. Charged particles subject to this component of the electric field experience gyration because of the magnetic field in the stellar wind plasma and are carried away from the planet, or re-impact the atmosphere and contribute to sputtering (another escape process). Hall electric fields ($\mathbf{J} \times \mathbf{B}$) have been documented to facilitate the escape of charged particles in regions with curved magnetic fields and contribute to the loss of plasma blobs via shear effects. Electron pressure gradients (∇p_e) constitute an essential ingredient of ambipolar electric fields, which arise from the fact that electrons typically move at faster velocities than ions, thus creating charge separation and driving the acceleration of ions to reimpose charge neutrality. These ambipolar electric fields are responsible for the outflow of ions from regions such as polar caps (and cusps), and this mechanism is

known as the *polar wind*. Polar caps are regions generally close to the magnetic poles that are characterized by the existence of open magnetic field lines therein.

Apart from processes such as the polar wind and ion pickup, there are several other avenues by which charged particles escape from planetary atmospheres. One such avenue is plasma instabilities, of which the Kelvin-Helmholtz instability (KHI) has been investigated in some detail. The KHI arises as a direct consequence of shear flow in fluids and drives the formation of vortices. These plasma vortices have been noticed to detach charged particles from the upper atmosphere, thereby driving atmospheric escape in the form of plasma clouds. In our Solar system, this phenomenon has been observed for Venus by means of the magnetometer on board the *Venus Express* (VEX) spacecraft. Another prominent mechanism entails the transfer of energy and momentum from the stellar wind to the upper ionosphere, thus resulting in the acceleration and escape of ions.

In principle, one can opt to analyze each of these processes separately and attempt to construct simple toy models to characterize their corresponding escape rates. However, such a strategy is rendered impractical in actuality because the mechanisms are numerous, and we lack semi-analytical prescriptions for all of them. Instead, it is more instructive to contemplate a toy model for weakly magnetized or unmagnetized planets—to wit, those with negligible magnetic moments. Venus is an archetypal example of an unmagnetized planet, while Mars could be placed in this category with some crucial caveats, since it possesses vestiges of crustal fields. Recent numerical modeling based on multispecies MHD simulations suggests that nonthermal ion escape mechanisms mediated by interactions with stellar winds are important for weakly magnetized Earth-sized planets (Brain et al. 2016; Airapetian et al. 2020).

Our approach partly mirrors the derivation presented in Zendejas et al. (2010). We commence our analysis by noting that the momentum carried by a single proton in the stellar wind is $m_H v_{sw}$, where m_H is the mass of a hydrogen atom and therefore approximately equal to the mass of a proton. Meanwhile, for a chemical species X in the planetary atmosphere to escape, the maximum momentum required is $m_X v_{\rm esc}$, where m_X is the mass of the particle and $v_{\rm esc}$ is the planet's escape velocity introduced in (4.28). Let us suppose that we are interested in the escape of O^+, which represents the major species lost from Earthlike as well as CO_2-dominated atmospheres (Dong, Jin, et al. 2018; Dong et al. 2019); for this species, it is apparent that

$m_X \approx 16\, m_H$. From (4.28), we see that the escape velocity for Earth-sized planets is $\mathcal{O}(10)$ km/s, whereas v_{sw} is typically $\mathcal{O}(100)$ km/s, although $v_{sw} \sim 10^3$ km/s is expected for M-dwarf exoplanets. Collectively, these heuristic relations indicate that $m_H v_{sw} \sim m_X v_{\rm esc}$ may be valid to an extent. Thus, the momentum carried by a stellar wind proton can be wholly imparted to a single unit of X for escaping the atmosphere.

In light of the above discussion, it is not unreasonable to suppose that the number of particles per unit time impinging on the planet is commensurate with the number of particles lost per unit time via atmospheric escape (denoted by $\dot{N}_p$). Recall that the stellar mass-loss rate is $\dot{M}_\star$ and the number flux at distance a is $\dot{M}_\star / \left(4\pi\, m_H a^2\right)$, assuming isotropic emission. Using the fact that the cross-sectional area of the unmagnetized planet is πR^2, we end up with

$$\dot{N}_p = \frac{\varepsilon}{4} \left(\frac{R}{a}\right)^2 \frac{\dot{M}_\star}{m_H}. \tag{4.31}$$

Naturally, not all of the incident protons in the stellar wind will contribute to atmospheric escape. This fact motivates the rationale behind the introduction of the efficiency factor ε in the right-hand side. On the basis of laboratory experiments and numerical simulations that have investigated the mixing of two fluids, we can adopt $\varepsilon \sim 0.1$ as the characteristic value (Zendejas et al. 2010).

Let us examine the validity of this formula for present-day Mars by choosing $R \approx 0.53\, R_\oplus$, $a \approx 1.52$ AU, and $\dot{M}_\odot \sim 2 \times 10^{-14}\, M_\odot\, \mathrm{yr}^{-1}$. Substituting these numbers into (4.31) leads us to $\dot{N}_p \sim 4.2 \times 10^{24}\, \mathrm{s}^{-1}$. In the current epoch, the total ion escape rate for Mars is $4.8 \times 10^{24}\, \mathrm{s}^{-1}$ as per both observational studies and detailed MHD simulations (Dong, Lee, et al. 2018). Next, let us consider Proxima b, whose parameters are specified to be $R \approx 1.07\, R_\oplus$, $a = 0.05$ AU, and $\dot{M}_\star \approx \dot{M}_\odot$. We end up with $\dot{N}_p \sim 1.6 \times 10^{28}\, \mathrm{s}^{-1}$, which is higher than the escape rate of $2.4 \times 10^{27}\, \mathrm{s}^{-1}$ calculated by Dong, Lingam, et al. (2017) but is close to the value of $1.1 \times 10^{28}\, \mathrm{s}^{-1}$ computed by Garcia-Sage et al. (2017). Thus, from the preceding two examples, we see that (4.31) yields reasonably accurate results at first glimpse.

An important point worth recognizing at this juncture is that $\dot{M}_\star$ evolves over time. The most common prescription used in the literature is $\dot{M}_\star \propto t^{-\beta}$, where t is the age of the star. For the Sun, based on empirical and theoretical data, it was proposed by Ó Fionnagáin and Vidotto (2018)

that β transitioned from 0.7 for $t < 2$ Gyr to 3.9 for $t > 2$ Gyr. An immediate consequence of (4.31) is that the atmospheric escape rates are substantially enhanced for young stars. This trend is confirmed by numerical investigations carried out by Dong, Lee, et al. (2018) for ancient and current Mars. The total ion escape rate at ~ 4 Ga was found to be $1.1 \times 10^{27}\ \mathrm{s}^{-1}$, whereas the corresponding value for modern Mars is $4.8 \times 10^{24}\ \mathrm{s}^{-1}$. Hence, it is likely that the ion escape rates for terrestrial planets in our Solar system at 4 Ga were more than two orders of magnitude higher relative to the present day. Nonthermal atmosphere escape could largely explain why Mars transitioned from a warmish climate and ~ 1 bar atmosphere to a cold one with surface pressure of $\sim 6 \times 10^{-3}$ bar today (Kite 2019; Jakosky 2019).

Next, we express the mass of the planetary atmosphere (M_{atm}) in terms of the surface pressure (P_s) as follows:

$$M_{\mathrm{atm}} = \frac{4\pi R^2 P_s}{g}, \tag{4.32}$$

where $g \approx g_{\oplus}\,(R/R_{\oplus})^{1.7}$ is the planet's surface gravity normalized to the Earth's value of $g_{\oplus} \approx 9.8\ \mathrm{m/s^2}$. From (4.31) and (4.32), we construct a characteristic timescale t_{SW} for atmospheric depletion via $t_{SW} = M_{\mathrm{atm}}/(\dot{N}_p \bar{m})$, with $\bar{m}$ signifying the average mass of the ionic species undergoing escape.[6] The advantage associated with formulating t_{SW} is that it can be expressed entirely in terms of basic stellar and planetary parameters by combining (4.15), (4.31), and (4.32).

If we estimate t_{SW} for Proxima b using the above formulae in conjunction with $M_{\mathrm{atm}} \sim 5 \times 10^{18}$ kg (mass of Earth's atmosphere), we find that the depletion timescale is $\mathcal{O}\left(10^8\right)$ yr. This result is consistent with detailed MHD simulations of both magnetized and unmagnetized cases that have yielded values ranging from $\mathcal{O}\left(10^7\right)$ to $\mathcal{O}\left(10^9\right)$ yr, with most converging on depletion timescales of $\mathcal{O}\left(10^8\right)$ yr. In contrast, for the TRAPPIST-1 system (with $P_s = 1$ atm assumed throughout), the innermost planet may be depleted of its atmosphere in $\mathcal{O}\left(10^8\right)$ yr, while the duration of atmospheric retention for the outermost planet appears to be $\mathcal{O}\left(10^{10}\right)$ yr (Dong, Jin, et al. 2018). The predicted rates of atmospheric

6. In formulating this timescale, we are expressly operating under the premise that the rate of outgassing and kindred processes (which increases M_{atm}) is much smaller, and therefore negligible, compared to the rate of atmospheric escape.

escape for the outer TRAPPIST-1 planets are lower than Proxima b, chiefly because TRAPPIST-1 is smaller and less active than Proxima b and was accordingly estimated to possess a comparatively weaker stellar wind.

Our discussion has hitherto been centered on unmagnetized planets. When it comes to magnetized planets, the situation is rendered much more complex because of the planet's internal magnetic field and the presence of a sizable magnetosphere. A crude method is to envision the magnetosphere as a spherical bubble and thus replace the geometric cross-sectional area πR^2 of the planet with πR_{mp}^2. In this case, (4.31) is preserved except for R being replaced by R_{mp}. Depending on the magnitude of R_{mp} relative to R, it is conceivable that the escape rate for magnetized planets could exceed those of their unmagnetized counterparts. Blackman and Tarduno (2018) formulated an analytic model in which the magnetized equivalent of (4.31) was derived. As the derivation lies beyond the scope of the book, only the final expression is provided:

$$\dot{N}_p^{(\mathrm{mag})} = \mathcal{Q} \cdot \dot{N}_p, \tag{4.33}$$

with $\dot{N}_p^{(\mathrm{mag})}$ denoting the escape rate for magnetized planets, the unmagnetized escape rate $\dot{N}_p$ is given by (4.31), and $\mathcal{Q}$ is defined as follows:

$$\mathcal{Q} \sim 7.1 \times 10^{-2} \left(\frac{\mathcal{K}}{0.1} \right) \left(\frac{R_{mp}}{R} \right)^2. \tag{4.34}$$

In the above formula, $\mathcal{K}$ is proportional to the ratio of magnetic reconnection (viz., changes in magnetic topology) velocity to the stellar wind velocity close to the planet. Hence, for $\mathcal{K} \sim 0.1$, we see that $\mathcal{Q} > 1$ is achieved when $R_{mp} \gtrsim 3.75\, R$. On the whole, the relationship between planetary magnetic fields and nonthermal atmospheric escape is subtle and nonmonotonic in nature and thus necessitates case-by-case studies (Gunell et al. 2018; Lingam 2019; Egan et al. 2019).

In summary, planets in the HZ of very low-mass stars ($\lesssim 0.1\, M_\odot$) with atmospheres comparable in mass to the Earth are susceptible to being depleted of their atmospheres over sub-Gyr timescales (i.e., on the order of 10 to 100 Myr). As we have seen in Chapter 3, the origin of life required a maximum timescale of ~ 0.8 Gyr, while technological intelligence (*H. sapiens*) emerged 4.5 Gyr after the Earth was habitable. Given that our powers of extrapolation from a single datum are highly limited, these timescales

ought not be considered sacrosanct since evolution on other worlds may operate at rates markedly different from those on Earth. Bearing this caveat in mind, it is plausible that Earth-analogs around low-mass stars (late-type M-dwarfs to be specific) might not give rise to complex, or even microbial, life since they might be stripped of their atmospheres. This difficulty could be bypassed by planets with thick atmospheres, since the depletion timescale t_{SW} is proportional to $M_{\rm atm}$.

There is yet another point that can be inferred at this stage. Earlier, we have seen how the duration of time that the planet may spend in the HZ (t_{HZ}) varies as a function of mass; it is quantified via (4.6). Now, suppose that the atmospheric depletion timescale is smaller than t_{HZ}. In this event, the maximum timescale over which biological evolution can unfold is dictated by t_{SW} instead of t_{HZ}. For the time being, let us suppose that $\mathcal{Q} \sim 1$, allowing us to use (4.31) and (4.33) interchangeably. By using $t_{SW} = M_{\rm atm}/(\dot{N}_p \bar{m})$ along with (4.5), (4.16), and (4.31) for Earth-analogs, we find

$$t_{SW} \sim 100 t_\odot \left(\frac{\Omega_\star}{\Omega_\odot}\right)^{-1.33} \left(\frac{M_\star}{M_\odot}\right)^{4.76}, \tag{4.35}$$

where the constant of proportionality has been calibrated on the basis of the depletion timescale for the Earth (Lingam & Loeb 2017d). Owing to the very high sensitivity to the stellar mass, it is easily verified that Earth-analogs around G-type stars like the Sun are not likely to have their atmospheres eroded by stellar winds during the stellar lifetime.

To recap, we have argued that there are two constraints on the maximum amount of time biological evolution can proceed: (1) duration of time the planet spends in the HZ and (2) duration of time required for atmospheric erosion by stellar winds. Apart from these two factors, numerous other constraints will exist, some of which we shall encounter later. Nevertheless, for the time being, we will define the maximum interval of habitability as $t_H = \min\{t_{HZ}, t_{SW}\}$. Choosing $\Omega \sim \Omega_\star$ in (4.35) and simplifying t_H, we end up with

$$t_H \sim 0.55\, t_\odot \left(\frac{M_\star}{M_\odot}\right)^{-2} \qquad M_\star > M_\odot,$$

$$t_H \sim 0.55\, t_\odot \left(\frac{M_\star}{M_\odot}\right)^{-1} \qquad 0.5 M_\odot < M_\star < M_\odot,$$

$$t_H \sim 0.46\, t_\odot \left(\frac{M_\star}{M_\odot}\right)^{-1.25} \qquad 0.41 M_\odot < M_\star < 0.5 M_\odot,$$
$$t_H \sim 100 t_\odot \left(\frac{M_\star}{M_\odot}\right)^{4.76} \qquad M_\star < 0.41 M_\odot. \tag{4.36}$$

The maximum duration of habitability has been plotted in the top panel of Figure 4.2. It can be verified from (4.36) or the figure that the upper bound on the abiogenesis timescale for Earth (0.8 Gyr) is attained only when $M_\star \gtrsim 0.22\, M_\odot$. Similarly, in order for $t_H \geq 4.5$ Gyr to be valid—the interval for the rise of technological intelligence on Earth—we require $0.32 \lesssim M_\star/M_\odot \lesssim 1.1$.

In Section 2.1, we argued that the total number of species on Earth is reasonably represented via exponential growth over time. Despite the fact that this exponential growth will not operate continuously over the entire interval of habitability (t_H), we shall work with this premise. The reason for doing so is that it provides us with a useful metric of peak biological complexity. In this case, the peak number of species ($N_{\rm peak}$) that can arise is given by

$$N_{\rm peak} \sim \exp\left(\frac{t_H}{\tau_c}\right) - 1, \tag{4.37}$$

where τ_c is the characteristic (*e*-folding) timescale. For now, we shall hold τ_c to be constant, which is tantamount to the assumption that the pace of evolution is similar on all habitable planets; we select $\tau_c \approx 163$ Myr in accordance with Section 2.1. We have plotted $N_{\rm peak}$ as a function of the stellar mass in the bottom panel of Figure 4.2. By inspecting the figure, we observe that $N_{\rm peak} < 1$ occurs when $M_\star \lesssim 0.15\, M_\odot$, thus implying that Earth-analogs around late-type M-dwarfs have a relatively lower likelihood of hosting life. This figure also indicates that only a subset of stars are predisposed toward having planets with biospheres comprising the same number of species as on Earth, where we have chosen the optimistic (but highly uncertain) estimate of $\sim 10^{12}$ species, as explained in Section 2.1.2; the corresponding mass range is $0.32 < M_\star/M_\odot < 1.1$.

4.2.3 Joule heating

In plasma physics, the resistivity η is responsible for the dissipation of energy in a system. The rate of dissipation of energy per unit volume $\dot{\mathcal{E}}$ for resistive MHD is expressible as

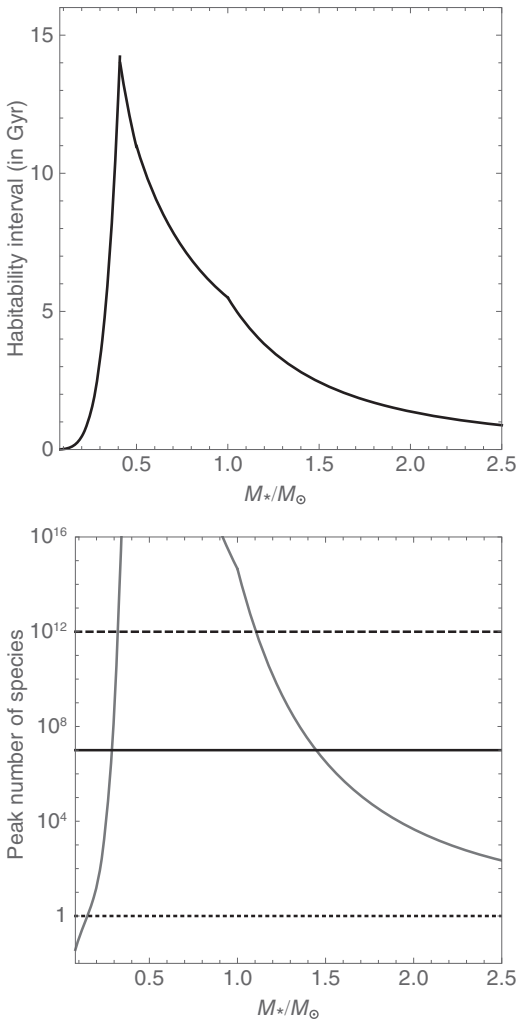


Figure 4.2 *Top:* Maximum habitability interval (in Gyr) as a function of stellar mass (in units of $M_\odot$). *Bottom:* Peak diversity $\left(N_{\rm peak}\right)$ attainable as a function of stellar mass. The black solid and dashed lines represent the presumed *current* number of eukaryotic and microbial species on Earth. The black dotted line denotes the regime where the peak number of species drops below unity, suggesting that abiogenesis is not feasible on such worlds. (© Manasvi Lingam and Avi Loeb.)

$$\dot{\mathcal{E}} = \mathbf{J} \cdot \mathbf{E} = \sigma_P |\mathbf{E}|^2, \tag{4.38}$$

where $\sigma_P = \eta^{-1}$ is the conductivity of the plasma. In SI units, the dimensions of $\dot{\mathcal{E}}$ are W/m^3. Instead, if we are interested in the energy flux, we must replace σ_P by Σ_P, where the latter represents the height-integrated conductivity: $\Sigma_P = \int \sigma_P \, dh$. Thus, we have

$$Q_E = \Sigma_P |\mathbf{E}|^2, \tag{4.39}$$

with Q_E denoting the energy flux dissipated in the plasma, with units of W/m^2. For the Earth, it is common to use $\Sigma_P \sim 1$–10 S (Kivelson & Ridley 2008); S denotes siemens, the SI unit used in measuring conductance.

The question of interest to us is, How much energy is transferred from the stellar wind to the ionosphere due to Joule heating? In tackling this problem, Kivelson and Ridley (2008) modeled the ionosphere as a spherical conductor (with finite height-integrated conductivity Σ_P) and the inflowing stellar wind as an electromagnetic wave. By doing so, it was demonstrated that the energy deposited in the ionosphere is equivalent to the transmitted energy from the incoming electromagnetic wave. We will not reproduce the entire derivation since it necessitates some specialized knowledge of plasma physics. The final expression for $|\mathbf{E}|$ is given by

$$|\mathbf{E}| = |\mathbf{E}_{sw}| \frac{2\Sigma_P^{-1}}{\Sigma_P^{-1} + \Sigma_A^{-1}}, \tag{4.40}$$

where $\Sigma_A^{-1} = \mu_0 v_A$ is the Alfvénic impedance and $v_A = B_{sw}/\sqrt{\mu_0 \rho_{sw}}$ is the Alfvénic velocity of the stellar wind. Inspecting the above expression, the reader may notice that this expression is identical to the magnitude of the transmitted electric field in the theory of transmission lines; this resemblance is not coincidental as the two models are formally equivalent. To complete our set of expressions, we observe that $|\mathbf{E}_{sw}| \sim v_{sw} B_{sw}$ is roughly valid (in the ideal MHD regime) after dropping an extra trigonometric factor arising from the cross product.

Clearly, there are a number of factors, each of which is subject to myriad uncertainties. Fortunately, we can consider three simple cases where the final expressions are much simplified. First, let us consider the scenario wherein $\Sigma_P \ll \Sigma_A$. In this event, we end up with

$$Q_E \sim 4\Sigma_P v_{sw}^2 B_{sw}^2. \tag{4.41}$$

Second, we consider the domain where $\Sigma_P \sim \Sigma_A$. Repeating the calculations, we find

$$Q_E \sim \Sigma_P v_{sw}^2 B_{sw}^2. \tag{4.42}$$

This result is identical to (4.41), apart from the missing factor of 4. Finally, we consider the case wherein $\Sigma_P \gg \Sigma_A$, which leads us to

$$Q_E \sim \frac{4\Sigma_P^{-1} \rho_{sw} v_{sw}^2}{\mu_0} \tag{4.43}$$

after using the definition for Σ_A. The key point worth appreciating here is that (4.41) and (4.42) are proportional to the magnetic pressure (P_{mag}), while (4.43) is proportional to the dynamic pressure (P_{dyn}). We analyzed these terms (kinetic and dynamic pressures) in Section 4.2.1 and concluded that they are several orders of magnitude higher for Earth-analogs of late-type M-dwarfs compared to the Earth.

In the case of the TRAPPIST-1 planets in the HZ, one can use (4.42) without sacrificing much accuracy. For these planets, it was found that $Q_E \sim 0.5$–1 W/m^2 (Cohen et al. 2018), implying that a significant fraction (between 10–50 percent) of the stellar wind input energy is transferred to the ionosphere as Joule heating. Moreover, this energy flux is around 1 percent of the total stellar irradiance received by the planets and about an order of magnitude higher than the extreme ultraviolet (EUV) flux that we shall tackle shortly. Thus, Joule heating is likely to be significant for habitable planets orbiting late-type M-dwarfs. The deposition of high quantities of energy in the ionosphere is expected to impact planetary habitability, but the biological effects have not been sufficiently investigated to date. Hence, we shall wrap up our analysis at this juncture, although we hope that this issue will be subject to further explorations in the future.

4.3 STELLAR ELECTROMAGNETIC RADIATION

Of the myriad facets of planetary habitability dependent on stellar factors, those connected to the electromagnetic radiation originating from the host star have been the most widely studied and appreciated. This field represents a vast topic, owing to which our discussion, by necessity, will be both brief

Table 4.2 List of *potential* positive and negative consequences for habitability associated with stellar electromagnetic radiation

Range	Ramifications	M-Earths
XUV	Land-water planets from water worlds through H_2O photolysis	High
XUV	Atmospheric O_2 buildup via H_2O photolysis for complex life	High
UV-C	Synthesis of building blocks for prebiotic chemistry	Low
UV-Bio	Selection agent for evolutionary innovations and speciation	Low
PAR	Facilitating photosynthesis	Low
XUV	Total desiccation of oceans via H_2O photolysis	High
UV-Bio	Damage to biomolecules (e.g., DNA)	Low
UV-Bio	Suppressing photosynthesis	Low

Notes: XUV ($\sim$ 0.6–120 nm), UV-C ($\sim$ 200–280 nm), UV-Bio ($\sim$ 200–400 nm), PAR ($\sim$ 400–750 nm). In the third column, *High* and *Low* denote the energy fluxes of electromagnetic radiation (in the appropriate wavelength range) received by Earth-analogs around M-dwarfs (M-Earths) with respect to the Earth. The first five rows represent the benefits attributable to electromagnetic radiation; the last three rows correspond to the detrimental effects.

and selective. A compact overview of the beneficial and detrimental impacts on bio-habitability stemming from electromagnetic radiation is presented in Table 4.2.

4.3.1 Evaporation of oceans and buildup of oxygen

It has been appreciated since the mid-twentieth century that extreme ultraviolet (EUV) photons and X-rays—collectively known as XUV radiation—are capable of driving atmospheric escape of neutral particles (Lammer 2013; Owen 2019). First, a clarification regarding the notation is in order. By *EUV radiation,* we refer to photons with wavelengths in the range of 10–120 nm, whereas *XUV radiation* encompasses photons with wavelengths of 0.6–120 nm.

To compute the atmospheric mass-loss rate $\dot{M}_p$ (in units of kg/s) facilitated by XUV radiation, let us embark on the construction of a toy model. We start by introducing the energy flux of XUV photons close to the planet by $\mathcal{F}_{XUV}$. The geometric cross-sectional area of the planet is πR^2. It is evident that the product of these two quantities yields the energy inflow rate (units of W). In equilibrium, this inflow of energy provides the impetus for particles to escape the planet's atmosphere. In other words, it must approximately equal $\frac{1}{2}\dot{M}_p v_{\rm esc}^2$, where $v_{\rm esc}$ is given by (4.28). Thus, we end up with

$$\frac{G\dot{M}_p M}{R} \sim \pi R^2 \mathcal{F}_{XUV}, \tag{4.44}$$

and by solving for $\dot{M}_p$, we obtain

$$\dot{M}_p = \frac{\eta_{XUV}}{K_{\text{eff}}} \frac{\pi R^3 \mathcal{F}_{XUV}}{GM}, \tag{4.45}$$

where η_{EUV} quantifies the heating efficiency of XUV photons and K_{eff} is an additional factor of order unity to account for the effects of stellar tidal forces (Erkaev et al. 2007). There exist several variants of (4.45) in the literature, but the functional dependence on planetary and stellar parameters is preserved. This regime is conventionally referred to as the *energy-limited* regime because of our reliance on energy balance. In reality, however, a number of distinct regimes are feasible (Owen & Alvarez 2016). For instance, the photon number could serve as the limiting factor instead of the energy, which leads to a different expression for the atmospheric mass-loss rate. In this case, the mass-loss rate is given by

$$\dot{M}_p \sim \frac{\pi \bar{m} R^2 \mathcal{F}_{XUV}}{h\bar{\nu}}, \tag{4.46}$$

where $h\bar{\nu}$ denotes the mean energy of the photons that contribute to atmospheric escape. This formula is obtained by first computing the photon *number* flux from the energy flux via $\mathcal{F}_{XUV}/(h\bar{\nu})$. This expression is multiplied with the geometric cross-sectional area to obtain the total particle-loss rate. Finally, in order to convert it into the mass-loss rate, we must multiply the expression with the mean mass $\bar{m}$ of the chemical species undergoing escape.

At first glimpse, it seems tempting to argue that $\dot{M}_p$ should be smaller for planets around low-mass stars. This conclusion would appear to follow if one treats the star as a classical black body, implying that $\mathcal{F}_{XUV}$ ought to be much smaller for low-mass stars since the bulk of their radiation is in the infrared. However, while this line of reasoning seems reasonable, it is not correct. In the case of solar-type stars, most of the emission is from the photosphere. On the other hand, when we consider M-dwarfs, the layers exterior to the photosphere—the chromosphere and transition regions—contribute significantly to the fluxes of high-energy photons (Linsky 2019). More precisely, an essential characteristic of M-dwarfs is that their ratios of far-UV

(117–175 nm) to near-UV (175–320 nm) fluxes tend to be $\sim$ 1000 times higher than the corresponding ratio for the Sun (Tian et al. 2014; France et al. 2016; Linsky 2019). Moreover, one must also account for the fact that the HZs of low-mass stars are situated much closer. Hence, it is not surprising that the EUV flux at Proxima b has been constrained to be $\sim$ 30 times the corresponding value at the Earth.

One more consideration has garnered a great deal of attention in recent times. It is well established that all stars pass through a pre-main-sequence (PMS) phase, during which their luminosity is powered primarily by gravitational contraction as opposed to the conventional route entailing the fusion of hydrogen. Two striking aspects of the PMS phase stand out for late-type M-dwarfs with $M_\star \lesssim 0.1\, M_\odot$. First, the stellar luminosity at the beginning of the PMS phase may be two orders of magnitude higher than the stage at which the star enters the main-sequence. Second, the duration of time that the star spends in the PMS phase can last up to a few Gyr. As a result, planets around such stars that would otherwise be situated in the HZ during the main sequence do not lie in this annular region during the PMS phase because of the enhanced luminosity. To put it differently, these planets would exceed the runaway greenhouse threshold, thereby resulting in significant water losses (Ramirez & Kaltenegger 2014).

The rationale behind the loss of H_2O is that the water vapor accumulated in the atmosphere as a result of the greenhouse effect would be subjected to photolysis to yield hydrogen and oxygen. As the former species is much less massive than the latter, it would be more susceptible to atmospheric escape. This is simple to verify by converting the mass-loss rate (4.45) to the particle-loss rate by dividing the former by the mass of the chemical species under consideration. Hence, hydrogen is expected to be preferentially lost to space, leaving behind massive O_2 atmospheres. It should be borne in mind that this picture has been deliberately simplified, and the actual amount of water lost will depend on the XUV flux, the planet's mass, and the water inventory.

Given the XUV flux, it is feasible to estimate the quantity of water lost or, equivalently, the rate at which abiotic O_2 will build up in the atmosphere. We refer the reader to Luger and Barnes (2015) for the full derivation, which we do not reproduce here since it necessitates a number of prerequisites. The basic idea, however, can be qualitatively understood as follows. Strong hydrodynamic flows involving a lighter species (hydrogen in our case) are capable of dragging along heavier species with them.

The criterion for determining whether a heavy species (oxygen in our model) is dragged along relies on the concept of the crossover mass (m_c), defined as

$$m_c = m_H + \frac{k_B T_F F_H}{b_0 g X_H}, \tag{4.47}$$

where m_H is the mass of a hydrogen atom, k_B refers to the Boltzmann constant, b_0 is the binary diffusion coefficient (which quantifies the degree of collisions between two species), T_F is the temperature of the flow, $X_H = 1/3$ denotes the molar mixing ratio at the base of the flow, and F_H is the average outward particle flux of hydrogen. The criterion for facilitating the escape of oxygen is $m_c > m_O$, where m_O is the mass of an oxygen atom. It turns out that this condition can be rewritten as $\mathcal{F}_{XUV} > \mathcal{F}_c$, where the critical XUV flux $\mathcal{F}_c$ is defined as

$$\mathcal{F}_c \sim 0.18\,\mathrm{J\,m^{-2}\,s^{-1}} \left(\frac{M}{M_\oplus}\right)^2 \left(\frac{R}{R_\oplus}\right)^{-3} \left(\frac{\eta_{XUV}}{0.30}\right)^{-1} \tag{4.48}$$

and can be further simplified by using the mass-radius relationship $M \propto R^{3.7}$. The rate at which O_2 builds up in the atmosphere ($\dot{P}_{O_2}$) is measured in units of bars/Myr. It depends on the magnitude of $\mathcal{F}_{XUV}$ and is given by

$$\begin{aligned} \dot{P}_{O_2} &\sim 5.35\,\mathrm{bars/Myr} \left(\frac{M}{M_\oplus}\right)^2 \left(\frac{R}{R_\oplus}\right)^{-4} && \mathcal{F}_{XUV} \geq \mathcal{F}_c, \\ \dot{P}_{O_2} &\sim 0.138\,\mathrm{bars/Myr} \left(\frac{\mathcal{F}_{XUV}}{\mathcal{F}_\oplus}\right) \left(\frac{R}{R_\oplus}\right)^{-1} \left(\frac{\eta_{XUV}}{0.30}\right) && \mathcal{F}_{XUV} < \mathcal{F}_c, \end{aligned} \tag{4.49}$$

where $\mathcal{F}_\oplus \sim 4.6 \times 10^{-3}\,\mathrm{J\,m^{-2}\,s^{-1}}$ is the XUV flux incident on the Earth. In comparison, the rate of O_2 supplied to Earth's atmosphere due to the burial of organic matter, which constitutes the major biological source, is ~ 0.06 bars/Myr.

For Proxima b, it has been estimated that $\mathcal{F}_{XUV} \approx 60\,\mathcal{F}_\oplus$ (Ribas et al. 2016). Choosing $M = 1.3\,M_\oplus$ and using the mass-radius relationship in (4.48) imply that $\mathcal{F}_c \approx 53\,\mathcal{F}_\oplus$. Hence, the first expression in (4.49) must be employed, since $\mathcal{F}_{XUV} > \mathcal{F}_c$. Next, let us consider TRAPPIST-1g—the

outermost planet of TRAPPIST-1 within the classical HZ. Given that $L_{XUV} \sim 7 \times 10^{19}$ W for the parent star (Bourrier et al. 2017), the XUV flux received at TRAPPIST-1g is computed from $\mathcal{F}_{XUV} = L_{XUV} / \left(4\pi a^2\right)$, which leads us to $\mathcal{F}_{XUV} \sim 0.11\,\mathrm{J\,m^{-2}\,s^{-1}}$ after using $a \approx 0.047$ AU. Next, we make use of $M \approx 1.148\,M_{\oplus}$ and $R \approx 1.148\,R_{\oplus}$ for TRAPPIST-1g (Grimm et al. 2018) in (4.48), thus resulting in $\mathcal{F}_c \sim 0.16\,\mathrm{J\,m^{-2}\,s^{-1}}$. Hence, we end up with $\mathcal{F}_{XUV} < \mathcal{F}_c$, indicating that the second formula from (4.49) should be utilized for TRAPPIST-1g.

In Section 3.3, we documented how oxygenic photosynthesis produces oxygen as a product, owing to which the latter has been extensively investigated as a signpost of biological activity. However, in our case, the buildup of O_2 in the atmosphere can become very significant without any underlying biological causes. Hence, in such instances, it would give rise to a *false positive* in the search for biosignatures—that is, a false indication of life. There have been a number of studies undertaken in this subject, which concluded that the abiotic production of massive amounts of O_2 as well as ozone (O_3) are feasible through the photolysis of H_2O and CO_2 by XUV radiation. In most instances, however, the dynamical evolution of the planetary interior was not taken into account. In particular, the formation of magma oceans could result in the efficient absorption of abiotic O_2 released via XUV photolysis (Wordsworth et al. 2018). In Section 6.4.1, we shall briefly examine some potential methods that seek to differentiate between abiotic and biotic O_2 in the atmosphere.

It was pointed out earlier that the UV photolysis of water is accompanied by the depletion of oceans, with the latter having important ramifications for habitability as liquid water constitutes one of the chief requirements for life. Before discussing the consequences of water loss, it is worth mentioning the studies that have been undertaken for Proxima b and the TRAPPIST-1 planetary system. For the former, Ribas et al. (2016) predicted that $< 1 M_{oc,\oplus}$ has been lost over its history, where $M_{oc,\oplus} \approx 1.4 \times 10^{21}$ kg denotes the mass of Earth's oceans today. The loss of H_2O from Proxima b may result in a thick O_2 atmosphere with a pressure on the order of 100 bars. For the TRAPPIST-1 planets, Bourrier et al. (2017) estimated that the current mass-loss rate of the oceans ranges from $8.2 \times 10^{-3}\,M_{oc,\oplus}$/Myr for TRAPPIST-1b to $2.9 \times 10^{-4}\,M_{oc,\oplus}$/Myr for TRAPPIST-1h. It was further estimated that the TRAPPIST-1 planets lying within the orbit of TRAPPIST-1g may have lost more than $20\,M_{oc,\oplus}$ over the age of the star (~ 7.6 Gyr).

This leads us to an important point that we shall elaborate further in the next chapter. One of the unique, and underappreciated, aspects about the Earth is that the fraction of the surface covered by land (0.3) is similar to that covered by water (0.7). Yet, in the majority of instances, there is no guarantee that this balance between the land and water fractions will be manifested. In other words, one can expect most planets to be covered almost entirely by water or land as a result of two distinct mechanisms at work. First, the initial water inventory of the planets is specified by a wide array of physical mechanisms that facilitate the delivery of H_2O during terrestrial planet formation. As a result, the water inventories of planets are anticipated to vary widely.

In turn, this implies that there will be several worlds with water inventories much higher than the Earth. One of the best examples in this regard are the TRAPPIST-1 planets themselves. The water mass fractions (mass of H_2O divided by the planetary mass) for these planets depend on the models used for the planetary interiors, but there are strong grounds for supposing that the water mass fraction for some of the TRAPPIST-1 planets might reach as much as 20–30 percent (Unterborn, Hinkel, et al. 2018). In this scenario, the water inventory would be more than two orders of magnitude higher than that of the Earth, resulting in worlds comprising only oceans on the surface. Such *water worlds* are anticipated to be fairly common on both observational and theoretical grounds.

On the other hand, we have seen earlier that high XUV fluxes during the long PMS phase of low-mass stars are very effective in terms of depleting several oceans' worth of water. It would therefore appear that worlds endowed with a moderate initial water inventory could be subject to desiccation, thus ending up as desert planets. Hence, to summarize, there exist multiple channels for water delivery as well as subsequent water depletion. In order for the surface fractions of water and land to become comparable to one another, a certain degree of fine-tuning is probably necessary to ensure that the water-loss mechanisms deplete just the right amount of H_2O to avoid desert planets or water worlds. Thus, as per our arguments, we may anticipate the following two points for low-mass stars: (1) worlds with Earthlike water inventories are rare, and (2) the water inventory is describable by a bimodal distribution. In the next chapter, we shall explore the validity of this conjecture and the ensuing biological ramifications for desert or ocean worlds.

4.3.2 Origin of life

It goes without saying that the steps involved in the origin of life on Earth, to say nothing of other planets and moons, are still shrouded in mystery. We examined the ongoing hypotheses, modeling, and experiments pertaining to abiogenesis in Chapter 2. In particular, we saw in Section 2.3.1 that UV radiation plays an important role in prebiotic chemistry by enabling the synthesis of the precursors of biomolecules such as proteins, lipids, and nucleic acids. Although we discussed these matters in depth, a quick recap of the salient points is in order.

- UV light confers a selective advantage to RNA-like molecules, implying that it may have been a key factor in enabling their polymerization.
- A large number of laboratory experiments give rise to complex organic mixtures ("tar") that are effectively dead ends insofar as the origin of life is concerned. This issue, sometimes dubbed the *asphalt problem,* might be overcome in specific geochemical environments wherein UV radiation plays a potentially vital role.
- The monomer units of RNA (nucleotides) are comparatively stable when irradiated by UV photons, which has led some scientists to propose that they originated in the high-UV environments on early Earth.
- Synthesizing the building blocks of biomolecules is very challenging in the absence of regular manual intervention and in settings that are ostensibly reminiscent of the early Earth. Remarkably, many of the recent breakthroughs in this area are reliant on UV light. The prebiotic compounds synthesized include (1) RNA and DNA nucleotides, (2) basic sugar-related molecules, (3) antecedents of nucleic acids, amino acids, lipids, and carbohydrates via interlinked pathways based on plausible feedstock molecules (e.g., hydrogen cyanide), and (4) iron-sulfur clusters, which are essential components of most metabolic networks.

To sum up, there is compelling, but not yet sufficient, evidence suggesting that UV light played an important role in abiogenesis.

For the sake of simplicity, let us suppose that this conjecture is correct and that UV radiation is essential for the origin of life on other worlds. This leads us toward the question of which Earth-analogs receive low fluxes of UV radiation. This question has a quantitative answer: planets in the HZ of M-dwarfs receive bioactive UV fluxes (in the wavelength range $200 < \lambda < 400$ nm) at the surface that are approximately 100–1000 times smaller compared to the UV flux incident on early Earth (Rugheimer et al. 2015). Since the bioactive UV flux reaching the surface depends on stellar properties,[7] the region around the host star where the planet receives a sufficiently high photon flux for efficient UV-mediated prebiotic chemistry (christened the *abiogenesis zone*) will not necessarily overlap with the classical HZ.

This matter has been investigated by several groups, of which the most comprehensive is the study by Rimmer et al. (2018). The authors based their analysis on two recent experiments in UV-based prebiotic chemistry, elucidated in Section 2.3.1, that led to the synthesis of biomolecular building blocks. There are two basic reactions where UV plays a direct role: (1) in the first network, it contributes to the formation of HS^- from hydrogen sulfide (H_2S), and (2) in the second network, it plays an analogous role in synthesizing SO_3^{2-} from sulfur dioxide (SO_2). In order for the photochemical reactions (entailing UV light) to dominate over the reactions that operate in the absence of radiation, the UV fluxes in the wavelength range $200 < \lambda < 280$ nm at the planetary surface must exceed a critical value. In the case of (1), it is given by

$$\mathcal{F}_c\left(\mathrm{HS}^-\right) \sim 10.6\,\mathrm{J\,m^{-2}\,s^{-1}}, \tag{4.50}$$

whereas for (2) we must use

$$\mathcal{F}_c\left(\mathrm{SO}_3^{2-}\right) \sim 4.5 \times 10^{-2}\,\mathrm{J\,m^{-2}\,s^{-1}}. \tag{4.51}$$

It is evident that the fluxes must be higher for the photochemical reaction network involving HS^- to function with respect to the one that relies on SO_3^{2-}. It turns out that the critical flux (4.51) is satisfied only when $T_\star > 4400$ K, which is roughly equivalent to $M_\star \gtrsim 0.7\,M_\odot$. In Figure 4.3,

7. The UV flux incident on the surface is also dependent on planetary characteristics such as the atmospheric composition and column density.

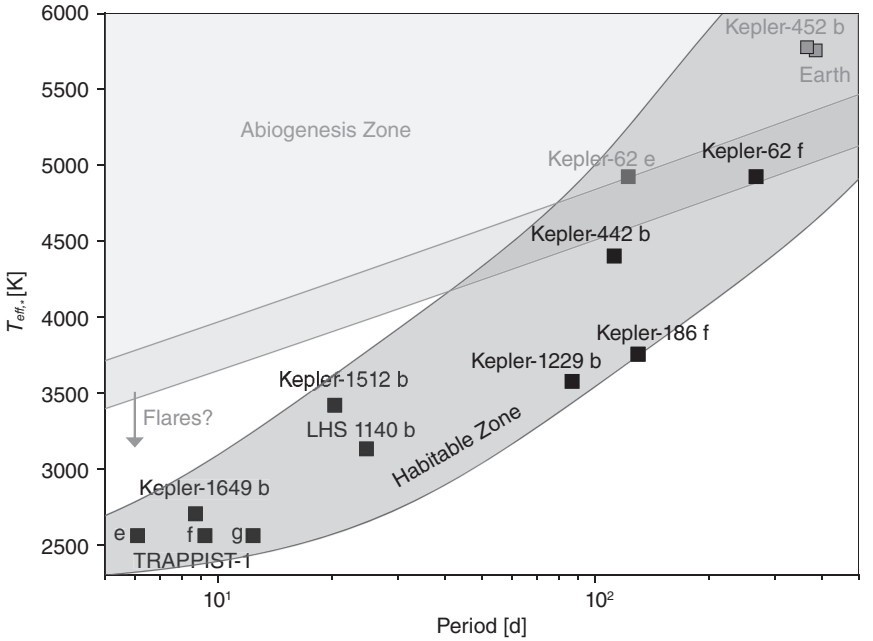


Figure 4.3 The effective temperature of the star ($T_\star$) is plotted on the *y*-axis, while the orbital period the planet is plotted on the *x*-axis. The lower shaded region denotes the classical HZ, whereas the upper shaded region signifies the Abiogenesis Zone, wherein the stellar UV flux is high enough to ensure that the yield of photochemical products is sufficiently high ($\sim$ 50 percent). The intermediate overlapping shaded region accounts for the error bars carried over from UV-centric laboratory experiments. (© The Authors. CC-BY-4.0. *Source:* Paul B. Rimmer, Jianfeng Xu, Samantha J. Thompson, Ed Gillen, John D. Sutherland, and Didier Queloz [2018], The origin of RNA precursors on exoplanets, *Science Advances* 4[8]: eaar3302, fig. 4.)

the abiogenesis and HZ zones have been plotted as a function of the stellar temperature and the orbital period of the planet. For all stars with $T_\star \lesssim 4000$ K, we see that the two zones do not overlap. Hence, this trend indicates that low-mass stars have a slender likelihood of facilitating abiogenesis via UV-driven prebiotic chemistry.

Hitherto, we have restricted ourselves to examining whether the incident UV fluxes are sufficient to power the synthesis of biomolecular precursors. However, it is important to appreciate the temporal aspect as well—namely, that the rates of prebiotic reactions are conceivably dependent on the bioactive UV fluxes (Ranjan et al. 2017). This hypothesis

implies that planets that receive lower UV fluxes may witness abiogenesis occurring over a longer timescale. On the basis of this idea, Lingam and Loeb (2018d) proposed a simple model wherein the reaction rates were proportional to the bioactive UV fluxes incident on Earth-analogs, such that the abiogenesis timescale (t_0) was inversely proportional to the bioactive UV flux for a given Earth-analog. Thus, it was found that

$$t_0 \sim t_{0,\oplus} \left(\frac{M_\star}{M_\odot} \right)^{-3} \qquad M_\star \lesssim M_\odot,$$
$$t_0 \sim t_{0,\oplus} \left(\frac{M_\star}{M_\odot} \right)^{-1} \qquad M_\star \gtrsim M_\odot, \tag{4.52}$$

where $t_{0,\oplus} \lesssim 0.8$ Gyr. Henceforth, we will adopt the conservative choice of $t_{0,\oplus} = 0.08\, t_\odot$ (0.8 Gyr) since the earliest definitive signatures of life documented to date are from ~ 3.7 Ga (see Chapter 2). We are now confronted with two distinct timescales: the putative timescale for the origin of life (t_0) and the maximum interval over which the planet is habitable (t_H). It is instructive to define a habitability likelihood $\Gamma = t_H/t_0$. The advantage of this metric is that $\Gamma < 1$ indicates that the corresponding Earth-analogs are unlikely to host life. We have plotted Γ as a function of the stellar mass in Figure 4.4. We observe that stars with $M_\star \lesssim 0.4\, M_\odot$ have $\Gamma < 1$, suggesting that planets in the HZs of these stars have a lower likelihood of hosting life. We can also see that this function peaks around $M_\star = M_\odot$, but it must be noted that this is partly due to our choice of fitting functions while deriving our heuristic timescales.

It would appear from our discussion thus far that planets around low-mass stars are not predisposed to host biospheres due to low fluxes of bioactive UV radiation. This potential downside may be ameliorated to an extent by stellar flares, which emit high fluxes of electromagnetic radiation transiently—we shall return to this topic in Section 4.4.1.

4.3.3 Major evolutionary events

The basic thrust of Chapter 3 was to identify a series of evolutionary innovations that have unfolded on the Earth. Five major evolutionary events (MEEs) were identified in total: the emergence of life, oxygenic photosynthesis, eukaryotes, animals and technological intelligence. We argued that, for a *subset* of habitable worlds, the same evolutionary breakthroughs may

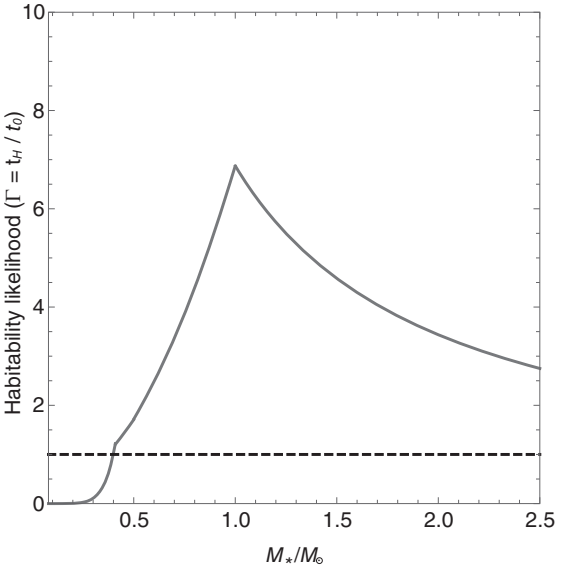


Figure 4.4 The propensity of a planet to host life (i.e., the ratio of its habitability duration to the abiogenesis timescale) as a function of the stellar mass $M_\star$. The dashed horizontal line demarcates the regime $\Gamma = 1$. (© Manasvi Lingam and Avi Loeb.)

unfold. This statement is by *no* means akin to claiming that these MEEs are inevitable or that they represent the only feasible trajectories in evolutionary space. Moreover, in Section 3.10, we presented a mathematical formulation in which each MEE was treated as a critical hard step.

We can apply the critical steps model to arrive at some interesting conclusions. For starters, let us define the total number of critical steps as n ($n = 5$ for Earth). In this chapter, we have delineated two timescales thus far for habitability: t_0 and t_H. Recall that abiogenesis is the first critical step, implying that there are $n-1$ critical steps after it. These $n-1$ critical steps must unfold over a timescale $\Delta t = t_H - t_0$, provided that $\Delta t > 0$. As mentioned previously, in the event that $\Delta t < 0$, the $n-1$ critical steps cannot emerge since there is insufficient time for even abiogenesis to occur. The cumulative probability $\mathcal{P}_n(M_\star)$ for this to take place is computed by replacing $r \to n-1$, $n \to n-1$, and $t \to \Delta t$ in (3.49)—this amounts to saying that the $n-1$ steps must unfold within Δt. Thus, we end up with

$$\mathcal{P}_n(M_\star) = \left(1 - \frac{t_0}{t_H}\right)^{n-1}. \tag{4.53}$$

An important point worth recognizing is that this only represents the *ideal* probability—that is, assuming that t_0 and t_H represent the only constraints and there are no other limitations set by other factors like climate, nutrient availability, and so on.

Now, there are two cases of interest, depending on whether we are interested in planets with detectable biosignatures or technosignatures. In this analysis, we shall focus only on the subset of planets whose MEEs are analogous to Earth's.

- **Probability of technological intelligence:** As per our preceding assumptions, we must work with $n = 5$. For such worlds, it will be theoretically feasible to detect signatures of technological intelligence by searching for technosignatures.
- **Probability of *detectable* microbial life:** When it comes to detecting microbial life, some of the most well-known biosignatures like oxygen and ozone are not detectable at low concentrations. As a result, although the Earth was teeming with life throughout much of its history, the vanishingly low levels of O_2 and O_3 in the atmosphere until the Great Oxidation Event would have yielded a *false negative*. Setting aside the false positives we encountered earlier for the time being, the evolution of oxygenic photosynthesis played a vital role in amplifying the oxygen levels. Thus, in this instance, it is necessary to choose $n \geq 2$.

We have therefore plotted (4.53) as a function of $M_\star$ in the top panel of Figure 4.5. The peak occurs at $M_\star = M_\odot$, which is a consequence of the ansatzen chosen for t_0 and t_H. It is also evident that the curves rise sharply at $M_\star \approx 0.5 M_\odot$. This figure predicts that an Earth-analog orbiting a G-type star has the highest probability of successfully completing the critical steps necessary for the emergence of detectable microbial or technological species.

Until now, we have only concerned ourselves with the prospects for life on a single Earth-analog around a given star. Yet, it must be recognized that the abundance of stars is dependent on their mass, since low-mass stars are known to be more numerous than high-mass ones. Another point to highlight is that the total number of Earth-sized planets in the HZ per star appears to be weakly dependent on the mass of the host star (Kaltenegger

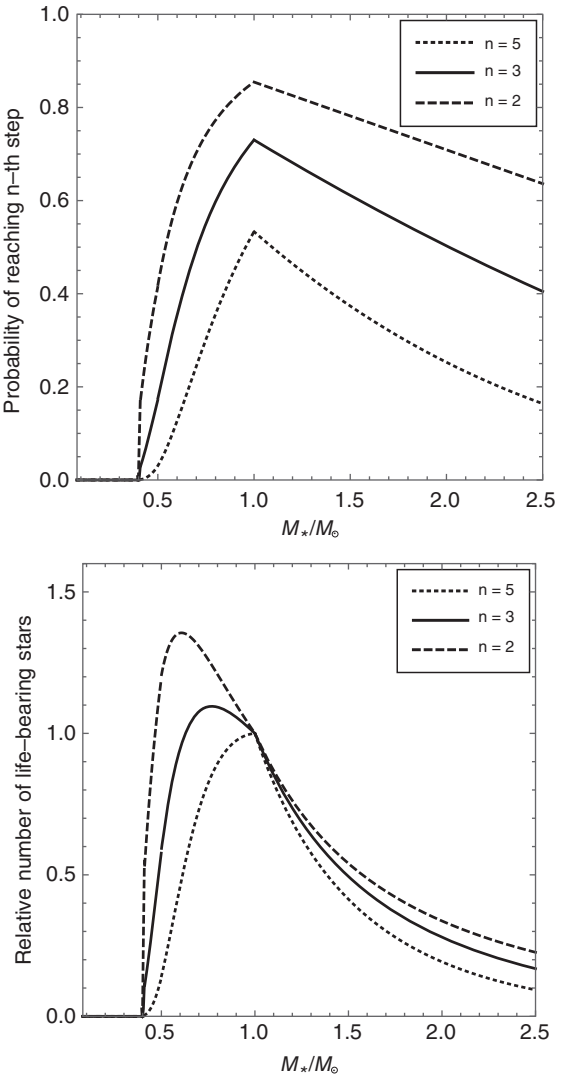


Figure 4.5 *Top:* The probability of attaining the n-th critical step as a function of the stellar mass $M_\star$ (in units of $M_\odot$) for different cases of n; this plot is based on equation (4.53). *Bottom:* The relative number of stars ($\zeta_\star$) with life-bearing planets, which have successfully passed through n critical steps as a function of the stellar mass $M_\star$ (in units of $M_\odot$). (© Manasvi Lingam and Avi Loeb.)

2017). Hence, we can calculate the relative number of Earth-analogs in our Galaxy with detectable microbial or technological species via $\zeta_\star = \mathcal{N}_\star / \mathcal{N}_\odot$, where the total number of stars $\mathcal{N}_\star$ is defined as follows:

$$\mathcal{N}_\star = \mathcal{P}_n (M_\star) \frac{dN_\star}{d (\ln M_\star)} \tag{4.54}$$

In the above equation, $dN_\star / d (\ln M_\star)$ refers to the number of stars per logarithmic mass interval and is readily determined from the stellar initial mass function (IMF), which is an empirical function that provides the number of stars in a given mass interval. We shall not go into the details here, as they have been worked out in Lingam and Loeb (2019f). The final result has been plotted in the bottom panel of Figure 4.5 from which, several conclusions immediately stand out. First, we find that the peak occurs at $M_\star \approx M_\odot$ for $n = 5$, thus implying that Sunlike stars in our Galaxy ought to be the most numerous insofar as having planets with technological intelligence is concerned. Second, when $n = 2$, the peak is manifested at $M_\star \approx 0.6 M_\odot$ (and at $M_\star \approx 0.77 M_\odot$ for $n = 3$), indicating that K-type stars are potentially the most numerous in terms of hosting planets with microbial, but detectable, life.[8] Finally, for our choices of n, it can be verified that $\zeta_\star$ does not vary much in the range $0.5 M_\odot < M_\star < 1.5 M_\odot$. As a result, it is conceivable that these stars represent the most promising avenues in the search for life.

4.3.4 Mutagenic effects of UV radiation

The fact that UV radiation, especially in the wavelength range of $180 < \lambda < 300$ nm, has deleterious effects is widely known. One of the best examples is the damage to human skin wrought by UV radiation, including its propensity for causing skin cancer. In addition, UV radiation has a number of harmful effects such as inhibiting photosynthesis in both land plants and plankton in the oceans (especially the latter). Moreover, it has been documented to cause damage to DNA and other biomolecules—this can occur when UV radiation promotes the formation of hydrogen peroxide, which leads to the formation of reactive oxygen species that are capable of

8. This category of stars also has the benefit of supporting long-lived and tangible biosignatures (Arney 2019), a field that we shall address in Chapter 6.

causing breaks in DNA as well as the alteration of the nucleobases and sugar molecules.

Although it is undeniable that UV radiation damages ecosystems at various trophic levels, there has been a tendency in some studies of exoplanetary habitability to overstate the dangers posed by UV radiation. It must be appreciated at the outset that a number of geochemical environments are capable of effectively shielding lifeforms from UV radiation. One of the best-studied examples is the shielding conferred by a mixture of prebiotic and inorganic compounds in oceans. At UV wavelengths of $\lambda \sim 260$ nm, only ~ 1 percent of the surface radiation survives after passage through $\sim 2 \times 10^{-3}$ m of ocean water (Cleaves & Miller 1998). Another important shielding agent is organic-rich hazes in the atmosphere—akin to those currently observed in Titan—that reduces the amount of UV flux reaching the surface by ~ 97 percent compared to the haze-free case (Arney et al. 2016). Apart from these naturally occurring environments, organisms on Earth (and presumably elsewhere) have evolved myriad adaptations to screen themselves from UV radiation; they have also found innovative solutions for repairing their DNA. A review of this fascinating subject can be found in Cockell and Knowland (1999).

A commonly used metric in studies involving UV radiation is the biologically effective irradiance (BEI), defined as follows:

$$F_{\rm eff} = \int_{\lambda_1}^{\lambda_2} \mathcal{F}_{\lambda}^{(\rm surf)} \mathcal{S}(\lambda)\, d\lambda \tag{4.55}$$

Here, $F_{\rm eff}$ denotes the BEI, $\mathcal{F}_{\lambda}^{(\rm surf)}$ is the spectral flux density at the surface, and $\mathcal{S}(\lambda)$ is the action spectrum for a given biomolecule. In qualitative terms, the action spectrum quantifies the relative biological response (and the extent of damage) and is wavelength dependent. The BEI is computed for a particular wavelength range. Table 4.3 provides the BEI for Earth-analogs around different stars relative to modern-day Earth. The action spectrum for DNA was utilized with $\lambda_1 = 182$ nm and $\lambda_2 = 370$ nm. From this table, two broad inferences can be drawn. First, regardless of the actual epoch, the BEI for an Earth-analog around a star with $M_\star \approx 0.1\, M_\odot$ is about two orders of magnitude lower than that of the Earth. Second, the presence of an ozone layer (related to the buildup of atmospheric oxygen) is crucial in determining the BEI. For instance, the BEI at 3.9 Ga on Earth was about 600 times higher than the present-day value.

Table 4.3 Biologically effective irradiance (F_{eff}) for Earth-analogs around stars with different effective temperatures ($T_\star$) relative to modern-day levels

$T_\star$ (in K)	BEI at 3.9 Ga	BEI at 2.0 Ga	BEI at modern Earth
7000	3520	2.75	0.367
6250	1680	10.8	0.524
Sun	611	41.3	1.00
5500	482	107	0.678
5000	181	56.5	0.671
4250	26.2	1.95	0.105
3600	44.5	1.13	3.17×10^{-2}
3200	27.8	1.26	2.45×10^{-2}
2400	2.33	0.182	3.93×10^{-3}

Notes: In the third row, $T_\star \approx 5780$ K for the Sun. The choice of 3.9 Ga (Gyr ago) corresponds to the Archean Earth dominated by prebiotic chemistry. The choice of 2.0 Ga reflects a planet where the buildup of atmospheric O_2 (and ozone) has been initiated, but the overall levels are still a few orders of magnitude below modern-day levels. (*Data source:* S. Rugheimer, A. Segura, L. Kaltenegger and D. Sasselov [2015], UV surface environment of Earth-like planets orbiting FGKM stars through geological evolution, *Astrophys. J.* 806[1]: 137.)

Other factors that can influence the BEI apart from the age of the star and the presence of ozone are the stellar activity and the atmospheric column density. Neither of these factors is unexpected, given that more active stars produce higher fluxes of UV radiation, thus resulting in enhanced values of the BEI. Similarly, when one considers planets with rarefied atmospheres (e.g., Mars), the amount of UV radiation reaching the surface is elevated, thereby engendering a higher value of the BEI. Yet, it is important to recognize, in light of our preceding discussion, that an appropriate combination of evolutionary adaptations and ecological niches may ensure that many planets ought to be theoretically capable of sustaining life.

Lastly, we have remarked that UV radiation contributes to DNA damage. Consequently, UV radiation is anticipated to have served as a selective agent in favoring those microbes that were capable of undertaking DNA repair. In the same spirit, it has been argued that sexual reproduction emerged as a response to UV-induced DNA damage, because it enables the removal of deleterious mutations, as discussed in Section 3.10. Thus, the mutations caused by UV radiation—either induced via direct exposure or during the processes of DNA repair—may have driven molecular evolution, possibly resulting in higher rates (Rothschild 1999). In addition, some empirical evidence, albeit tentative, suggests that higher rates of molecular

evolution and speciation (formation of new species) are correlated with the amount of energy (e.g., UV radiation) accessible to the environment (K. L. Evans & Gaston 2005).

On the basis of these arguments, it is instructive to build a simple model. Let us operate under the premise that the flux of UV radiation received by an Earth analog dictates the rate of speciation, only provided that all other factors are held equal. We saw in (4.37) that the number of species is characterized by a typical timescale (τ_c). Therefore, we could interpret τ_c as the characteristic timescale for speciation, implying that its inverse would be proportional to the flux of UV radiation in our model. In this event, τ_c will no longer be constant but will depend on stellar properties. As the stellar mass serves as our proxy, by analogy with (4.52), we have

$$\tau_c \sim \tau_{c,\oplus} \left(\frac{M_\star}{M_\odot} \right)^{-3} \qquad M_\star \lesssim M_\odot,$$

$$\tau_c \sim \tau_{c,\oplus} \left(\frac{M_\star}{M_\odot} \right)^{-1} \qquad M_\star \gtrsim M_\odot, \tag{4.56}$$

where $\tau_{c,\oplus} = 163$ Myr for the Earth. By substituting (4.56) into (4.37), we can estimate the peak number of species that is theoretically possible under the assumption of continuous exponential growth. We have plotted this function in Figure 4.6, from which it is readily verified that the maximum occurs at $M_\star \approx M_\odot$. Although the exact position of this peak is partly attributable to our functional choices for t_H and τ_c, this plot nevertheless indicates that Sunlike stars might have a higher tendency to host planets with complex biospheres and high biodiversity. Similarly, as per Figure 4.6, planets around low-mass stars may have a lower likelihood of sustaining diverse biospheres.

4.3.5 Photosynthesis

Life on Earth is dominated by photosynthesis as far as its biomass is concerned. Of the many variants of photosynthesis, oxygenic photosynthesis has proven to be the most prominent. It is reliant on splitting water molecules (which are commonly available) to yield oxygen, and the net reaction can be expressed as

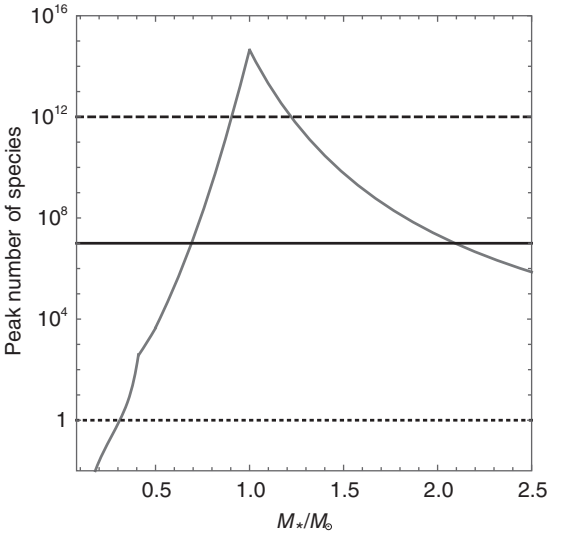


Figure 4.6 The peak number of species $\left(N_{\rm peak}\right)$ attainable as a function of stellar mass $M_\star$ (in units of $M_\odot$). The black solid and dashed lines represent the presumed *current* number of eukaryotic and microbial species on Earth. The black dotted line denotes the regime where the peak number of species drops below unity, suggesting that abiogenesis is not feasible on such worlds. (© Manasvi Lingam and Avi Loeb.)

$$\mathrm{CO_2} + 2\mathrm{H_2O} + h\nu \xrightarrow{\text{pigments}} \mathrm{CH_2O} + \mathrm{H_2O} + \mathrm{O_2}, \tag{4.57}$$

where the presence of $\mathrm{H_2O}$ on both the left- and right-hand sides underscores the fact that water constitutes both reactant and product. The evolution of oxygenic photosynthesis was a major evolutionary breakthrough that transformed the subsequent trajectory of geochemical and biological evolution on our planet. Moreover, the majority of carbon fixation (synthesis of organic compounds) on Earth—expressed via a quantity known as the net primary productivity (NPP)—takes place via oxygenic photosynthesis. Hence, our subsequent analysis will be wholly centered on oxygenic photosynthesis, unless stated otherwise.

4.3.5.1 The range of photosynthetically active radiation

On Earth, most organisms that employ oxygenic photosynthesis rely on photons in the wavelength range of $400 < \lambda < 700$ nm. However, recent

experiments investigating the pigment chlorophyll f have confirmed that cyanobacteria can utilize photons with $\lambda \sim 750$ nm for the purpose of carrying out oxygenic photosynthesis (Nürnberg et al. 2018). Hence, the range of photosynthetically active radiation (PAR) for oxygenic photosynthesis on Earth is roughly 400 to 750 nm. This empirical fact immediately raises the question of whether the same range is applicable to oxygenic photosynthesis on other worlds (Kiang, Segura, et al. 2007).

It must be admitted that we do not have a clear understanding of the molecular machinery that could evolve on other worlds for the purposes of harvesting light energy and converting it into chemical energy. If we suppose that the photosystems—discussed in Sections 3.2 and 3.3—resemble those on Earth, the lower bound can be estimated. It has been shown that ultraviolet photons, characterized by $\lambda < 400$ nm, are responsible for causing ionization and cellular damage (especially to photosystem II), consequently inhibiting the rate and efficiency of photosynthesis. As a result, it is conventional to adopt the lower bound for PAR as $\lambda_{\rm min} = 400$ nm, although this limit may be lowered by 50 nm or thereabouts for some organisms.

The upper bound for PAR is subject to even more uncertainty. In Section 3.3, we described how the splitting of water takes place via the water-oxidizing complex (WOC) in photosystem II. An essential feature of the WOC is that two photons are consumed for the oxidation of a single water molecule. Each of these two photons has a wavelength of ~ 700 nm, implying that the total energy is $E_{\rm tot} \sim 3.54$ eV. Instead, let us suppose that an alternative mechanism exists in which χ photons are consumed to split one water molecule. The corresponding wavelength $\lambda_{\rm max}$ is given by

$$\lambda_{\rm max} \sim 350\, \chi \text{ nm}. \tag{4.58}$$

Thus, it may be theoretically possible for the evolution of oxygenic photosynthesis that entails three photons with $\lambda_{\rm max} \sim 1.05$ μm, or four photons with $\lambda_{\rm max} \sim 1.4$ μm, and so on (Wolstencroft & Raven 2002). At first glimpse, it appears as though the upper bound on PAR is arbitrary; however, physics may offer a way out of this conundrum.

The spectral radiance B_λ for a blackbody is known to be

$$B_\lambda = \frac{2hc^2}{\lambda^5} \left[\exp\left(\frac{hc}{\lambda k_B T_\star} \right) - 1 \right]^{-1}, \tag{4.59}$$

when expressed in terms of the wavelength λ. The spectral photon flux n_λ is obtained from dividing B_λ by the photon energy hc/λ, which yields

$$n_\lambda = \frac{2c}{\lambda^4}\left[\exp\left(\frac{hc}{\lambda k_B T_\star}\right) - 1\right]^{-1}. \tag{4.60}$$

If we determine the maximal value for n_λ using $dn_\lambda/d\lambda = 0$, we end up with

$$\lambda_{\rm peak} \approx 635\,{\rm nm}\left(\frac{T_\star}{T_\odot}\right), \tag{4.61}$$

where $T_\odot \approx 5780$ K is the effective temperature of the Sun. On Earth, the peak absorbance of photosynthetic pigments is typically around 680 nm, which is not far removed from the estimate of 635 nm obtained from (4.61). In this calculation, we have ignored the role of the planetary atmosphere. Due to the buildup of atmospheric oxygen, it has been established that the peak of the incident spectral photon flux was shifted to ~ 685 nm, which is very close to the peak absorbance wavelength documented for photosynthetic pigments on our planet (Kiang, Siefert, et al. 2007). Thus, to leading order, it is reasonable to surmise that the absorbances of photosynthetic pigments are optimized to attain their peaks at the wavelengths where the maximum value of the spectral photon flux occurs.

A more sophisticated approach along these lines was presented by Björn (1976, 2015), who argued that the optimal frequency is achieved when the power output of the photosynthetic apparatus is maximized. By employing this criterion, the power output was estimated to be

$$P(\nu) = \mu_{\rm eff}(\nu) \times \nu^2 \exp\left(-\frac{h\nu}{k_B T_\star}\right), \tag{4.62}$$

where $\mu_{\rm eff}(\nu)$ denotes the effective potential with the following definition:

$$\mu_{\rm eff} = \frac{\mu}{1 + k_B T_{\rm Chl}/\mu}, \tag{4.63}$$

where $T_{\rm Chl} \sim 300$ K is the temperature of the pigment and μ is given by

$$\mu = \mu_c + k_B T_{\rm Chl} \ln\left(\frac{k_B T_{\rm Chl}}{\mu_c + k_B T_{\rm Chl}}\right), \tag{4.64}$$

with μ_c defined as

$$\mu_c = k_B T_{\rm Chl} \ln\left(\frac{R_\star^2}{12a^2}\right) + h\nu\left(1 - \frac{T_{\rm Chl}}{T_\star}\right). \tag{4.65}$$

We leave it as an exercise for the reader to calculate the frequency (and the corresponding wavelength) at which the maximum value of P occurs, by making use of the criterion $dP/d\nu = 0$ in conjunction with (4.62)–(4.65). It can be shown, for instance, that choosing $T_\star \sim 3000$ K yields a wavelength of ~ 1.3 μm for the peak power.

Although these models yield some interesting results, they are oriented toward finding the optimal wavelength. Hence, estimating this wavelength is *not* the same as determining the upper bound for PAR. A number of different hypotheses abound in the literature for the latter, but none of them are considered definitive (Kiang et al. 2007). The amount of energy required to split one molecule of water is 1.23 eV, which would naively suggest that a photon with wavelength $\gtrsim 1$ μm is unsuitable for photosynthesis. In actuality, however, the photon(s) is not directly employed for splitting the water molecule; instead, it takes part in a complex biochemical pathway outlined in Section 3.3. Another criteria often regarded as a prerequisite for photosynthesis is that the photons should be capable of engendering electronic excitations, and current data suggest that photons with wavelengths $\gtrsim$ 1.1 μm must be excluded. Finally, there have been arguments on thermodynamic grounds that photons with $\lambda > 0.9$ μm are inefficient for carbon fixation due to the high efficiency of backreactions during the night.

Therefore, in light of the many uncertainties involved, we shall henceforth operate under the conservative assumption that the limits for PAR on other worlds are the same as on Earth. This premise is tantamount to selecting $\lambda_{\rm min} = 400$ nm and $\lambda_{\rm max} = 750$ nm.

4.3.5.2 *Fluxes of PAR on Earth-analogs and their implications*

From (4.57), it is apparent that the flux of photons in the PAR range received at the surface of the planet will dictate the NPP of the planet. Let us suppose that the availability of photons constitutes the only limiting factor for photosynthesis. We will focus only on Earth-analogs herein, implying that

our results are solely dependent on the stellar mass; other stellar factors such as the age are not taken into consideration.

The photon flux $F_\star$ (in units of photons $\mathrm{m}^{-2}\ \mathrm{s}^{-1}$) received at the Earth-analog is

$$F_\star = \frac{\dot{N}_\star}{4\pi a_\star^2}, \tag{4.66}$$

where $a_\star$ can be calculated by using (4.5).[9] In this formula, the photon production rate $\dot{N}_\star$ for the star is defined as

$$\dot{N}_\star = 4\pi R_\star^2 \times \Psi_B \times \int_{\lambda_{\min}}^{\lambda_{\max}} n_\lambda \, d\lambda, \tag{4.67}$$

with the spectral photon flux n_λ given by (4.60), and the factor $4\pi R_\star^2$ accounts for the area of the star. Note that $\Psi_B \approx 2$ is a fudge factor introduced to account for deviations from the blackbody spectrum and ensure that PAR photon flux at Earth is consistent with observations. The integral (4.67) can be evaluated analytically and substituted into (4.66). What interests us is the ratio $F_\star/F_\oplus$, where $F_\oplus$ represents the critical value of the PAR photon flux received by the planet that is necessary for sustaining a NPP equal to that of the Earth. The value of $F_\oplus$ is extracted from the fraction of photons in the PAR range that are employed for photosynthesis on Earth; it has been estimated that $F_\oplus \approx 4 \times 10^{20}$ photons $\mathrm{m}^{-2}\ \mathrm{s}^{-1}$ (Lehmer et al. 2018).

With this data, we have plotted $F_\star/F_\oplus$ as a function of the stellar mass in the left panel of Figure 4.7. To begin with, it is important to recognize that $F_\star/F_\oplus \approx 2.5$ for $M_\star = M_\odot$; therefore, this ratio does not equal unity for the Earth. This result is a consequence of the fact that the NPP of the Earth is not limited by the availability of PAR but by the concentration of nutrients. Second, it can be seen that Earth-analogs around stars with $M_\star \lesssim 0.21 M_\odot$ do not satisfy the criterion $F_\star/F_\oplus > 1$. As a consequence, Earth-analogs orbiting late-type M-dwarfs are expected to have a lower likelihood of sustaining Earthlike biospheres insofar as NPP is concerned.

In our analysis, we have assumed that the star can be effectively modeled as a blackbody. This assumption is not accurate for M-dwarfs because a sizable fraction of them (up to two-thirds) are characterized by frequent stellar

9. We also utilize the formula $T_\star \propto M_\star^{0.35}$, which follows by combining the mass-radius ($R_\star \propto M_\star^{0.8}$) and mass-luminosity ($L_\star \propto M_\star^3$) scaling relations with (4.2).

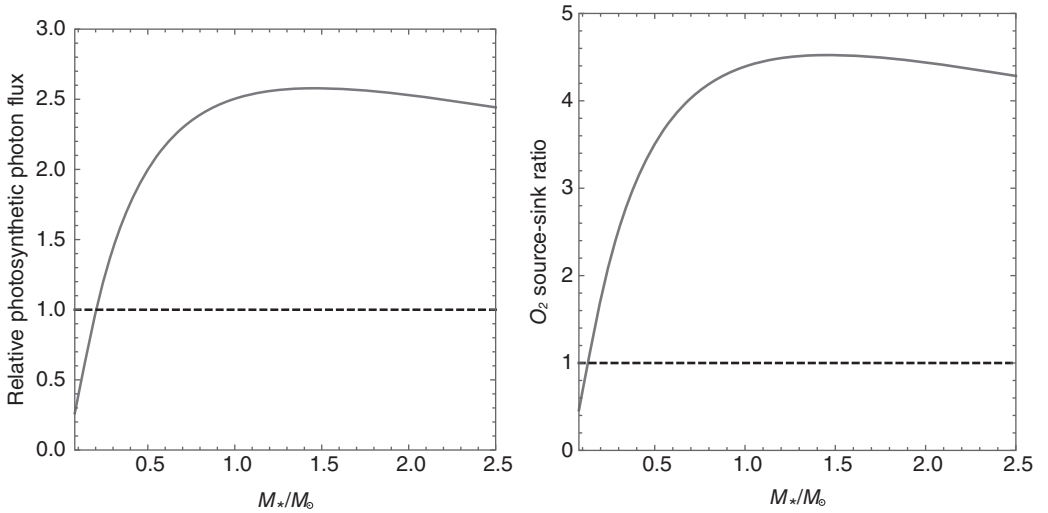


Figure 4.7 *Left:* The photon flux in the wavelength range of 400 to 750 nm received by a habitable Earth-analog as a function of the stellar mass (in units of $M_{\odot}$). The horizontal dashed line denotes the minimum flux that is ostensibly necessary for ensuring that the NPP of the biosphere is close to that of the Earth. *Right:* The maximal oxygen source-to-sink ratio (Δ_{O_2}) as a function of the stellar mass ($M_{\star}$) provided that the PAR flux serves as the sole limitation. When it comes to Earth-analogs around stars that fall below the dashed line (denoting the balance of the O_2 source and sink), oxygen depletion becomes dominant. (© Manasvi Lingam and Avi Loeb.)

flares and associated space weather phenomena. It is not easy to formulate a general expression for $F_{\star}$ that encompasses the contributions from flares, since the flaring activity has been documented to vary by orders of magnitude even for stars with the same stellar mass. Despite this caveat, a couple of broad statements are feasible regarding the impact of flares on $F_{\star}$. First, in the case of quiescent stars, the contribution to $F_{\star}$ arising from flares is negligible compared to the blackbody spectrum because the time-averaged photon production rate from flares will be typically much lower than the bolometric luminosity (Lingam & Loeb 2019c). Second, when it comes to the highly active M-dwarfs, the contribution from flares could enhance $F_{\star}$ by as much as one order of magnitude (Mullan & Bais 2018). The fraction of such stars is not precisely known, but it might be on the order of 10 percent. Therefore, our results are probably more valid for comparatively quiescent stars.

Now, let us turn our attention to the ramifications that stem from the buildup of atmospheric O_2. Oxygenic photosynthesis not only enables the synthesis of organic compounds but also yields oxygen as a product. On Earth, oxygenic photosynthesis functions *indirectly* as a major source of atmospheric oxygen when organic matter is buried. The latter is responsible for ensuring that backreactions, such as respiration and oxidative decay of organic matter, do not deplete oxygen from the atmosphere. If there was no organic burial, the oxygen produced by photosynthesis would be consumed fully by the aforementioned sources. Another important source of O_2 is based on the burial of pyrite (FeS_2) as it would otherwise react with oxygen and reduce its abundance, which is not considered herein.

Let us consider the amount of O_2 gained from organic burial by adopting the methodology presented in Lehmer et al. (2018). On Earth, the production rate of organic compounds (in the form of CH_2O) has been estimated to be around 3.5×10^{15} mol/yr. As remarked earlier, only a small fraction of the total organic carbon is buried, with the upper bound on the burial efficiency being 2.9×10^{-3}. By multiplying this factor with the rate of synthesis of organic matter, we can arrive at an estimate for the rate of O_2 supplied to the atmosphere, which is found to be 10^{13} mol/yr. In Section 3.3, it was noted that eight photons are required per CO_2 fixed (for ideal conditions). Now, under the assumption that the photosynthetic apparatus is similar on other planets, it follows that the production rate of atmospheric O_2 is proportional to the photon flux, since the latter dictates the rate of carbon fixation. With this choice of ansatz, we employ

$$\mathcal{S}_{O_2} \sim 10^{13}\,\text{mol/yr}\,\frac{F_\star}{F_\oplus}, \tag{4.68}$$

where $\mathcal{S}_{O_2}$ denotes the production rate of O_2. There are a couple of major O_2 sinks, of which only one of them is anticipated to be prominent in a predominantly anoxic world. This sink is based on the depletion of O_2 via rapid reactions with reducing gases released due to volcanism and submarine weathering of the ocean floor. It is essential to recognize that the rate of oxygen removal is a time-dependent quantity and will vary across different planets in accordance with their geological properties (e.g., internal heat budget). As we have no information about any of these variables, we shall suppose that the depletion flux is constant and equal to that of modern

Earth. In this event, we end up with

$$\mathcal{L}_{O_2} \sim 5.7 \times 10^{12}\,\mathrm{mol/yr}, \tag{4.69}$$

with $\mathcal{L}_{O_2}$ denoting the O_2 depletion rate. In order for O_2 to build up in the atmosphere, the source must dominate over the sink, i.e., we require $\mathcal{S}_{O_2} > \mathcal{L}_{O_2}$. It is therefore instructive to define the source-to-sink ratio, $\Delta_{O_2} = \mathcal{S}_{O_2}/\mathcal{L}_{O_2}$.

A crucial point worth appreciating is that Δ_{O_2} represents the maximum possible value of the source-to-sink ratio, since it presupposes that the entire PAR flux is utilized by photosynthesis and that no other limitations are prevalent. We have plotted Δ_{O_2} as a function of the stellar mass in the right panel of Figure 4.7, where we find that the condition $\Delta_{O_2} < 1$ is satisfied when $M_\star \lesssim 0.13\,M_\odot$. To put it differently, Earth-analogs orbiting M-dwarfs with masses below this threshold—which translates to a sizable fraction of all M-dwarfs and includes the local examples of Proxima Centauri and TRAPPIST-1—have a low probability of accumulating oxygen in the atmosphere via oxygenic photosynthesis. It is also evident from the left panel of Figure 4.7 that such worlds are unlikely to host biospheres whose productivity is similar to that of the Earth.

This result has two important consequences for the detection of biosignatures and technosignatures. First, if the buildup of atmospheric O_2 is not feasible, it is likely that searches for oxygen would yield *false negatives*. These planets may host biospheres, but their detectability by seeking signatures of O_2 and O_3 is questionable because the concentrations of these gases in the atmosphere would be too low.[10] Second, we analyzed the significance of high atmospheric O_2 levels for the emergence of complex, motile, multicellular life in Section 3.4.1. Although it is overly simplistic to claim that increasing O_2 levels were the sole trigger for the diversification of animals in the Neoproterozoic era, it is not unreasonable to contend that the former was at least partly responsible for enabling the latter. If high oxygen levels can therefore be regarded as a genuine prerequisite for complex life, they will also serve as a precondition for the evolution of technological intelligence.

10. An implicit assumption in this statement is that false positives arising from UV photolysis, as discussed in Section 4.3.1, do not complicate matters further.

In turn, this breakthrough opens up the possibility of finding life through technosignatures.

Our analysis has therefore yielded the following broad conclusions under the explicit and narrow assumption that photon fluxes are the only limiting factor. Quiescent stars with $M_\star \lesssim 0.13 M_\odot$ have a higher likelihood of hosting Earth-analogs with anoxic atmospheres since the buildup of O_2 in the atmosphere via oxygenic photosynthesis may be suppressed. Stars in the range $0.13 M_\odot \lesssim M_\star \lesssim 0.21 M_\odot$ allow for Earth-analogs on which the buildup of atmospheric oxygen is possible, but the timescale for attaining modern Earth levels could be long (possibly exceeding the age of the Universe); for these worlds, the overall NPP is expected to be smaller compared to the Earth. Finally, Earth-analogs around stars with $M_\star \gtrsim 0.21\, M_\odot$ are theoretically capable of sustaining biospheres with the same productivity as that of modern Earth.

In this simplified model, it bears reiterating that we have exclusively dealt with worlds that are exactly akin to the Earth in all respects. One important point to note, however, is that numerous characteristics of aquatic photosynthesis are strongly dependent on the stellar type (Wolstencroft & Raven 2002). For instance, a theoretical model developed by Lingam and Loeb (2020b) suggests that the depth of the euphotic zone (i.e., the layer where photosynthesis could take place) diminishes by more than an order of magnitude as one moves from Earth-analogs around Sunlike stars to those orbiting late-type M-dwarfs. Hence, there is a pressing need to investigate exo-oceanography seriously and advance our understanding of this field.

4.3.6 The oxygenation of the atmosphere

In the preceding paragraphs, we recalled the importance of atmospheric oxygen in the context of facilitating the diversification of complex life. Until now, we have encountered two different pathways by which atmospheric O_2 could be generated. The first involved the *abiotic* buildup of O_2 via UV photolysis in Section 4.3.1, thus potentially leading to very thick O_2 atmospheres. The biological habitability of such worlds remains unresolved, as not enough studies have been undertaken in this subject. The second avenue, tackled in Section 4.3.5, relied on the evolutionary innovation of oxygenic photosynthesis that eventually leads to the rise in atmospheric O_2 levels.

In Section 3.4, we described how the Earth's atmosphere transitioned from a predominantly anoxic state to roughly $\lesssim 1$ percent of the present atmospheric level (PAL) around 2.4 Ga. This event was christened the Great Oxidation Event (GOE). Numerous mechanisms to explain its causation and timing have been proposed, some of which were reviewed in Section 3.4. One of the more well-known hypotheses entails a combination of both UV photolysis and the emergence of oxygenic photosynthesis, proposed by Catling et al. (2001). The central idea is that organic matter is generated through oxygenic photosynthesis via (4.57), and the resultant compounds are subsequently decomposed by other microbes to release methane (CH_4). Methane undergoes UV photolysis with hydrogen escaping to space and the carbon combining with O_2 to form CO_2. If all of the appropriate chemical reactions are summed up, the net effect is

$$2H_2O + \text{“biosphere”} + h\nu \rightarrow O_2 + 4H\,(\uparrow \text{space}). \tag{4.70}$$

The important point to recognize here is that the above reaction is *not* the same as the abiotic photolysis of H_2O because the existence of a microbial biosphere is necessary for mediating some of these reactions.

As a result, it is plausible that the characteristic timescale for the atmospheric O_2 levels to reach a certain value (e.g., 1 percent PAL) via (4.70) may be smaller on planets with high UV fluxes in the appropriate wavelength range. In order to quantify these fluxes, it is reasonable to use the Lyα flux (Φ_L) as a proxy, since it drives a large fraction of the total water and methane photolysis. One can therefore work out Φ_L for Earth-analogs by using the empirical data for the Lyα fluxes. This analysis was carried out by Lingam and Loeb (2018d), who introduced the ansatz: $\Phi_L \propto M_\star^\kappa$, where $\kappa \approx -2.3$ for $M_\star \lesssim M_\odot$ and $\kappa \approx 3.3$ for $M_\star \gtrsim M_\odot$. For stars more massive than the Sun, the flux received by Earth-analogs is higher because a larger fraction of the blackbody radiation is emitted as UV light. For Earth-analogs around stars less massive than the Sun, especially M-dwarfs, we have documented how the dual effects of emission from the outer layers of the star and closer orbital radii collectively enhance the far-UV fluxes in Section 4.3.1.

Next, we shall adopt the prescription that the timescale for the oxygen levels to increase (t_I) is inversely proportional to Φ_L, along the lines of our

approach in Section 4.3.2. Thus, we end up with

$$t_I \sim t_{I,\oplus} \left(\frac{M_\star}{M_\odot}\right)^{2.3} \qquad M_\star \lesssim M_\odot,$$
$$t_I \sim t_{I,\oplus} \left(\frac{M_\star}{M_\odot}\right)^{-3.3} \qquad M_\star \gtrsim M_\odot, \tag{4.71}$$

where $t_{I,\oplus}$ represents the characteristic timescale for the initiation of the GOE for the Earth *after* oxygenic photosynthesis has been invented; the latter constraint must be included since oxygenic photosynthesis is a prerequisite for (4.70). On the basis of our analysis in Chapter 3, we adopt ~ 2.7 Ga for the evolution of oxygenic photosynthesis and ~ 2.4 Ga for the GOE on Earth; the choice of this timeline leads us to $t_{I,\oplus} \sim 0.3$ Gyr.

We are now in a position to formulate the overall time required for the atmospheric oxygen to attain post-GOE (i.e., shortly after the GOE) levels on Earth-analogs around other stars. Let us denote this timescale by t_{O_2} and decompose it into three distinct components: (1) the timescale for the origin of life (t_0), (2) the time required for oxygenic photosynthesis to arise subsequent to abiogenesis (t_{OP}), and (3) the timescale for the oxygenation of the atmosphere via UV photolysis after the emergence of oxygenic photosynthesis (t_I). Of this trio, we have already documented the expressions for (1) and (3) in (4.52) and (4.71), respectively. This leaves us with t_{OP}, the temporal duration between two major evolutionary breakthroughs, which remains unknown. Nevertheless, for the sake of argument, suppose that the rate of molecular evolution is dictated by the UV flux at the surface, as proposed in Section 4.3.4. In this event, adopting the functional dependence on $M_\star$ from (4.56), we have

$$t_{\mathrm{OP}} \sim t_{\mathrm{OP},\oplus} \left(\frac{M_\star}{M_\odot}\right)^{-3} \qquad M_\star \lesssim M_\odot,$$
$$t_{\mathrm{OP}} \sim t_{\mathrm{OP},\oplus} \left(\frac{M_\star}{M_\odot}\right)^{-1} \qquad M_\star \gtrsim M_\odot, \tag{4.72}$$

where $t_{\mathrm{OP},\oplus} \sim 1$ Gyr, obtained by utilizing the putative evolutionary timeline adumbrated in Chapter 3, which posited that life originated $\gtrsim 3.7$ Ga and oxygenic photosynthesis arose ~ 2.7 Ga.

Therefore, by summing up the three components of t_{O_2} introduced above, we arrive at the following expression:

$$t_{O_2} \sim 1.8\,\mathrm{Gyr}\left(\frac{M_\star}{M_\odot}\right)^{-3} + 0.3\,\mathrm{Gyr}\left(\frac{M_\star}{M_\odot}\right)^{2.3} \qquad M_\star \lesssim M_\odot,$$
$$t_{O_2} \sim 1.8\,\mathrm{Gyr}\left(\frac{M_\star}{M_\odot}\right)^{-1} + 0.3\,\mathrm{Gyr}\left(\frac{M_\star}{M_\odot}\right)^{-3.3} \qquad M_\star \gtrsim M_\odot. \tag{4.73}$$

We have plotted t_{O_2} as a function of the stellar mass $M_\star$ in the left panel of Figure 4.8. The trend is along expected lines, i.e., more massive stars display shorter timescales because of the enhanced UV fluxes incident on their Earth-analogs. It is also possible to calculate the function t_H/t_{O_2} in order to determine which stars can give rise to sufficiently oxygenated atmospheres on Earth-analogs before their duration of habitability is truncated. This variable has been plotted in the right panel of Figure 4.8. We find that $t_H/t_{O_2} < 1$ occurs when $M_\star \lesssim 0.57\,M_\odot$, indicating that stars in the range may not have Earth-analogs with biogenic oxygen levels akin to those on Earth after the GOE. We also observe that the peak occurs at $M_\star \approx M_\odot$, possibly implying that G-type stars have a relatively higher likelihood of hosting Earth-analogs with oxygenated atmospheres. It goes without saying that a multitude of geological, chemical, and biological factors dictate the necessary time for oxygenation, which are not incorporated in our crude model.

4.3.7 Exergy of stellar radiation

Until now, we have primarily dealt with energy fluxes arising from stellar radiation, winds, or Joule heating. This focus is merited since the *quantity* of energy accessible to a planet will influence its propensity for hosting biospheres. However, another issue of import is the *quality* of energy, which can be heuristically envisioned as the efficiency at which the supplied energy is utilizable by the planet. It is therefore appropriate to introduce the notion of *exergy* at this stage. In qualitative terms, exergy represents the maximum amount of work that could be extracted from a system when it is brought into equilibrium with its surroundings. The interested reader may check out Dincer and Rosen (2013) for an overview of this subject. Our analysis will closely parallel the recent analysis of this topic by Scharf (2019).

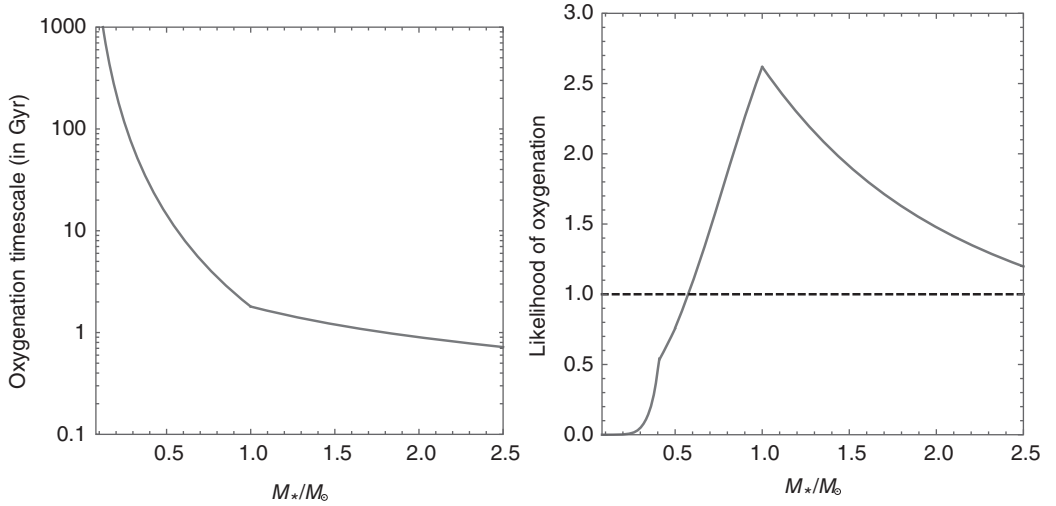


Figure 4.8 *Left:* The timescale (in Gyr) for an Earth-analog to attain atmospheric oxygen levels comparable to that of post-GOE Earth as a function of the stellar mass $M_\star$. *Right:* The likelihood of Earth-analogs to achieve oxygen levels similar to post-GOE Earth as a function of the stellar mass. The likelihood function (γ-axis) is defined as the ratio t_H/t_{O_2}, where t_H and t_{O_2} are given by (4.36) and (4.73), respectively. The black dashed line denotes the regime where the likelihood is below unity, suggesting that oxygenation may not be feasible on such worlds. (© Manasvi Lingam and Avi Loeb.)

Quantitatively speaking, the exergy (E_x) is defined as

$$E_x = U_0 + PV_0 - TS_0 - \sum_i \mu_{0i} M_i, \tag{4.74}$$

where U_0, V_0, S_0, and μ_{0i} are the internal energy, volume, entropy, and chemical potential (of species i) of the system; M_i denotes the number of moles of species i; and P and T are the pressure and temperature of the surroundings. This formula shares a close resemblance to the Gibbs free energy, but the exergy is dependent on both internal (system) and external (surroundings) parameters.

The exergy for a blackbody (integrated over all wavelengths) per unit area at temperature T_0 immersed within a blackbody radiation field at effective temperature $T_{\rm eff}$ can be calculated. The details can be found in Candau (2003), with the final expression given by

$$E_x = \sigma \left(T_{\rm eff}^4 - \frac{4}{3} T_{\rm eff}^3 T_0 + \frac{1}{3} T_0^4 \right), \tag{4.75}$$

and it is easy to verify that the exergy reduces to the standard expression σT_{eff}^4 in the limit $T_0 \to 0$. The efficiency (denoted by $\mathcal{W}$) associated with the conversion of radiation to work over the whole spectrum is defined as

$$\mathcal{W} = \frac{E_x}{\int B_\lambda \, d\lambda} = 1 - \frac{4}{3}\left(\frac{T_0}{T_{\text{eff}}}\right) + \frac{1}{3}\left(\frac{T_0}{T_{\text{eff}}}\right)^4. \tag{4.76}$$

The Carnot efficiency ($\mathcal{W}_C$) is conventionally, but not always correctly, regarded as the upper bound on the efficiency attainable by a thermodynamic system. In our case, it has the definition

$$\mathcal{W}_C = 1 - \frac{T_0}{T_{\text{eff}}}. \tag{4.77}$$

It is straightforward to verify that $\mathcal{W}_C > \mathcal{W}$ using (4.76) and (4.77) in concurrence with $T_0/T_{\text{eff}} < 1$.

Now, let us introduce the assumption that the Earth-analog and the host star are blackbodies, which is apparently reasonable to leading order. In this event, we choose $T_0 \approx 255$ K (the effective temperature of the Earth) and $T_{\text{eff}} = T_\star$. Alternatively, if we treat the putative biosphere as our system and model it as a blackbody, we must work with $T_0 \approx 288$ K (average surface temperature of the Earth). Before proceeding further, we observe that the analog of (4.76) can be generalized from the whole-spectrum case to include a specific range of wavelengths (e.g., for photosynthesis). The resultant expression is analytical, but we shall not reproduce it here because of its complexity (involving polylogarithmic functions).

In Figure 4.9, we have plotted the functions $\mathcal{W}$ and $\mathcal{W}_C$ as a function of the stellar temperature $T_\star$. First, along expected lines, it is observed that $\mathcal{W} < \mathcal{W}_C$ is always valid. Second, it is found that both $\mathcal{W}$ and $\mathcal{W}_C$ increase with the stellar temperature, thus implying that hotter (and typically more massive) stars enable higher efficiencies with regard to converting the stellar radiation into work. For TRAPPIST-1, choosing $T_\star = 2511$ K yields $\mathcal{W} \approx 0.865$ and $\mathcal{W}_C \approx 0.898$. In the case of the Sun, we find $\mathcal{W} \approx 0.941$ and $\mathcal{W}_C \approx 0.956$ after using $T_\star = 5780$ K. It is therefore apparent that the difference in efficiencies between TRAPPIST-1 and the Sun is around 7.6 percent. Although this variation appears to be very small when viewed in isolation, when multiplied by the energy input to the planets, it could give rise to nontrivial effects. In conclusion, our analysis demonstrates that habitable planets around low-mass stars are characterized by lower photon conversion efficiencies. In turn, this might lower their prospects for

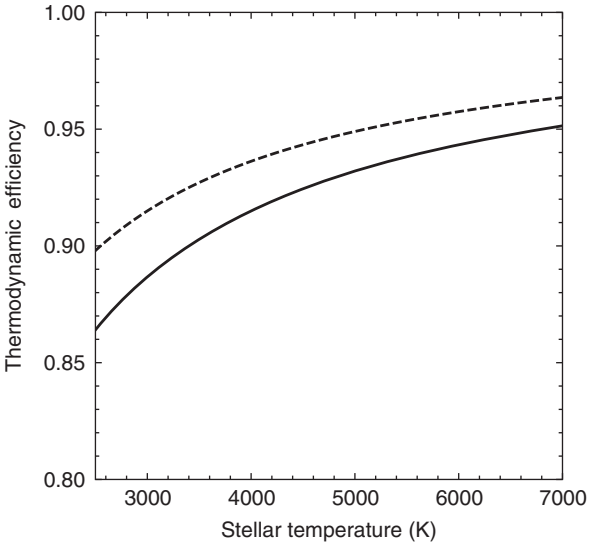


Figure 4.9 The efficiency $\mathcal{W}$ of converting radiation to work (solid curve) and the Carnot efficiency $\mathcal{W}_C$ (dashed curve) as functions of the effective stellar temperature (K). In applying the formulae (4.76) and (4.77), we have chosen $T_0 \approx 255$ K. (© Manasvi Lingam and Avi Loeb.)

habitability to some degree, although more research is necessary in order to assess the veracity of this hypothesis.

4.4 STELLAR FLARES AND ASSOCIATED SPACE WEATHER PHENOMENA

Stellar flares are explosive phenomena that occur close to the surface of the star, resulting in the rapid release of multiple forms of energy as well as energetic charged particles. A comprehensive overview of this subject, from both the observational and theoretical perspectives, was expounded in Priest (2014). The underlying mechanism behind this explosive emission of energy is suspected to be magnetic reconnection. The latter entails alterations in magnetic topology via the breaking and reconnection of field lines, consequently facilitating the rapid conversion of magnetic energy into other variants of energy. Classical models of magnetic reconnection, which were first introduced in the 1950s, predicted that energy release during flares occurred slowly on timescales of $\sim 10^6$ s, in contradiction with observations which revealed eruptions of $\sim 10^3$ s. However, recent theories of magnetic

reconnection have begun to bypass this conundrum. For instance, the disruption of elongated current sheets due to a potent instability (dubbed the plasmoid instability) results in energy conversion over timescales that are commensurate with observations (Comisso et al. 2016, 2017).

The biggest solar flares documented in modern history are characterized by energies of $\sim 10^{25}$ J, with the most prominent one on record being the Carrington event from 1859, which released a total energy of order 10^{25} J. However, this value is not the upper bound: flares with energies $\geq 10^{26}$ J (known as superflares) have been documented for other stars and have been explained via theoretical models. The *Kepler* and *TESS* missions have collectively recorded much observational data regarding the statistics of superflares (Maehara et al. 2015; Davenport 2016; W. S. Howard et al. 2019; Günther et al. 2020). In general, the occurrence rate ($\dot{N}_f$) of flares is expressible as

$$\frac{d\dot{N}_f}{dE_f} \propto E_f^{-\Upsilon}, \tag{4.78}$$

where E_f denotes the total energy of a given flare; $\dot{N}_f$ is a function of E_f. The power-law exponent does not exhibit significant variation across different stars or the bolometric flare energy, and it generally obeys $\Upsilon \approx 1.5$–2. On the other hand, the constant of proportionality in (4.78) is sensitive to myriad stellar properties.

It is instructive to contemplate a simple model to determine the maximum amount of magnetic energy that can be converted into flare energy. In order to do so, it is important to recognize that the stellar flares occur at *active* regions in the vicinity of sunspots. Thus, to leading order, we express the area of the flaring sites in terms of the fraction f_A of the star covered in active regions. The total magnetic energy released from the active region is represented as

$$E_{\text{mag}} = \frac{B_A^2}{2\mu_0}\mathcal{V}_A, \tag{4.79}$$

where B_A and $\mathcal{V}_A$ represent the magnetic field strength and volume of the active region(s). In dimensional terms, we know that $\mathcal{V}_A \sim (\text{Area})^{3/2}$ and the area of the active region(s) is given by $f_A \times 4\pi R_\star^2$; the latter follows from noting that $4\pi R_\star^2$ is the total area and using the definition of f_A from the previous paragraph. Finally, we introduce the efficiency factor ϵ_f for the conversion of magnetic energy into flare energy. Using these relations in

conjunction with (4.79), we obtain

$$E_f \sim 10^{28}\,\mathrm{J}\left(\frac{\epsilon_f}{0.1}\right)\left(\frac{B_A}{0.1\,\mathrm{T}}\right)^2\left(\frac{f_A}{0.3}\right)^{3/2}\left(\frac{R_\star}{R_\odot}\right)^3. \tag{4.80}$$

An important point worth bearing in mind is that we have normalized f_A by the *maximal* fraction encompassed by active regions, whereas B_A and ϵ_f have been normalized by their characteristic values. In principle, the magnetic fields in active regions might reach $B_A \sim 1$ T. In this event, (4.80) predicts that the maximum value of E_f ought to be $\sim 10^{30}$ J, which is potentially consistent with in-depth empirical and theoretical analyses.

For the Sun, it has been suggested that superflares with $E_f \sim 10^{27}$ J occur over a timescale of ~ 2000 yr (Shibayama et al. 2013). Unfortunately, it is not easy to verify this hypothesis owing to the sparsity of direct evidence (Usoskin 2017). The most encouraging evidence to date concerns a spike in the concentration of the radioactive isotope ^{14}C in tree rings dated to 774–775 CE that is about twenty times higher than fluctuations arising from normal solar activity (Miyake et al. 2012). A number of astrophysical phenomena were invoked to explain this observation, of which one of the most promising candidates is that this spike was caused by the eruption of a solar superflare whose energy might have been $\sim 10^{27}$ J. Another spike in the concentration of ^{14}C that dates from 993–994 CE was also interpreted as the outcome of a potential superflare by some scientists. In each of these instances as well as other anomalies, alternative hypotheses exist and the available evidence is riddled with ambiguities.

Henceforth, we will primarily focus our attention on planets in the HZ of M-dwarfs as a subset of these stars are very active and often, albeit *not* always, display flaring rates that are much higher than G-type stars like the modern Sun (Maehara et al. 2012). In this respect, it is worth remarking that TRAPPIST-1 and Proxima Centauri are known to flare regularly. Observations have revealed that the cumulative rate of flares with energies $\gtrsim 10^{26}$ J is $\sim 10^{-2}$ per day; in contrast, the corresponding value for modern Sun is on the order of $\sim 10^{-5}$ per day. Recall that exoplanets in the HZs of these stars will be situated much closer, might possess weak magnetic fields, and might be subject to rapid atmospheric erosion from a combination of strong stellar winds and initially high XUV fluxes.

4.4.1 Electromagnetic radiation

One of the best-known consequences of stellar flares is that they are responsible for producing enhanced fluxes of high-energy electromagnetic radiation. Most studies in this area have been oriented toward quantifying the UV fluxes received by planets in the HZs of their host stars. When it comes to M-dwarfs, some studies have perceived the high UV fluxes as a deterrent for habitability, but it behooves us to bear in mind the varied UV shielding mechanisms outlined in Section 4.3.4. Nevertheless, it is instructive to quantify the UV fluxes in order to comprehend the negative effects on unshielded lifeforms before addressing the positive consequences.

For starters, it is necessary to introduce some nomenclature. The UV light that reaches the surface is conventionally divided into three distinct bands: UV-A (315–400 nm), UV-B (280–315 nm), and UV-C (< 280 nm). The presence of ozone in the atmosphere ensures that most of the UV photons that penetrate to the surface are in the UV-A range, which is responsible for much lower DNA damage (by a factor of ~ 100) with respect to shorter UV wavelengths. Thus, the UV fluxes incident at the surface are heavily influenced by the presence or absence of the ozone layer. For the time being, let us suppose that the atmospheres are mostly devoid of ozone—this premise is justified for reasons that shall be explored in Section 4.4.3.

Segura et al. (2010) carried out a detailed study for a hypothetical Earth-analog orbiting the star AD Leonis (with $M_\star \approx 0.36\, M_\odot$) at a distance of 0.16 AU subjected to a flare with a total energy of $\sim 10^{27}$ J. By self-consistently accounting for the effects of ozone depletion during the flare (see Section 4.4.3), they found that the overall UV flux in the range ~ 200 to 400 nm reaching the surface at the peak of the flare was ~ 50 times higher than the background value. At the peak of the flare, the fluxes of UV-A, UV-B, and UV-C radiation were given by 120.77 W/m^2, 3.15 W/m^2, and 1.93×10^{-14} W/m^2, respectively. These fluxes are only somewhat higher than the normal UV fluxes received on modern Earth: 31.5 W/m^2, 0.798 W/m^2, and 2.30×10^{-16} W/m^2 for UV-A, UV-B, and UV-C radiation, respectively (Rugheimer et al. 2015). Hence, when UV radiation is the sole parameter at play, the biological impact of the flare is anticipated to be minimal.

In a more recent study, Estrela and Valio (2018) investigated the effects of superflares on a hypothetical Earthlike planet at a distance of 1 AU from

the star Kepler-96. The mass of Kepler-96 is virtually identical to the Sun, but its age has been estimated to be ~ 2.34 Gyr. The biologically effective irradiance (BEI), introduced in (4.55), was calculated for three different superflares, of which the biggest had $E_f \sim 1.8 \times 10^{28}$ J. The biologically effective fluence (BEF), denoted by E_{eff}, was calculated by multiplying the BEI with the duration of the flare (about 580 s) for two microorganisms: *Escherichia coli* and *Deinococcus radiodurans*; the latter bacterium has gained renown for being a hardy polyextremophile whose resistance to radiation is truly remarkable.

The critical UV doses for *E. coli* and *D. radiodurans* at ~ 254 nm are 22.6 J/m^2 and 553 J/m^2, respectively, where the critical threshold demarcates the limit at which only 10 percent of the bacteria survive. In the absence of ozone, for the superflare mentioned above, it was found that $E_{\text{eff}} \approx 2 \times 10^4$ J/m^2 for *E. coli* and $E_{\text{eff}} \approx 1.3 \times 10^4$ J/m^2 for *D. radiodurans*. In contrast, in the presence of an ozone layer, E_{eff} was reduced by roughly three orders of magnitude for both these microbes. Even in the absence of an ozone layer, theoretical models based on radiative transfer indicate that microbes akin to *D. radiodurans* and *E. coli* may be able to survive at ocean depths of ~ 12 m and ~ 28 m, respectively.

In mid-2018, the first superflare at naked-eye brightness was reported for Proxima Centauri with a corresponding energy of $E_f \sim 10^{26.5}$ J (W. S. Howard et al. 2018). In the presence of elevated UV radiation, the effects on the putative lifeforms of Proxima b were argued to be profound in the absence of ozone. The UV-B fluence deposited at the surface during the flare was estimated to be $\sim 3.5 \times 10^4$ J/m^2. This fluence lies just below the lethal dose of $\sim 4 \times 10^4$ J/m^2 for *D. radiodurans,* but it would be lethal to the majority of UV-resistant microorganisms on Earth. For example, a dose of $\sim 1.5 \times 10^4$ J/m^2 UV-B radiation will drive approximately 50 percent of freshwater invertebrates (at a depth of $\lesssim 0.5$ m) to extinction. The fluence of UV-C radiation at the surface was predicted to be $\sim 3.6 \times 10^4$ J/m^2, which is a factor of ~ 65 higher than the critical threshold for *D. radiodurans* specified in the previous paragraph.

It is important, however, to recognize that either evolutionary adaptations (such as those evinced by lichens) or suitable geochemical niches (e.g., in the ocean depths) may still permit life to exist on Proxima b. In fact, even the very concept of the critical dose introduced earlier is not robust. Current laboratory experiments suggest instead that the survival curves of microorganisms, when exposed to radiation, are innately biphasic in nature. Hence,

despite being subjected to extremely high doses, a small (but nonnegligible) fraction of microbes might possess the capacity to survive and proliferate. Hence, in contrast to discussion supported by previously cited publications, experiments performed by Abrevaya et al. (2020) and theoretical modeling by J. T. O'Malley-James and Kaltenegger (2019b) bolster the prospects for the survival of Earth-like extremophiles on Proxima b.

Now, let us turn our attention to the positives linked with flares. We documented one of them in some detail in Section 4.3.5, where we pointed out that flares could enhance the time-averaged photon flux of photosynthetically active radiation by up to one order of magnitude for highly active M-dwarfs. As a result, for this category of stars with frequent large flares, one may choose to multiply (4.66) with the phenomenological factor of $(M_\star/M_\odot)^{-1}$, albeit on the basis of admittedly sparse data (Mullan & Bais 2018). The inclusion of this extra factor might ensure that a small fraction of late-type M-dwarfs, if sufficiently active, have the potential to host Earthlike biospheres. The second aspect of flares playing a positive role involves raising the UV fluxes received by the planets.

Section 4.3 elucidated how the incident UV flux at the surface can, in principle, regulate important evolutionary phenomena such as abiogenesis, biodiversity, and the buildup of atmospheric O_2. In each of these instances, the broad conclusion was that M-dwarf exoplanets are limited by the availability of UV photons. The paucity of UV radiation could be ameliorated by the enhanced, albeit transient, radiation fluxes from stellar flares. Consequently, M-dwarf exoplanets may experience intermittent periods of high mutation rates, thereby promoting ecological and evolutionary change during such intervals. This mode of evolution, characterized by hypermutation episodes, might favor rapid speciation that compensates for the low background mutation rates predicted for M-dwarf exoplanets in Section 4.3.4.

We can estimate the UV photon flux arising from flares as follows. To start with, let us suppose that the occurrence rate of flares with total energy E_f is $\dot{N}_f$ (with units of inverse time). Of the total energy, only a fraction f_{UV} is emitted in the wavelength range of interest, the latter of which will vary depending on the specific phenomenon we seek to focus on. In this event, the time-averaged luminosity is given by $\dot{N}_f E_f$, and the UV flux incident on the planetary surface (sans atmospheric ozone) due to flares is given by

$$\mathcal{F}_{UV}^{(f)} = \frac{f_{UV}\dot{N}_f E_f}{4\pi a^2}, \tag{4.81}$$

where $\mathcal{F}_{UV}^{(f)}$ (units of W/m^2) denotes the planetary UV flux contribution from flares. We can further simplify this expression for Earth-analogs by substituting (4.5) for the orbital radius a in (4.81).

In order to comprehend how (4.81) could be gainfully employed, let us consider the case of UV-mediated abiogenesis outlined in Section 4.3.2. Recall that there exists a minimum UV flux that is necessary in order to facilitate crucial prebiotic reactions that require UV light; we will work with (4.51) in particular. Thus, by comparing this equation with (4.81), we can calculate the lower bound on the occurrence rate ($\dot{N}_f$) required to ensure that the critical UV threshold is exceeded. After some rearrangement, we end up with

$$\dot{N}_f \gtrsim 3.8 \times 10^2 \,\mathrm{day}^{-1} \left(\frac{E_f}{10^{27}\,\mathrm{J}}\right)^{-1} \left(\frac{M_\star}{M_\odot}\right)^3, \tag{4.82}$$

where we have used $f_{UV} \approx 6.7 \times 10^{-2}$ under the assumption that the flare behaves akin to a blackbody at temperature 9000 K and considering the wavelength range of $200 < \lambda < 280$ nm (Ducrot et al. 2020). We have expressed $\dot{N}_f$ in day^{-1} to preserve continuity with the units employed in the literature. We leave it to the reader to carry out a similar calculation for f_{UV} over the entire bioactive UV range (200–400 nm) or the XUV range (0.6–120 nm) under the same set of assumptions. On the basis of the data from the *TESS* mission, it was found that the criterion (4.82) was fulfilled for 62 M-dwarfs out of a total sample of 632 M-dwarf flare stars (Günther et al. 2020). The analog of (4.82) is easy to calculate for photosynthesis as well, with the cutoff being set by the number of photons required for the sustenance of an Earthlike biosphere ($F_\oplus$); the PAR range is ~ 400 to 750 nm, based on Section 4.3.5. Using these criteria in conjunction with modeling the flare as a blackbody, the flaring frequency should satisfy

$$\dot{\mathcal{N}}_f \gtrsim 9.4 \times 10^3 \,\mathrm{day}^{-1} \left(\frac{E_f}{10^{27}\,\mathrm{J}}\right)^{-1} \left(\frac{M_\star}{M_\odot}\right)^3. \tag{4.83}$$

As (4.83) is higher than (4.82) by more than one order of magnitude, it is unlikely that most M-dwarfs would fulfill the above condition.

4.4.2 Coronal mass ejections

Coronal mass ejections (CMEs) refer to large structures comprising plasma and magnetic fields that are expelled from stellar coronae. Direct

observations of extrasolar CMEs are challenging, but a few examples exist. By performing high-resolution X-ray spectroscopy of the star HR 9024, Argiroffi et al. (2019) detected Doppler shifted spectral lines and inferred the presence of a large CME with mass $\sim 10^{18}$ kg and kinetic energy of 5×10^{27} J. The causal link between stellar flares and CMEs is not yet precisely resolved, but it can be said that a subset of CMEs are closely associated with stellar flares. The reader should consult D. F. Webb and Howard (2012) for a detailed empirical description of CMEs. An interesting point worth noting about CMEs is that they are responsible for geomagnetic storms that disturb the Earth's magnetosphere. In turn, these storms are capable of disrupting electrical power grids and satellite communications, and it has been estimated that the worldwide financial losses resulting from a large geomagnetic storm could be on the order of ~$1 to $10 trillion (US) (Eastwood et al. 2017).

Most of the typical CMEs tend to carry a total mass of $\lesssim 10^{13}$ kg and travel at velocities of order 100 to 1000 km/s. Among the various phenomena linked with CMEs, one of the most pertinent is that they have been documented to facilitate the acceleration of energetic particles through shock waves, which is tackled in more detail in Section 4.4.3. Here, we shall focus on the consequences of CMEs with regard to their effects on planetary magnetospheres and atmospheric erosion. To begin with, it is instructive to contemplate an empirical scaling between the CME mass ($M_{\rm CME}$) and the total flare energy (E_f) constructed from solar observations (Aarnio et al. 2012):

$$M_{\rm CME} \sim (6.9 \pm 3.1) \times 10^{-2}\, E_f^{0.63 \pm 0.04}, \tag{4.84}$$

where $M_{\rm CME}$ is measured in units of kg and E_f in units of J. It is apparent that the above relation is close to the theoretical prediction of $M_{\rm CME} \propto E_f^{2/3}$ (Takahashi et al. 2016) insofar as the power-law exponent is concerned. We do not yet have a clear picture of when the above scaling breaks down, but it should occur at some stage (Odert et al. 2017). For instance, by drawing on a combination of empirical relations and numerical modeling, Moschou et al. (2019) determined that CMEs linked to the most energetic flares appear to have masses of order 10^{19} kg. The velocity of the CME ($v_{\rm CME}$) is estimated by balancing the kinetic energy of the CME with a fraction of the magnetic energy inherent in the active regions,

$$\frac{1}{2} M_{\rm CME} v_{\rm CME}^2 \sim f_{\rm CME} E_{\rm mag}, \tag{4.85}$$

where f_{CME} denotes the efficiency factor of converting magnetic energy into the kinetic energy of the CME and E_{mag} was defined in (4.79); note that $f_{\mathrm{CME}} \sim 0.1$ is permitted. The maximum kinetic energy of CMEs may be limited by the action of strong large-scale stellar magnetic fields (Alvarado-Gómez et al. 2018; Vida, Leitzinger, et al. 2019).

Let us now consider the parameters for a sizable CME—namely, comparable to the one that impacted the Earth during the famous 1859 Carrington event mentioned earlier. For such a CME, the number density was ostensibly ~ 50 times higher than the current solar wind density near the Earth (8.7×10^6 m^{-3}) while the velocity was ~ 4 times higher than the solar wind velocity at the Earth ($\sim 5 \times 10^5$ m/s), as per simulations by Ngwira et al. (2014). Thus, after substituting the above values into (4.11), the dynamic pressure will be ~ 800 times higher than that of the contemporary solar wind. If we employ this result in (4.13) and suppose that the stellar wind pressure is dominated by the kinetic contribution, we find that the magnetopause distance for an Earth-analog would be compressed to roughly one-third of its steady-state value for modern solar wind conditions.

The most important effect that CMEs have, insofar as our analysis of stellar habitability is concerned, is to enhance the rates of atmospheric escape. For the CME event described above, using (4.14), we see that the mass-loss rate should be enhanced by a factor of ~ 200. In turn, utilizing (4.31) implies that the atmospheric escape rate should increase by the same factor with respect to the normal solar wind conditions. This prediction aligns quite closely with sophisticated MHD simulations that yielded an overall enhancement of ~ 110 (Dong, Huang, et al. 2017). Moreover, data collected by the Mars Atmosphere and Volatile Evolution (MAVEN) mission indicates that comparatively smaller CMEs boost the Martian atmospheric ion escape rates by roughly an order of magnitude.

Although we have focused on planets within our Solar system (or around Sunlike stars), the same considerations are applicable to other exoplanets mutatis mutandis. Kay et al. (2016) carried out numerical simulations and concluded that planets in the HZs of active M-dwarfs are susceptible to ~ 0.5 to 5 major CME impacts per day; broadly similar results were obtained by Kay et al. (2019) for young Solar-type stars. The high frequency of such CME impacts is a direct consequence of the enhanced activity of M-dwarfs in concurrence with the much smaller orbital radii of the planets. Recall from Section 4.2.1 that M-dwarf exoplanets may have

weak magnetic fields, which could further elevate the atmospheric escape rates under certain circumstances.

Several publications have sought to compute the mass-loss rates arising from expulsion of CMEs ($\dot{M}_{\rm CME}$). Loosely speaking, these studies indicate that $\dot{M}_{\rm CME} \gg \dot{M}_\star$ for magnetically active stars, where $\dot{M}_\star$ represents the mass-loss rate attributable to stellar winds. In particular, for young and active Solar-type stars, theoretical models suggest that $\dot{M}_{\rm CME}/\dot{M}_\star \sim 10$ to 100 is feasible (J. J. Drake et al. 2013; Cranmer 2017; Odert et al. 2017). This relation stands in sharp contrast to modern Sun, for which $\dot{M}_{\rm CME} \sim 2 \times 10^{-2} \dot{M}_\odot$. Hence, by applying (4.31) with $\dot{M}_{\rm CME}$ instead of $\dot{M}_\star$, we surmise that atmospheric escape rates for some planets might be enhanced by a few orders of magnitude compared to the situation wherein only stellar winds are considered. As a result, the typical timescale for complete atmospheric erosion would be lowered by the same factor.

In addition, the magnetospheres will experience compression due to the increased dynamic pressure exerted by CMEs, as seen from inspecting (4.13). Hence, it appears reasonable to conclude that CMEs exacerbate the same issues associated with stellar winds in Section 4.2.

4.4.3 Stellar proton events

The majority of energetic stellar flares are accompanied by the emission of stellar / solar energetic particles (SEPs); the corresponding phenomena have been dubbed stellar / solar proton events (SPEs). The physical mechanisms underlying the production of SEPs can be classified into two broad categories: "impulsive" events reliant on magnetic reconnection and "gradual" events that necessitate fast shock waves driven by CMEs (Reames 2013). The latter are often responsible for producing high-fluence SPEs and have therefore been argued to regulate planetary habitability to a greater degree. The kinetic energies of SEPs (mostly protons or electrons) tend to fall within the keV and MeV ranges, but the maximum values for the Sun have been documented to reach a few GeV. Since direct measurements of extrasolar SEP fluences do not currently exist, most studies make use of extrapolations based on empirical scaling relations constructed from solar observations (Youngblood et al. 2017). It is important, however, to appreciate that the resultant predictions may break down at high enough SEP fluences (H. S. Hudson 2015).

It has been roughly half a century since the role of SEPs in the formation of hydrogen and nitrogen oxides was recognized. Two representative reactions leading to the formation of nitric oxide (NO) are as follows (Crutzen et al. 1975):

$$N_2 + e \rightarrow 2N + e \tag{4.86}$$

$$N + O_2 \rightarrow NO + O \tag{4.87}$$

Note that the first reaction is contingent on the availability of energetic electrons (viz., SEPs). Nitrogen oxides, in particular, have been observed to facilitate the depletion of ozone via catalytic reactions, with one such example being

$$NO + O_3 \rightarrow NO_2 + O_2, \tag{4.88}$$

where NO_2 refers to nitrogen dioxide. As a result, high-fluence SPEs are capable of driving the destruction of the ozone layer over time. In turn, the loss of ozone would result in much higher fluxes of UV radiation (especially UV-C) reaching the planetary surface. As we have seen in Section 4.4.1, this may affect surface habitability, although it does not rule out the prospects for life underwater or in other subsurface environments.

Segura et al. (2010) analyzed the effects of a putative SPE accompanying the AD Leonis flare (Section 4.4.1). It was found that the atmosphere ozone depletion on the Earth-analog reached a maximum of 94 percent under the assumption that the planet was unmagnetized. More recently, Howard et al. (2018) investigated the cumulative effect of SPEs impacting Proxima b. By taking the frequency of flares and their correlation with SPEs into consideration, it was argued that more than 90 percent of the ozone layer could be lost within five years. In a similar vein, Tilley et al. (2019) studied the role of repeated SPEs in depleting the ozone layer of an unmagnetized Earth-analog around the active star GJ 1243 (with $M_\star \approx 0.24\, M_\odot$) via numerical simulations. It was concluded that an ozone depletion of 94 percent may occur in ten years.

Collectively, these analyses indicate that near-total ozone depletion might transpire over a comparatively short timescale of $\mathcal{O}\left(10^5\right)$ yr on Earth-analogs near active stars. This timescale varies significantly depending on the flaring frequency, planetary magnetic field, and atmospheric composition. Table 4.4 presents the timescale for ≥ 99 percent ozone depletion as a function of the flare energy and occurrence rate for an Earth-analog around GJ 1243. Two interesting trends can be deduced from the data. First, as long as the frequency of large flares is high (i.e., more than one per month),

Table 4.4 Timescale for near-total ozone depletion as a function of flare frequency and energy

Inter-flare separation (s)	t_{25} (s)	t_{27} (s)
7.20×10^3	3.50×10^7	1.89×10^7
8.64×10^4	7.38×10^7	3.22×10^7
6.05×10^5	2.55×10^8	8.33×10^7
2.63×10^6	2.64×10^{14}	1.35×10^8

Notes: Inter-flare separation refers to the interval between two flares of a given energy. t_{25} and t_{27} are the timescales for achieving $\geq 99\%$ ozone depletion when subjected to flares of energy $\sim 10^{25}$ J and $\sim 10^{27}$ J, respectively. The timescale for ozone depletion is calculated for an Earth-analog at a distance of 0.16 AU. For Earth-analogs around other stars, in order to preserve the same SEP fluences, the flare energies must be roughly scaled by $(a/0.16\,\mathrm{AU})^2$, where a denotes the orbital radius. (*Data source:* Adapted from M. A. Tilley, A. Segura, V. S. Meadows, S. Hawley, and J. Davenport [2019], Modeling repeated M-dwarf flaring at an Earth-like planet in the habitable zone: I Atmospheric effects for an unmagnetized planet," *Astrobiology* 19[1]: 64–86.)

the timescale for ozone depletion does not change much; it does, however, manifest a weak dependence on the flare energy. Second, once a critical threshold is exceeded (which depends on the flare energy), the timescale becomes very sensitive to the flare energy.

Even a single superflare causes substantial, but transient, ozone depletion. By making use of empirical data and theoretical models for the extent of ozone depletion, the associated SEP fluences, and flare energies, Lingam and Loeb (2017e) proposed a phenomenological relation for the degree of ozone depletion ($\mathcal{D}_{O_3}$) as a function of the flare energy and orbital radius:

$$\mathcal{D}_{O_3} \sim 2.8\% \left(\frac{E_f}{10^{25}\,\mathrm{J}}\right)^{9/25} \left(\frac{a}{1\,\mathrm{AU}}\right)^{-18/25}. \tag{4.89}$$

Using this formula, it is straightforward to verify that a superflare of energy $E \sim 2 \times 10^{29}$ J would result in complete ozone depletion ($\mathcal{D}_{O_3} \approx 100\%$) for a terrestrial planet at 1 AU; in contrast, the appropriate value for a planet at 0.1 AU is $E \sim 2 \times 10^{27}$ J. We can further specialize to the case of Earth-analogs and employ (4.5) in (4.89), thus obtaining

$$\mathcal{D}_{O_3} \sim 2.8\% \left(\frac{E_f}{10^{25}\,\mathrm{J}}\right)^{9/25} \left(\frac{M_\star}{M_\odot}\right)^{-27/25}. \tag{4.90}$$

Apart from facilitating ozone depletion, the impact of SEPs with planetary atmospheres may give rise to another problematic issue. Protons that

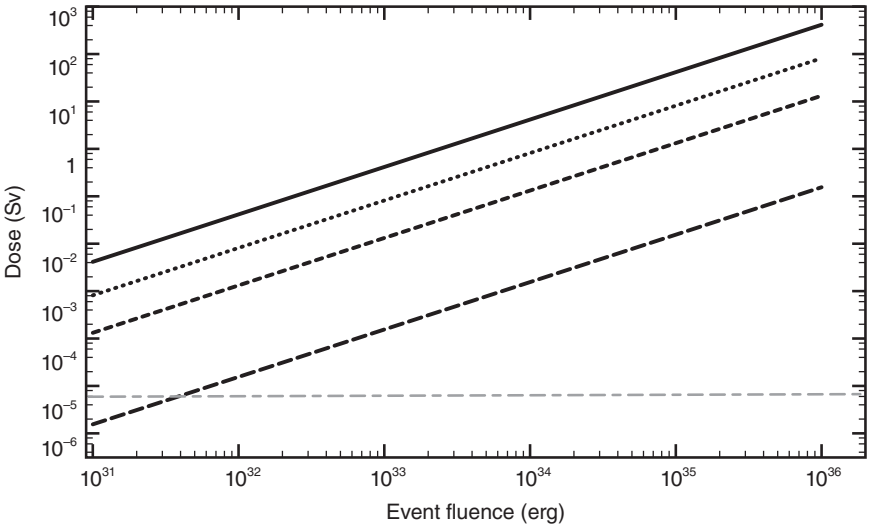


Figure 4.10 Potential radiation dose received at the surface due to particle cascades as a function of the flare energy for Proxima b. The solid, dotted, short-dashed, and long-dashed curves correspond to atmospheric column densities of 10^3 kg m^{-2}, 3×10^3 kg m^{-2}, 7×10^3 kg m^{-2}, and 10^4 kg m^{-2} (Earthlike), respectively. The magnetic moment of the planet has been fixed to equal that of the Earth. The horizontal dashed line denotes the average daily radiation dose received on Earth. Radiation doses below a few Sv are not lethal to complex multicellular life on Earth. (© The Authors. By permission of Oxford University Press. *Source:* Dimitra Atri [2016], Modelling stellar proton event-induced particle radiation dose on close-in exoplanets, *Monthly Notices of the Royal Astronomical Society* 465[1]: L34–L38, fig. 8.)

have energies above 290 MeV (known as the pion production threshold) initiate the formation of secondary particle cascades that can reach the surface and increase the radiation dose received therein. This topic was subject to detailed numerical investigations by Atri (2017, 2020) and Yamashiki et al. (2019). Broadly speaking, the radiation dose at the surface was enhanced at (1) low atmospheric column density, (2) high flare energy, (3) low planetary magnetic moment, and (4) smaller orbital radii. On the basis of the preceding analysis in this chapter, it is conceivable that one or more of (1)–(4) are valid for exoplanets around M-dwarfs.

The radiation dose received at Proxima b due to the particle cascades initiated by an SPE is depicted in Figure 4.10. The lethal radiation dose for humans and small mammals is ~ 5 to 10 Sv, whereas it is two to three orders

of magnitude higher for insects.[11] Inspecting the figure makes it apparent that the radiation dose contributed by SPEs is not likely to be lethal to complex multicellular organisms on Earth unless one considers the combination of rarefied atmospheres and large superflares. Reducing the magnetic moment from $\mathcal{M}_\oplus$ (adopted in the figure) to $0.05\,\mathcal{M}_\oplus$ is anticipated to increase the radiation dose by roughly one order of magnitude, but even this enhancement ought not adversely impact most of Earth's lifeforms.

Hitherto, we have listed only the downsides attributed to SEPs, but there are potential benefits as well. In order for the synthesis of biological building blocks to occur, the availability of suitable energy sources and pathways is known to be necessary. Numerical simulations undertaken by Airapetian et al. (2016) indicate that hydrogen cyanide (HCN)—one of the vital feedstock molecules (see Section 2.3.1)—is produced at concentrations of tens of ppmv (parts per million by volume) in the lower atmosphere during SPEs. The synthesis of nitrogen oxides via SEPs (as previously noted) and subsequent delivery of these electron acceptors to water bodies might facilitate the polymerization of RNA, emergence of protometabolic pathways, and the origin of life thereafter in either submarine hydrothermal vents or shallow ponds (Wong et al. 2017; Ranjan et al. 2019).

Moving beyond feedstock molecules, we saw in Section 2.3.5 that key biomolecular building blocks such as amino acids and nucleobases could be produced by SEPs at relatively high efficiencies based on laboratory experiments involving the bombardment of gaseous mixtures with high-energy protons. The time-averaged energy flux ($\Phi_{\rm SEP}$) incident on the surface of planets with 1 bar atmospheres due to SEPs was estimated in Lingam et al. (2018):

$$\Phi_{\rm SEP} \sim 50\,{\rm J\,m^{-2}\,yr^{-1}} \left(\frac{\dot{N}_{25}}{1\ {\rm day^{-1}}}\right)\left(\frac{a}{1\ {\rm AU}}\right)^{-2}, \tag{4.91}$$

where $\dot{N}_{25}$ refers to the number of SPEs linked to flares with $E_f \gtrsim 10^{25}$ J that impact the planet per day. This rate is possibly of order unity for both young Sunlike stars and active M-dwarfs (Kay et al. 2016, 2019), although it is orders of magnitude lower for modern Earth. An important point worth bearing in mind is that $\Phi_{\rm SEP}$ is higher for planets in the HZs of low-mass

11. The sievert (Sv) represents the SI unit used for measuring the effective dose of ionizing radiation.

stars because of the smaller orbital radii. Hence, for such planets, SPEs might provide an alternative means of bypassing the paucity of UV radiation for prebiotic chemistry outlined in Section 4.3.2. Lingam et al. (2018) drew on data from laboratory experiments to suggest that the production rates of prebiotic molecules was roughly given by

$$\dot{\mathcal{M}}_A \sim 10^7 \,\mathrm{kg/yr} \left(\frac{\Phi_{\mathrm{SEP}}}{100 \,\mathrm{J\,m^{-2}\,yr^{-1}}} \right) \left(\frac{R}{R_\oplus} \right)^2 \tag{4.92}$$

and

$$\dot{\mathcal{M}}_N \sim 10^4 \,\mathrm{kg/yr} \left(\frac{\Phi_{\mathrm{SEP}}}{100 \,\mathrm{J\,m^{-2}\,yr^{-1}}} \right) \left(\frac{R}{R_\oplus} \right)^2, \tag{4.93}$$

with $\dot{\mathcal{M}}_A$ and $\dot{\mathcal{M}}_N$ representing the rates of synthesizing amino acids and nucleobases, respectively. On early Earth (~ 4 Ga), it was estimated that $\dot{\mathcal{M}}_A$ may have been comparable to the production rate of amino acids via lightning (electrical discharges), and three to four orders of magnitude higher than the exogenous delivery of these molecules by meteorites.

Lastly, we turn our attention to an important geological conundrum—to wit, the faint young Sun problem encountered in Section 2.1.1. Several authors have proposed that the primordial Earth necessitated a high concentration of greenhouse gases in order to prevent oceans from freezing due to the diminished solar luminosity (~ 70 percent of the present-day value). Nitrous oxide (N_2O) is synthesized in sizable quantities during SPEs (Airapetian et al. 2016) and is characterized by an innate greenhouse potential that is ~ 300 times higher relative to CO_2 over a 100 yr period. Together, these points make it plausible that N_2O may have contributed significantly to the greenhouse warming of early Earth and other temperate exoplanets. In this event, SPEs could serve as a vital mechanism for sustaining clement conditions during such epochs.

4.5 CONCLUSION

We are the stars which sing;
We sing with our light;
We are the birds of fire,
We fly over the sky.
Our light is a voice;
. . . . . .

We look down on the mountains.
This is the Song of the Stars.

—"Song of the Stars," in *The Algonquin Legends of New England; or Myths and Folk Lore of the Micmac, Passamaquoddy, and Penobscot Tribes*

Ever since the dawn of civilizations, and probably earlier, humanity has worshipped the Sun as a giver of life. It is not much of an exaggeration to state that the majority of complex multicellular life on Earth, either directly (photosynthesis) or indirectly, is dependent on the Sun's existence. These considerations lead us to a crucial and underappreciated question: Why do humans, in their capacity as intelligent observers, find themselves on a planet orbiting a G-type star? Before addressing this question, we may pose a closely related question: Is the Sun anomalous in any respect compared to other stars in our Galaxy? If the Sun were a typical star in all respects, then our presence around the Sun could be viewed as a matter of random chance.

To answer this second question, Robles et al. (2008) conducted a comprehensive survey of eleven stellar characteristics such as the age, metallicity, elemental abundances, and rotational velocity. The study concluded that the Sun was atypical in two respects: (1) mass and (2) Galactic orbit. The significance of (2) is that the Sun has a less eccentric orbit than most stars, thus ensuring that the Solar system experiences a more stable Galactic "climate". As (2) falls within the domain of the Galactic Habitable Zone, we shall not discuss it further. Instead, let us turn our attention to (1)—namely, the datum that the Sun is more massive than $\sim 95\%$ of other stars in its neighborhood (T. J. Henry et al. 2006, 2018).

In other words, the Sun is uncommon with respect to its mass. The most common type of stars in this category are M-dwarfs. These stars are not only more common but also much more long-lived than the Sun. For example, consider two stars with $M_\star \sim 0.1\, M_\odot$ and $M_\star \sim M_\odot$. The former is about ~ 10 times more abundant than the latter and may live up to ~ 1000 times longer. As a result, if the probability of the emergence of intelligent observers (e.g., humans) did not depend on stellar parameters, one should expect them to be much more prevalent in the vicinity of M-dwarfs in the cosmic future. This statement was quantified in A. Loeb et al. (2016), who sought to evaluate the habitability of the Universe as a function of cosmic time by taking stellar formation rates, abundances, lifetimes, and other factors into account. The probability of cosmic habitability as a function of time is plotted in Figure 4.11.

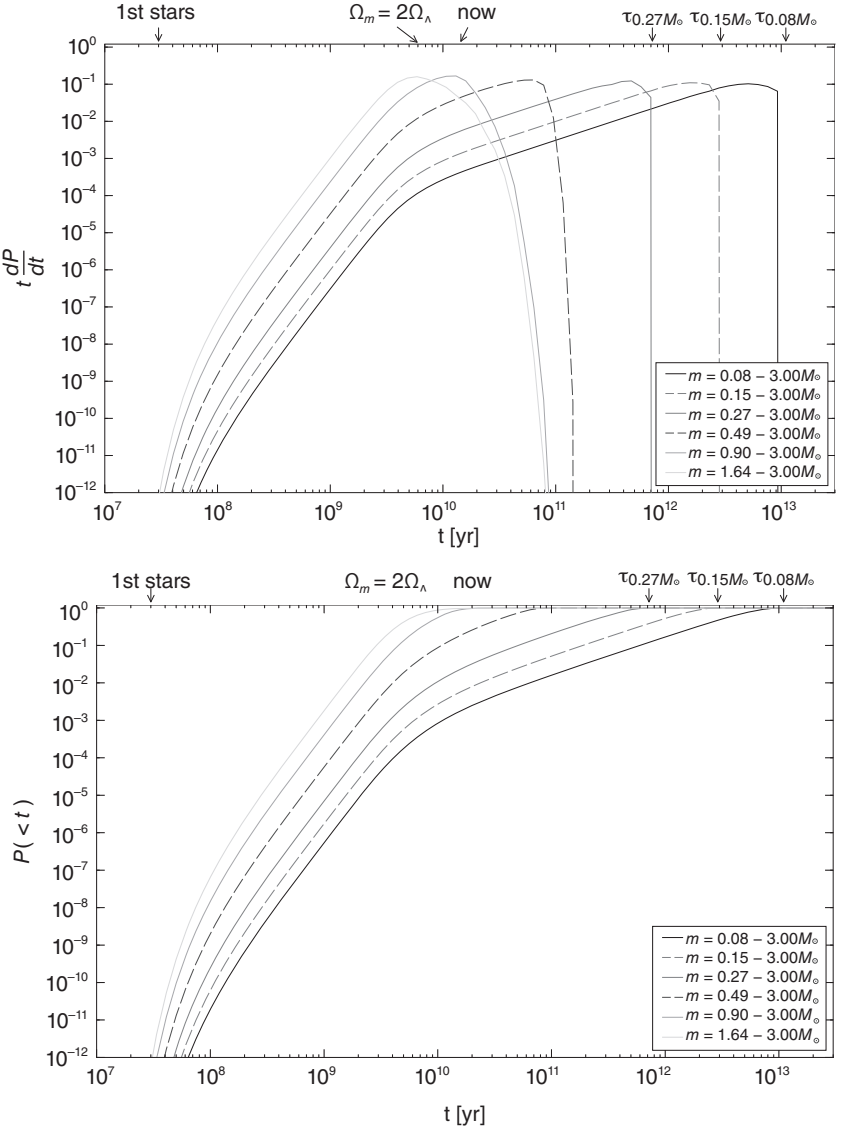


Figure 4.11 Probability distribution for the emergence of life within a fixed volume of the Universe as a function of cosmic time. Probability per log time tdP/dt (*top*) and cumulative probability $P(<t)$ (*bottom*) are shown for different *minimum* stellar masses for habitability. Labels on the top-axes depict formation of the first stars, start of cosmic acceleration ($\Omega_m = 2\Omega_\Lambda$), present day, and lifetimes of stars with masses $0.27\,M_\odot$, $0.15\,M_\odot$, and $0.08\,M_\odot$. The rightmost and leftmost curves correspond to $m = 0.08{-}3.00\,M_\odot$ and $m = 1.64{-}3.00\,M_\odot$, respectively; as one moves from left to right, the lower bound on the mass is successively reduced. (*Source:* Abraham Loeb, Rafael A. Batista, and David Sloan [2016], Relative likelihood for life as a function of cosmic time, *Journal of Cosmology and Astroparticle Physics* 08: 040, fig. 4.)

Hence, we can frame the first question from the above paragraph as follows: Why do humans find themselves near a Solar-type star instead of an M-dwarf in the cosmic future? This question has two broad classes of explanation:

- The probability of intelligent observers evolving on planets is independent of stellar mass. In this case, our emergence is anomalous, because the odds in favor of this occurring in the current epoch are $\lesssim 10^{-3}$, as seen in the bottom panel of Figure 4.11.
- Alternatively, the evolution of intelligent observers may depend on stellar mass. This scenario could materialize if physical mechanisms act in tandem to selectively lower the likelihood of intelligent life near M-dwarfs relative to Sun-like stars by at least $\sim$ three to four orders of magnitude. In this case, our origination in this particular eon will be fairly typical.

As of now, we do not have a clear picture of which of the above two hypotheses offers the correct explanation. However, on the basis of the models we have delineated in this chapter, there are a number of potential reasons why the second alternative might be favored, i.e., the habitability of M-dwarf exoplanets could be suppressed for microbial (and therefore technological) species.[12] The chief candidates in this respect include rapid atmospheric erosion via intense stellar winds and regular coronal mass ejections; a paucity of photons for achieving crucial evolutionary breakthroughs such as the origin of life, biodiversity, and photosynthesis; and a hostile surface environment because of stellar flares and stellar proton events.

On the other hand, we emphasize that counteracting physical mechanisms are capable of canceling out the detrimental effects caused by each of them. For example, M-dwarf exoplanets can start out with massive hydrogen-helium atmospheres that are ostensibly inimical to habitability and subsequently lose them through atmospheric escape driven by UV radiation or stellar winds; in this context, the latter two phenomena would not

12. An alternative explanation that merits further scrutiny is that M-dwarf exoplanets are conducive to hosting microbial life but present obstacles to the evolution of technological intelligence.

be detrimental but actually positive to habitability. Similarly, it is conceivable that planets with initially high water inventories (water worlds) were depleted via UV photolysis, thus giving rise to worlds with landmasses and oceans on the surface. In addition, planets may have formed outside the HZ, thereby escaping the brunt of the extended and desiccating pre-main-sequence phase of M-dwarfs, before migrating inward to the HZ at a later juncture. Lastly, M-dwarf planetary systems could possess inherent advantages such as enhanced lithopanspermia (i.e., transport of life between planets through rocky ejecta) that is several orders of magnitude more efficient than within our Solar system (Lingam & Loeb 2017a). On the other hand, in each of the above instances, either subtle fine tuning might be required or the plausibility of the proposed physical mechanisms remains indeterminate.

Finally, we wish to highlight some caveats regarding the bulk of our analyses in this chapter. For starters, we have ignored the intricate and interconnected array of planetary processes that can shape habitability over much shorter timescales than those encountered herein. One of the most notable examples in this realm is biogeochemistry, which is responsible for initiating nonlinear feedbacks that may result in drastic changes to the evolutionary landscape at a comparatively rapid rate. Another example concerning our neglect of planetary factors pertains to our analysis of photon availability for abiogenesis and photosynthesis in Section 4.3. We assumed that the photon fluxes represented the only limitation, but the availability of sufficient nutrients and reactants on the planet is equally vital. Moving on to stellar factors, we have focused primarily on stellar mass herein as a proxy, but it is worth appreciating that the likes of stellar age, activity, and rotation (all of which are time dependent) also regulate planetary habitability. To sum up, we have extensively employed the principle of ceteris paribus (holding all variables but one fixed) in this chapter. Yet, by doing so, we run the risk of missing out on dynamical evolution that necessitates the simultaneous interplay of more than one factor.

The questions we have posed are not merely of philosophical import, as they also possess great practical value. The rationale behind the latter is that humanity's observational resources, time, and funding are all limited. Hence, the selection of the optimal target stars and planets in the search for extraterrestrial life is of the utmost importance. The development

of theoretical models for assessing the habitability of exoplanets around different stars is arguably necessary for identifying the targets worthy of highest consideration. Moreover, such modeling can assist in identifying the putative biospheres that may exist and, therefore, in determining what types of biosignatures (and technosignatures) are to be expected.

With the impending launch of the James Webb Space Telescope (JWST) and the ongoing development of ground-based Extremely Large Telescopes (ELTs), we will soon be in a position to characterize terrestrial exoplanets and search for signatures of life around various stars. Thus, it is not unreasonable to hope that the answer to our previous question—Why do we find ourselves near a Sun-like star?—will emerge in the coming decades. If so, we might be able to determine the rarity or commonality of life in the Universe and which epoch in cosmic history (see Figure 4.11) is likely to be the most habitable.

Chapter 5

HABITABILITY: PLANETARY FACTORS

The fields, the lakes, the forests, and the streams,
Ocean, and all the living things that dwell
Within the daedal earth; lightning, and rain,
Earthquake, and fiery flood, and hurricane
The torpor of the year when feeble dreams
Visit the hidden buds, or dreamless sleep
Holds every future leaf and flower;—the bound
With which from that detested trance they leap;
. . . . . .
And *this,* the naked countenance of earth,
On which I gaze, even these primaeval mountains
Teach the adverting mind.

—Percy Bysshe Shelley, *Mont Blanc*

What constitutes a habitable planet? We sought to answer this question to some degree in the previous chapter, where we chose to focus almost exclusively on stellar factors. In doing so, we restricted ourselves to worlds that were essentially identical to that of modern Earth. However, even a cursory examination of our planet's evolutionary history, as laid out in Chapter 3, reveals that its propensity to host and sustain microbial and complex multicellular life has evolved drastically from one epoch to the next. Hence, we hold it self-evident that assessing the biological habitability of other worlds requires a proper consideration of planetary mechanisms and not just stellar processes.

If we accept this entirely reasonable premise, it raises another crucial question: What factors merit the highest attention? Needless to say, answering this question, provided that it is meaningful in the first place, is not an easy task. Any attempt to construct a fully dynamical picture that encapsulates variegated phenomena operating across multiple spatial and temporal scales and coevolving together via nonlinear feedbacks is well-nigh impossible: such a picture remains currently difficult to assemble even when it

comes to Earth (Lapôtre et al. 2020). On the other hand, resorting to crude metrics that scale basic physical parameters of a given world (e.g., size and effective temperature) with respect to the Earth is also problematic, because these prescriptions yield minimal information about habitability (Tasker et al. 2017). Therefore, from a scientific standpoint, we must navigate the waters carefully, between the Scylla of labyrinthine, tangled reality and the Charybdis of reductive oversimplification, to construct models that are reasonably accurate and tractable in tandem. In order to do so, we shall mostly exclude three broad categories of processes from this chapter, as described below.

The first ingredient that we do not address is the question of whether a given world can support a stable and clement climate over Gyr timescales. Naturally, this criterion represents one of the cornerstones for a planet to host life on the surface. Our reason for omitting the climate stems from its complexity and the large number of variables involved. For example, numerical models have illustrated that the climate depends on the orbital eccentricity and axial tilt (obliquity), the fraction of the surface covered by landmasses, the extent of oceans and their salinity, and many other abiotic factors (Shields 2019; X. Zhang 2020). Apart from these, there is growing empirical and theoretical evidence that life shapes its own environment via biological feedback mechanisms. It is therefore apparent that an exhaustive treatment of planetary climate lies beyond the scope of this book. Nevertheless, we do briefly investigate some aspects pertaining to this area in Section 5.3.

Moving ahead, biogeochemistry is a comparatively new field, albeit with old roots, that seeks to study how geological, biological, chemical, and physical processes collectively regulate the dynamical states of the natural environment. By its very nature, biogeochemistry is clearly of paramount importance for comprehending habitability, while simultaneously very difficult to model accurately, even when it concerns our planet. Hence, while there exists no doubt that biogeochemistry must constitute a central component of habitability studies in the future, we do not explore this topic in much detail herein. Although we do not address the question of how biogeochemical cycles of bioessential elements operate on other worlds, we do examine the issue of oceanic phosphorus and biomass abundance in Section 5.5.

Finally, we shall delve into neither the architecture of planetary systems nor orbital parameters such as obliquity and eccentricity. With regard to

the former, let us consider our Solar system. The presence of the Moon has been alleged to confer benefits ranging from the stabilization of the Earth's obliquity to generating modulated oceanic tides that stimulated abiogenesis, nutrient redistribution in oceans, biological clocks, and the evolution of tetrapods on land; the reader is referred to Lingam and Loeb (2018b) for a summary. Likewise, the presence of Jupiter and Saturn could have shielded the Earth from frequent impacts by comets and asteroids. Alternatively, they may have redirected planetesimals toward the Earth and assisted in the delivery of volatiles such as water. However, it does not automatically follow that planetary systems sans large moons, Jupiter, or Saturn would have a low likelihood of complex life. Thus, we opt not to tackle planetary architecture since it merits investigation on a case-by-case basis. Turning our attention to orbital factors, there is broad consensus that they play a pivotal role in regulating habitability (Shields et al. 2016), but which regions of the orbital parameter space preclude the existence of biospheres altogether remains unsettled.

Bearing these caveats in mind, we carry out an exploration of how the distribution of landmasses and oceans on the surface, the average surface temperature, plate tectonics, atmospheric composition, and tidal locking may influence the habitability of other worlds. We find that these factors could collectively either enhance or suppress biological habitability by orders of magnitude. We recommend readers peruse Lammer et al. (2009), Kasting (2010), Cockell et al. (2016), and Ehlmann et al. (2016) for reviews of exoplanetary habitability.

5.1 THE MYRIAD ROLES OF TEMPERATURE

In Section 4.1, we encountered the concept of the habitable zone (HZ), which represents the region surrounding the star(s) that is theoretically capable of hosting liquid water on the surface of a rocky planet. One must, however, be cognizant of two important points: (1) liquid water will exist outside the standard range of 273–373 K depending on the pressure, and (2) the range over which water is liquid at 1 bar (273–373 K) does *not* coincide with the range in which life survives on Earth. As per recent estimates, extremophiles on Earth are metabolically active across the temperature range of $-20\,^{\circ}\mathrm{C}$ (253 K) to $122\,^{\circ}\mathrm{C}$ (395 K); the reader should consult Clarke (2014), C. P. McKay (2014), and Merino et al. (2019) for

additional details. The lower limit is conventionally believed to arise from a combination of desiccation and vitrification induced by freezing. Vitrification, also known as the glass transition, results in the liquid essentially behaving like a solid. Due to the much higher internal viscosity of cells post-vitrification (at roughly $-20\,^{\circ}\mathrm{C}$), the diffusion of oxygen and metabolic products is slowed to such an extent that metabolism effectively grinds to a halt. The upper thermal limit for life on Earth is expected to stem from the destabilization of lipid bilayers (present in cell membranes) and proteins, partly due to the changes in the dielectric constant of water.

Hitherto, we have restricted ourselves to biological considerations based on liquid water as the solvent. We may seek to advance further, and ask ourselves whether physical chemistry itself sets limits on life. The answer may lie in the unique properties of hydrogen bonds, owing to their capacity to participate in (1) chain-pairing interactions that are necessary for both catalytic and replicative functions and (2) intermolecular interactions that are required for synthesizing new molecules, catalysis and exchanging information (Desiraju & Steiner 1999). We will therefore briefly outline the hypothesis presented in Vladilo & Hassanali (2018), which relies on the aforementioned attributes of hydrogen bonds. First, let us recall that the reaction rate (k) of a chemical reaction is expressible as

$$k = A \exp\left(-\frac{E_a}{k_B T_r}\right), \tag{5.1}$$

where A is a constant pre-exponential factor, k_B is the Boltzmann constant, E_a denotes the activation energy, and T_r represents the temperature at which the reaction takes place. In theory, the reaction rate will never become zero provided that E_a and T_r remain finite and nonzero. However, the reaction rate can be considered "slow" when $k/A < 10^{-3}$ is valid, which translates to

$$T_r \lesssim \frac{E_a}{6.9 k_B}. \tag{5.2}$$

Likewise, the upper bound $k \to A$ cannot be approached as long as E_a and T_r are nonzero and finite. It is therefore reasonable to suppose that the reaction can be "fast" if $k/A > 0.1$ holds true, thereby amounting to

$$T_r \gtrsim \frac{E_a}{2.3 k_B}. \tag{5.3}$$

The upper thermal limit ($T_{\max}$) is calculated by assuming that the N-H$\cdots$O bonds that exist in proteins and nucleic acids, or functionally equivalent biomolecules, must not be disrupted; in other words, the corresponding reaction must be sufficiently slow. By plugging $E_a \sim 0.26$ eV for this reaction into (5.2), we obtain $T_{\max} \approx 437$ K. Next, in order to estimate the lower thermal limit ($T_{\min}$), the activation (and deactivation) of hydrogen bonds for molecular signaling should be fast enough. Thus, by using (5.3) along with the fact that the minimum activation energy for such reactions is $E_a \sim 4.3 \times 10^{-2}$ eV, we end up with $T_{\min} \approx 217$ K.

To sum up, by solely restricting ourselves to assaying the properties of hydrogen bonds, we wound up with the heuristic limits $2.2 \times 10^2 \lesssim T \lesssim 4.4 \times 10^2$ K. This range is somewhat broader than the limits $2.5 \times 10^2 \lesssim T \lesssim 4.0 \times 10^2$ K for life on Earth. Yet, an important point worth appreciating is that the theoretical bounds calculated herein are *not* explicitly dependent on the solvent, and they impose minimal constraints on the chemical composition and structure of the putative biomolecules. Taking additional restrictions such as the choice of solvent (e.g., water) and the thermal stability of specific metabolic products into account will probably narrow the range further.

The importance of temperature is not merely confined to identifying the regimes at which life could exist. Instead, it plays a key role in determining the nature of ecosystems, the physiology of organisms, and the rates of biochemical reactions. Only a few select examples are addressed below, but Cossins and Bowler (1987) and Angilletta (2009) comprehensively explore the centrality of temperature in biology, for those who wish to delve further into this subject.

5.1.1 The Metabolic Theory of Ecology and temperature

The Metabolic Theory of Ecology (MTE) is arguably one of the most influential quantitative ecological models to emerge in the past few decades. Further reviews of this topic can be found in Brown et al. (2004), Price et al. (2012), and A. Clarke (2017). The MTE is based on the premise that the metabolic rate B_0 of organisms regulates biological processes spanning molecular, organismal, and ecological scales. The MTE has the advantage of possessing a physicochemical perspective, without having to invoke intricate and specialized biological factors. Naturally, this approach has attracted

criticism for reasons that we shall encounter shortly (D. S. Glazier 2015). Despite the limitations, we will work with the MTE since it may potentially help us understand the basic features of organisms and ecosystems on other worlds to leading order.

It is presumed that the metabolic rate for a given organism obeys the scaling

$$B_0 \propto m_0^{3/4} \exp\left(-\frac{\bar{E}}{k_B T_0}\right), \tag{5.4}$$

where m_0 and T_0 refer to the mass and the temperature of the organism, and $\bar{E}$ signifies the average activation energy that is determined from the appropriate rate-limiting metabolic step (Gillooly et al. 2001). On Earth, experiments indicate that $\bar{E}$ is typically 0.6 to 0.7 eV for respiration-dependent processes in aerobic microorganisms, plants, and animals (Dell et al. 2011). The mean value, corresponding to $\bar{E} = 0.65$ eV, is virtually equal to the activation energy of 0.66 eV estimated from the rate of ATP synthesis in isolated mitochondria (Gillooly et al. 2006). In contrast, when it comes to photosynthesis, the MTE predicts $\bar{E} \approx 0.32$ eV.

The following caveats should be borne in mind with reference to (5.4). For starters, the power-law exponent of 3/4 has been the subject of controversy for many decades: the current consensus favors a conspicuous degree of variation across organisms (DeLong et al. 2010; Glazier 2015). Second, by inspecting (5.4), it is evident that B_0 is a monotonically increasing function of T_0. This expression constitutes a clear idealization because the trend cannot continue ad infinitum because thermal adaptation, especially for complex multicellular organisms, breaks down after a certain point (Schulte 2015). As opposed to the Boltzmann factor ansatz, a number of alternatives have been investigated, such as

$$B_0 \propto \exp\left(-\frac{\bar{E}}{k_B T_0}\right)\left[1 + \exp\left(-\frac{E_h}{k_B T_0}\right) + \exp\left(-\frac{E_l}{k_B T_0}\right)\right]^{-1}, \tag{5.5}$$

where E_h and E_l are interpreted as the high- and low-temperature deactivation energies of enzymes, respectively (Sharpe & DeMichele 1977; Pawar et al. 2016). Third, it should be noted that $\bar{E}$ is not truly constant since it varies across organisms and different temperatures. To put it somewhat differently, the Boltzmann factor does not accurately represent the thermal dependence in all instances. Lastly, it is not merely the temperature

that matters but also the timescales over which it fluctuates. Rapid changes in temperature are liable to cause physiological damage as well as mass extinctions and irreversible changes in the biosphere.

As mentioned earlier, the metabolic rate is at the heart of the MTE, with the implication that several biological processes are regulated by the former. Many empirical studies have been conducted to assess the temperature dependence inherent in (5.4), and the appropriate references were cataloged in Lingam and Loeb (2018g). We briefly highlight some of them below:

- The maximum rates for population growth and genetic (i.e., molecular) evolution.
- The rates at which genetic divergence and the formation of new species (speciation) occur.
- The rates of embryonic and postembryonic development for multicellular organisms.
- The rates at which the production, storage, and turnover of biomass happen in ecosystems. For instance, this includes the rate at which organic compounds are produced via photosynthesis.
- The metabolic balance, fluxes of energy and materials, and trophic interactions in ecosystems are mediated by temperature.
- Higher biodiversity is predicted for habitats at hotter average temperatures. This pattern has been confirmed on Earth, with the number of species (i.e., species richness) declining as one moves from the equator toward the poles. It goes without saying that other factors, like the degree of precipitation, also play an important role.

The Boltzmann factor inherent in (5.4) predicts that most of the above quantities are expected to monotonically increase with temperature. Until a certain optimal value, the general rule of thumb is "*Hotter is better*" (Kingsolver & Huey 2008), although striking exceptions can and do exist by virtue of evolutionary adaptations (Kingsolver 2009). Therefore, in this regime, a higher temperature is positively correlated with elevated growth, fitness, and diversity.

On the basis of our discussion thus far, we may ask ourselves this question: Is it possible to construct a coarse-grained metric that depends solely

on basic physical parameters? In order to resolve this matter, let us turn our attention to (5.4). The value of T_0 depends on the local habitat under question for ectotherms (organisms whose body temperature depends on the environment), whereas it is governed by biological factors when it comes to endotherms (organisms that maintain a favorable body temperature through internal sources). Since we wish to design a planetary metric, we propose replacing T_0 with the average surface temperature of the planet (T_s). It follows, as a matter of course, that this ansatz does not capture spatial heterogeneity nor does it account for extraterrestrial endotherms. The advantage, however, arising from this substitution is that T_s represents a physical parameter that is measurable by upcoming observations.

We are primarily interested in normalizing the Boltzmann factor from (5.4) relative to the Earth. Thus, we introduce the function

$$\begin{aligned}\Delta_T &= \exp\left(\frac{\bar{E}_\oplus}{k_B T_\oplus} - \frac{\bar{E}}{k_B T_s}\right)\theta\,(T_s - T_L)\,\theta\,(T_U - T_s)\\ &= \exp\left[26.6\left(1 - \frac{\bar{E}}{\bar{E}_\oplus}\frac{T_\oplus}{T_s}\right)\right]\theta\,(T_s - T_L)\,\theta\,(T_U - T_s)\,, \qquad (5.6)\end{aligned}$$

where $\bar{E}_\oplus = 0.66$ eV, and $T_\oplus \approx 288$ K denotes the average surface temperature of the Earth. In the above expression, θ is the Heaviside step function, thereby ensuring that Δ_T becomes zero when $T_L < T_s < T_U$ is violated. T_L and T_U can be duly interpreted as the lower and upper thermal limits constraining the validity of the Boltzmann factor model. From (5.6), we see that the activation energy remains an unknown quantity. If we suppose that other worlds evolve metabolic pathways akin to those on Earth, we may set $\bar{E}/\bar{E}_\oplus \sim 1$. We will employ this scaling henceforth and drop the Heaviside functions by positing that $T_L < T_s < T_U$.

We contend that (5.6) is a simple and useful proxy, ceteris paribus, for estimating the biological potential of other worlds. It is worth reiterating here that the global surface temperature is the only variable, with all other parameters held constant. Our rationale is based on the observation that (5.6) resembles the functional form of the (relative) metabolic rate insofar as the temperature dependence is concerned, which, in turn, has been theorized to regulate a diverse array of crucial functions according to the MTE. Hence, in the event that the origin of life was successful, the odds for

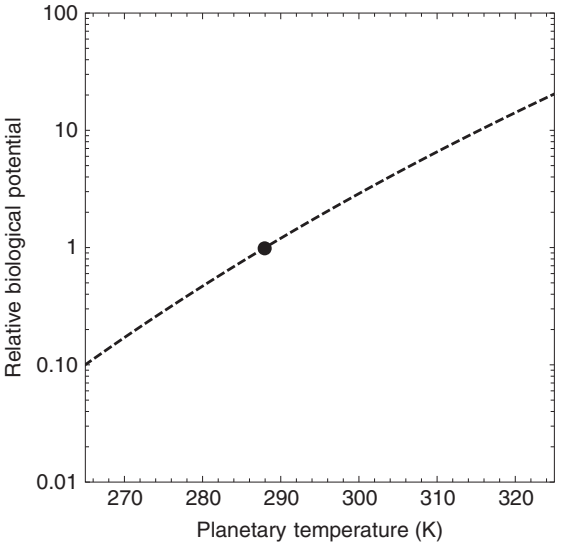


Figure 5.1 The prospective biological potential (5.6) for a given Earthlike planet in comparison with our planet as a function of the average surface temperature (T_s), with all other factors held fixed. The black dot represents the Earth's surface temperature of 288 K at which (5.6) becomes equal to unity. (© Manasvi Lingam and Avi Loeb.)

the emergence of complex biospheres might be enhanced at higher global (steady-state) surface temperatures because (1) the rates of genetic evolution and speciation are amplified and (2) greater species diversity is potentially sustainable.

We have therefore plotted the function (5.6) in Figure 5.1. The minimum surface temperature chosen (260 K) preserves consistency with the lower thermal limit documented for life on Earth, while the maximum (330 K) is based on experiments that have revealed that the monotonic trend breaks down at $\gtrsim 300$ K for multicellular organisms (Schulte 2015) and $\gtrsim 330$ K for thermophilic microbes (Brock 1985). Drawing on the preceding discussion, we may loosely interpret this plot as the relative likelihood of hosting a complex biosphere in the event that the Earth's surface temperature differed from its current value. Due to the exponential relationship in (5.6), the likelihood varies over two orders of magnitude despite T_s spanning a comparatively narrow range. This result underscores the fact that numerous macroecological processes are quite sensitive to the ambient

temperature, implying that the latter merits a central role in studies of biological habitability.

5.1.2 Thermal ramifications of stellar flares

We investigated the impact of stellar flares on habitability in Section 4.4, but we did not discuss the thermal impact on planets therein. The mean bolometric luminosity L_f can be estimated from the total flare energy E_f and its duration τ_f via $L_f \sim E_f/\tau_f$. We will make use of the theoretical ansatz $\tau_f \propto E_f^{1/3}$, since this scaling exhibits good agreement with observations of solar flares as well as stellar superflares (Maehara et al. 2015); recall that superflares typically obey $E_f \gtrsim 10^{26}$ J. Determining the proportionality constant is harder because it varies across flares, stellar cycles, and stars. We adopt the relation $\tau_f \sim 10^4 \text{ s } \left(E_f/10^{27}\,\text{J}\right)^{1/3}$, as it yields $\tau_f \sim 36$ min for $E_f \sim 10^{25}$ J, which is virtually equal to the average duration of X-class flares (whose energies can reach 10^{25} J) in this solar cycle, based on data collected by Geostationary Orbital Environmental Satellites (Swalwell et al. 2018). Thus, we end up with

$$L_f \sim 10^{23}\,\text{W} \left(\frac{E_f}{10^{27}\,\text{J}}\right)^{2/3}, \tag{5.7}$$

but we wish to caution that the above expression may actually represent an underestimate for superflares on Solar-type stars (Maehara et al. 2015). Next, we introduce the ratio $\delta_f = L_f/L_\star$, with $L_\star$ denoting the stellar luminosity. By combining (5.7) with the mass-luminosity relationship $L_\star \propto M_\star^3$, we obtain

$$\delta_f \sim 2.6 \times 10^{-4} \left(\frac{E_f}{10^{27}\,\text{J}}\right)^{2/3} \left(\frac{M_\star}{M_\odot}\right)^{-3}. \tag{5.8}$$

Thus, for late-type M-dwarfs with $M_\star \sim 0.1 M_\odot$, we see that $\delta_f \sim 1$ is attainable when $E_f \sim 10^{27}$ J. The largest superflares detected on Proxima Centauri (with $M_\star \approx 0.12 M_\odot$) and TRAPPIST-1 (with $M_\star \approx 0.09 M_\odot$) have energies of $10^{26.5}$ J and $10^{26.1}$ J, respectively, thus indicating that superflares of this magnitude are possible on active late-type M-dwarfs. The *Kepler* and *TESS* missions have also empirically confirmed the existence of superflares on M-dwarfs with energies $\gtrsim 10^{27}$ J. In consequence, achieving $\delta_f \sim 1$ is feasible for stars from this category.

We saw in Section 4.1 that the effective temperature $T_{\rm eq}$ of the planet equals

$$\sigma T_{\rm eq}^4 = \frac{L_\star \left(1 - A_p\right)}{16\pi a^2}, \tag{5.9}$$

where A_p and a represent the albedo and orbital radius of the planet, respectively. However, this model only yields the effective temperature and not the actual surface temperature (T_s). In order to determine the latter, we will adopt a simple greenhouse model outlined in D. J. Jacob (1999) that applies to planets receiving Earthlike stellar fluxes. The basic idea is that the incoming stellar radiation is distributed across two different systems: the atmosphere and the planet (sans its atmosphere). The former is viewed as being transparent to stellar radiation but absorbs radiation from the surface due to the presence of greenhouse gases, with the absorption fraction denoted by f_a. The atmosphere is modeled as having both upward- and downward-facing surfaces, each of which emits a radiation flux $f_a \sigma T_a^4$, where T_a is the temperature of the atmosphere. The first energy balance equation as seen by an observer from space is

$$\left(1 - f_a\right) \sigma T_s^4 + f_a \sigma T_a^4 = \frac{L_\star \left(1 - A_p\right)}{16\pi a^2}, \tag{5.10}$$

while the second energy balance equation is formulated for the atmospheric layer,

$$2 f_a \sigma T_a^4 = f_a \sigma T_s^4, \tag{5.11}$$

which simplifies to $T_a^4 = T_s^4/2$. Substituting this relation in (5.10), we find

$$\sigma T_s^4 = \frac{1}{1 - 0.5 f_a} \frac{L_\star \left(1 - A_p\right)}{16\pi a^2} = \frac{1}{1 - 0.5 f_a} \sigma T_{\rm eq}^4. \tag{5.12}$$

For the Earth, specifying $f_a \approx 0.77$ yields $T_s \approx 288$ K and $T_a \approx 242$ K, both of which are in good agreement with observations (D. J. Jacob 1999), provided that the atmospheric temperature is evaluated at the scale height above ground level. A clear limitation of this model is that it does not account for other factors, such as the size, surface pressure, and global topography of the planet; collectively, these variables could modify the average surface temperature by $\lesssim 20$ K even when the incident stellar flux is held fixed.

We observe that surface temperature prior to the flare is T_s, whereas its mean value during the flaring interval is denoted by T_s'. The elevation in surface temperature due to the flare is $\Delta T_s = T_s' - T_s$. From (5.12), it is apparent that the surface temperature is proportional to $L_\star^{1/4}$. If we posit that f_a does not change during flares, we obtain

$$\Delta T_s = T_s \left[\left(1 + \delta_f\right)^{1/4} - 1 \right] \approx 0.25 \delta_f T_s, \tag{5.13}$$

where the last equality strictly applies only when $\delta_f \ll 1$. Now, let us select an Earthlike terrestrial planet with $T_s \approx T_\oplus$ orbiting a $\sim 0.1 M_\odot$ star on which a superflare with $E_f \sim 10^{27}$ J has erupted. Using (5.8), we obtain $\delta_f \approx 0.26$, and substituting this result into (5.13), we arrive at $\Delta T_s \approx 17$ K. The key point worth appreciating here is that this increase in temperature is comparatively rapid—that is, on the order of hours. Another crucial point is that δ_f quantifies the average enhancement in luminosity over the duration of the stellar flare. In contrast, the peak luminosity of the flare may exceed δ_f by an order of magnitude in certain instances. In fact, a peak value of $\delta_f \approx 16$ has been documented for the most extreme superflares based on the Bayesian analysis of data collected by the *TESS* mission (Günther et al. 2020).

To summarize, we have seen that Earth-analogs around late-type M-dwarfs will experience significant increases in the surface temperature due to superflares. When liver mitochondria extracted from the mummichog (*F. heteroclitus macrolepidotus*), a fish acclimatized to living at 15 °C ($T_\oplus$), were subjected to acute thermal stress, it was discovered that their functioning broke down at approximately 38 °C (Schulte 2015). The increase in temperature by ~ 23 K is very close to that calculated in the preceding paragraph. A recent theoretical study by Strona and Bradshaw (2018) determined that heating the Earth by 20 K may reduce the global biodiversity to ~ 10 percent of its original value, whereas incorporating the effects of coextinction stemming from the interconnected nature of food webs was found to reduce this estimate even further. Moreover, recent numerical and empirical models indicate that the Permian-Triassic extinction event, the largest mass extinction event recorded in our planet's history, was characterized by $\Delta T_s \sim 10$ K in the oceans and temperature-dependent anoxia. Collectively, these factors could have been responsible for $\gtrsim 50$ percent of the total losses during this period (J. L. Penn et al. 2018).

Although the above discussion can seem ominous, a couple of key points should be borne in mind. First, the temperature changes documented for mass extinctions persisted for timescales *at least* on the order of years,[1] whereas the thermal fluctuations induced by superflares will occur on timescales of hours to days. Thus, it is apt to regard the former as being chronic, whereas the latter are classifiable as acute. Second, our discussion was exclusively centered on Earth-based organisms, but there is no reason a priori to suppose that the same applies to other worlds as well. In particular, owing to the potentially higher frequency of thermal fluctuations on M-dwarf exoplanets, it is plausible that complex multicellular organisms have evolved sophisticated thermoregulation mechanisms akin to those found mostly in microbes on our planet. Hence, in balance, thermal perturbations due to superflares may pose impediments to biological habitability for planets in the HZ of late-type M-dwarfs, but they are not likely to be insurmountable in nature.

Earlier, we provided reasons why superflares with energies 10^{27} J could erupt on late-type M-dwarfs. The next question has to do with their frequency of occurrence. In the event that these superflares are truly rare, we can dismiss the above potential risks for the most part. Therefore, by positing that superflares with energies 10^{27} J are feasible, let us focus on the well-known Proxima Centauri and TRAPPIST-1 to deduce the occurrence rate of such phenomena. In the former case, after presuming that the power-law scaling for the cumulative flare frequency distribution presented in Davenport et al. (2016) and Vida, Oláh, et al. (2019) holds true at high energies, we find that superflares exceeding the desired threshold may occur ~ 0.5 to 2 times per year. By repeating this calculation for TRAPPIST-1 using the cumulative flare frequency distribution derived in Vida et al. (2017) and A. L. Glazier et al. (2020), we obtain a frequency of ~ 1 per year.

While we should not mistake Proxima Centauri and TRAPPIST-1 as being representative of all active late-type M-dwarfs, these calculations nevertheless suggest that superflares of $\gtrsim 10^{27}$ J on such stars are not uncommon. In fact, when compared to geological timescales that typically span

1. The Cretaceous-Paleogene extinction event that killed the dinosaurs was probably accompanied by global cooling that lowered the surface temperature by 11 to 28 K for a span of 3 to 4 years due to 1.5×10^{13} kg of soot being injected into the atmosphere following the impact of an asteroid that was ~ 10 km in diameter (Bardeen et al. 2017).

$\sim 10^3$ to 10^9 yr, one event per year comes across as very frequent. Thus, on the basis of the preceding analysis, further work is undoubtedly necessary to properly assess the thermal responses of lifeforms situated on temperate planets to superflares on late-type M-dwarfs.

5.2 PLATE TECTONICS AND HABITABILITY

In qualitative terms, *plate tectonics* describes the dynamics of the lithosphere (i.e., the outermost solid shell of a planet), which is divided into plates moving at relative velocities to each other. The reader is referred to Schubert et al. (2001) and Korenaga (2013) for reviews of this subject. Mantle convection is manifested as two different modes of tectonics: (1) plate tectonics involving fragmented plates, and (2) stagnant-lid tectonics, wherein the entire planetary surface behaves as a rigid spherical shell. The latter is much more common than the former—in our Solar system, the Earth is the only object with radius $\gtrsim 250$ km that displays unambiguous plate tectonics, although there are some indications that Europa may fall in this category. The preponderance of stagnant-lid convection stems from the strong dependence of the mantle viscosity (η) on the temperature (T); one such common ansatz for the mantle viscosity has the form

$$\eta(T) \propto \exp\left(\frac{E_a}{k_B T}\right), \tag{5.14}$$

where E_a is the appropriate activation energy. Thus, as one approaches the surface where the temperatures are comparatively low, we see that the mantle viscosity becomes very high. As a result, it becomes much harder for the spherical "lid" to undergo deformation and fragmentation. Current models therefore seek to explain how the effects of temperature-dependent viscosity may be offset by other mechanisms that weaken the lithosphere. Water has been theorized to play a crucial role because it can percolate into the mantle and hydrate rocky material, thereby making the shell more susceptible to bending. Moreover, it serves as a lubricant that enables the sinking of plates into the interior. This process, known as subduction, occurs because oceanic lithosphere that is $\gtrsim 30$ Myr old has an overdensity of roughly 1 percent compared to the asthenosphere, the region of the upper mantle underlying the lithosphere.

There exists considerable ambiguity about when plate tectonics was first instigated on Earth (M. Brown et al. 2020). Resolving this conundrum requires an in-depth understanding of when and how subduction was initiated, consequently facilitating the formation of new subduction zones. It is conceivable that this question is ill posed—that is, multiple tectonic episodes (stagnant-lid and plate tectonics) may very well have transpired in the past instead of a single distinct transition. Broadly speaking, there are three conditions that must hold true over large regions of the lithosphere in order for widespread plate tectonics to occur: (1) the lithosphere must be denser than the asthenosphere, (2) the lithosphere must remain intact near subduction zones, and (3) the lithosphere must have weak regions that can be ruptured to form new plates (Stern & Gerya 2018).

The onset of plate tectonics on Earth, as remarked above, has been the subject of intense debate. On the one hand, a number of mechanisms operational on Hadean-Archean ($\gtrsim 4$ Ga) Earth—for example, impacts by asteroids and comets (Maruyama et al. 2018)—are invoked to hypothesize that subduction and plate tectonics arose in this era. On the other hand, factors such as the higher mantle temperature provide theoretical grounds for contending that plate tectonics was not feasible during this epoch. This issue is further compounded by the paucity of reliable geological data from this period, which makes it harder to falsify or validate predictions or construct new hypotheses. While certain scientists have advocated that some variant of plate tectonics was initiated on Earth early in its history—namely, in the Hadean (Turner et al. 2020)—others like Stern (2018) favor a recent origin in the Neoproterozoic (~ 0.54–1.0 Ga). Saddling these two extremes, there are tentative grounds for conjecturing that steady plate tectonics commenced on Earth around 3 Ga (Hawkesworth et al. 2019).

5.2.1 The advantages of plate tectonics

One of the basic requirements for habitability is the sustenance of stable temperatures that permit the existence of liquid water over sufficiently long ($\sim$ Gyr) timescales. Here, it should be remembered that the origin of life, especially complex multicellular life, ostensibly requires the maintenance of habitable conditions over such timescales. On Earth, the regulation of surface temperature has been feasible by imposing strict controls on the

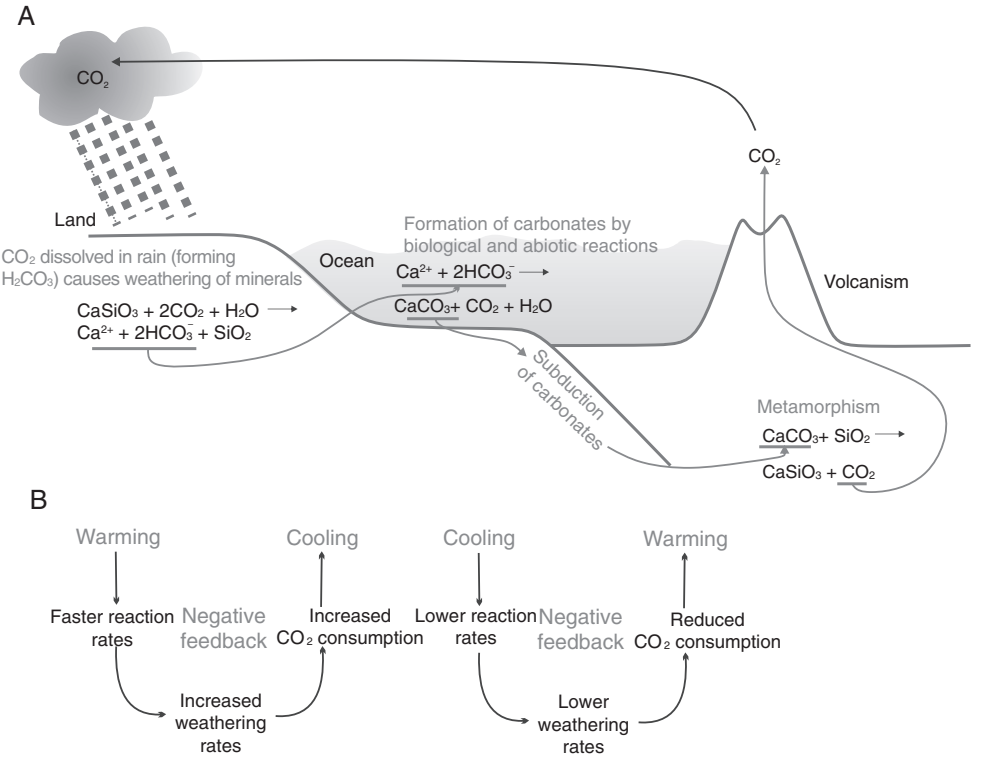


Figure 5.2 *A:* The carbonate-silicate cycle capable of regulating the surface temperature via feedback controls on the classic greenhouse gas carbon dioxide. The central role of plate tectonics in this cycle—namely, the subduction of carbonates—is also evident. *B:* How the carbonate-silicate cycle operates by enforcing a negative feedback loop. (© Mary Ann Liebert, Inc. *Source:* C. S. Cockell, T. Bush, C. Bryce, S. Direito, M. Fox-Powell, J. P. Harrison, H. Lammer, H. Landenmark, J. Martin-Torres, N. Nicholson, L. Noack, J. O'Malley-James, S. J. Payler, A. Rushby, T. Samuels, P. Schwendner, J. Wasdworth and M. P. Zorzano [2016], Habitability: A review, *Astrobiology* 16[1]: 89–117, fig. 6.)

abundance of carbon dioxide via the carbonate-silicate cycle (Walker et al. 1981; Kasting 2019; Isson et al. 2020). The mechanisms and feedback controls underlying the carbonate-silicate cycle are depicted in Figure 5.2.

CO_2 in the atmosphere dissolves in rainwater to form carbonic acid (H_2CO_3) that reacts with minerals to drive weathering. The ions liberated during this process are transported into the oceans, where they participate in abiotic and biological reactions to form carbonates. The latter are buried in marine sediments that are subducted into the interior,

where they undergo metamorphic reactions that liberate CO_2. The ensuing carbon dioxide is expelled into the atmosphere by virtue of volcanism. In the event that surface temperature increases due to higher concentrations of CO_2, it stimulates the enhancement of weathering rates, thereby resulting in the consumption of more CO_2 and reducing the temperature in turn. Along similar lines, the converse occurs when the surface temperature decreases, thus decreasing weathering rates and eventually driving up the temperature.

As we can see in Figure 5.2, plate tectonics enables the subduction of carbonates from the ocean, consequently opening up a pathway for returning CO_2 to the atmosphere through volcanic outgassing. Moreover, it has been argued to play a vital role in providing fresh material for weathering by uplifting calcium-silicate rocks to the surface, through orogeny or volcanic resurfacing.[2] Thus, if the carbonate-silicate cycle is viewed as a prerequisite for long-term habitability, it would appear as though the existence of plate tectonics is rendered significant since it constitutes a crucial component of this cycle. However, proceeding with this premise raises an immediate question: How important is the carbonate-silicate cycle itself?

First, suppose that we consider worlds with only oceans and no landmasses, which we shall return to in Section 5.5. In the absence of weathering or outgassing, one may therefore arrive at the conclusion that such worlds ought to be uninhabitable because the stabilizing carbonate-silicate cycle is not operational. Indeed, this line of reasoning was supported by many numerical studies in the 2000s, but more recent publications reversed this trend. For example, Kite and Ford (2018) developed a comprehensive model that took carbon partitioning and atmosphere-lithosphere interactions into account and determined that ocean worlds are habitable over Gyr timescales. The chief reason behind this result was the suppression of geochemical cycles (viz., cycle-independent habitability) due to the curtailment of carbon exchange between the mantle and oceans. Apart from the geochemical aspects, the planetary mass also plays a prominent role in regulating long-term habitability: numerical modeling by Arnscheidt et al. (2019) suggests that worlds with $\lesssim 2.7 \times 10^{-2}\, M_\oplus$ are unlikely to sustain habitable climates over Gyr timescales.

2. Orogeny involves the crumpling (folding and deformation) of continental plates that leads to the generation of mountain ranges.

Second, let us turn our attention toward worlds with stagnant-lid tectonics; if the Solar system and theoretical projections comprise viable sources of data, stagnant-lid worlds are probably much more prevalent than those with plate tectonics. In light of the preceding discussion regarding the carbonate-silicate cycle, stagnant-lid worlds do not seem very promising when it comes to long-term habitability, insofar as maintaining stable temperatures is concerned. However, state-of-the-art numerical models that incorporate thermal evolution of the mantle, volcanic outgassing, and climate favor the notion that worlds with stagnant-lid tectonics are habitable even up to Gyr timescales. In particular, numerical simulations performed by Foley and Smye (2018) indicate that Earth-sized worlds with initial rates of radiogenic heating (i.e., heat supplied by the decay of radioactive isotopes) that are ~ 1 to 4 times higher than the Earth's rate and total carbon dioxide budgets that are ~ 0.01 to 1 that of the Earth's inventory can remain habitable for approximately 1 to 5 Gyr. Looking beyond plate tectonics, the concentration of radionuclides exerts a multifaceted influence on habitability (Lugaro et al. 2018), as seen in Fig. 5.3.

Apart from the abundance of radioactive isotopes, the composition of the interior and surface—in particular, the inventory of H_2O at these locations—and the temporal evolution of the stellar luminosity play a major role in determining long-term habitability of stagnant-lid worlds. Numerical models indicate that habitability over Gyr timescales is feasible provided that these worlds possess adequate initial concentrations of H_2O and appropriate outgassing rates (Godolt et al. 2019).

Third, the dichotomy between stagnant-lid and mobile-lid plate tectonics is probably simplistic. Instead, due to the intrinsically heterogeneous nature of volcanism as well as the lithosphere, it seems plausible that some regions could be stagnant at a given moment in time while others are active (Weller & Kiefer 2020). Last, other sources of energy such as tidal heating can drive volcanic outgassing and thereby pave the way for stable climate over long timescales, even in the absence of conventional plate tectonics (Valencia et al. 2018; Zanazzi & Triaud 2019).

Plate tectonics may also influence habitability through other channels, delineated below (see Figure 5.3). For starters, the subduction of water-bearing rocks into the mantle could lead to the mantle's cooling, thus sustaining the core-mantle temperature gradient, core convection, and the planetary dynamo (B. J. Foley & Driscoll 2016). As we saw in Section 4.2, it is conceivable that magnetic fields generated by the dynamo can shield

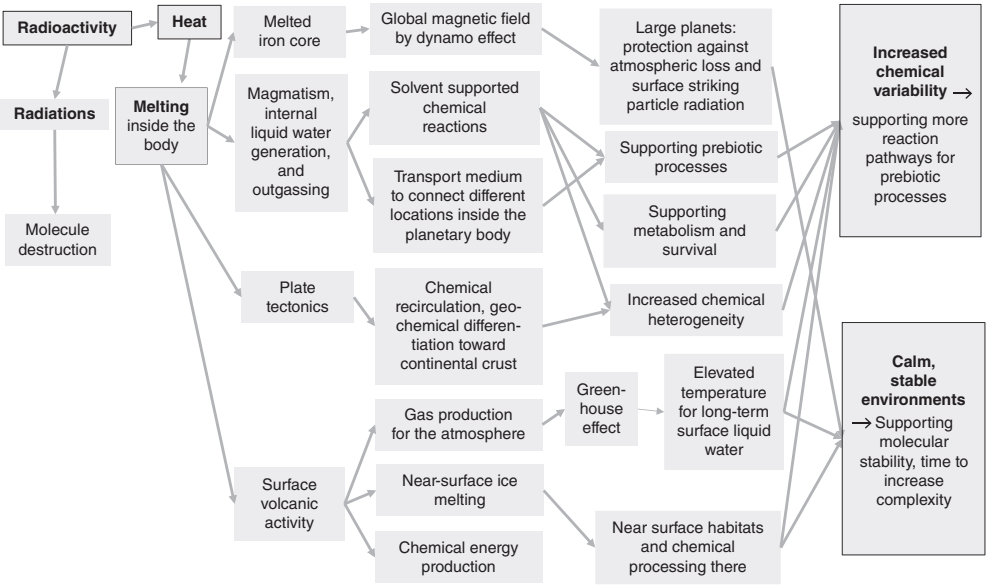


Figure 5.3 Overview of the complex array of interactions between plate tectonics and other factors that act concomitantly to regulate habitability. (© The Authors. CC-BY-NC-ND. *Source:* M. Lugaro, U. Ott, and Á. Kereszturi [2018], Radioactive nuclei from cosmochronology to habitability, *Progress in Particle and Nuclear Physics* 102: 1–47, fig. 6.)

the planet from the dual effects of high-energy radiation and atmospheric escape, although many uncertainties exist regarding the latter. Water delivered to the mantle during subduction also facilitates hydrous melting to produce felsic rocks (e.g., granites) that largely constitute the basis of continents. It is essential, however, to recognize that continental crust might be formed even in the absence of plate tectonics.

The production of granite during tectonic activity, in particular, has several important ramifications. First, not all rock types are equally rich in bioessential elements such as phosphorus. For instance, mid-ocean ridge basalt and gabbro that are widespread in the oceanic crust are depleted not only in phosphorus but also in other elements (e.g., potassium) employed by organisms on Earth. Hence, it has been theorized that the supply of nutrients from continental landmasses is $\sim 10^6$ times higher compared to that provided by mid-ocean ridges (sites where new oceanic crust is generated) on Earth (Maruyama et al. 2013). It is conceivable that worlds with

stagnant-lid tectonics will have relatively homogeneous surface composition that is dominated by mafic rocks akin to Earth's oceanic crust; modern Mars is an excellent example in this context (Ehlmann & Edwards 2014). Therefore, on account of the above reasons, such worlds may have a paucity of nutrients.

Furthermore, quartz is possibly widespread in sediments on worlds with plate tectonics because the granites duly formed are subsequently eroded by water, winds, and glaciers leading to the dispersal of quartz grains. An interesting consequence of quartz-rich sediments (partly translucent) is that the flux of photosynthetically active radiation (PAR) penetrating to a particular depth is much higher compared to their volcanogenic (mostly opaque) counterparts (Parnell 2004). To be specific, at the depth of 2×10^{-3} m, the PAR flux in quartz-rich soils is about two orders of magnitude higher than in the latter. Hence, the presence of quartz boosts the chances for putative microbes to survive in sediments at depths where they are shielded from ultraviolet radiation but still receive enough PAR for photosynthesis to take place.

Looking beyond quartz, plate tectonics played a major role in enhancing the diversity of minerals on Earth. More precisely, it could have given rise to ~ 500 minerals that had not existed until the onset of plate tectonics (Hazen & Ferry 2010). Apart from the genesis of new minerals, plate tectonics probably assisted in the spatiotemporal redistribution of existent minerals. In particular, the concentration of radioactive minerals in local environments would have created promising sites for the synthesis of prebiotic molecules. Thus, in view of the manifold benefits accorded by minerals in the origin and evolution of life (see Section 2.4.3), it seems plausible to surmise that worlds with plate tectonics have an advantage in this regard.

Finally, plate tectonics is anticipated to directly affect biological evolution itself in many ways (R. J. Stern 2016). To begin with, the extent of continental shelves (submerged continental regions) is proportional to the number of continents. Hence, the presence of plate tectonics can enable the fragmentation of supercontinents, in turn increasing the aggregate of continental shelves. These environments are particularly important, since they are distinguished by relatively shallow waters, thereby serving as bridges between oceans and continents that permit aquatic organisms to adapt and settle on land. Plate tectonics also causes the dispersal and separation of landmasses and oceans, drives the formation of new mountain ranges, and brings about the reconnection of landmasses and oceans. Collectively,

these changes in the lithosphere regulate the rates of speciation (Valentine & Moores 1970), with plate tectonics potentially amplifying the species richness after fragmentation and conserving it post-recombination; this hypothesis has gained recent corroboration through the analysis of marine invertebrate fossils (Zaffos et al. 2017). Therefore, viewed in toto, there are compelling grounds to argue that worlds with plate tectonics will be more habitable compared to those with stagnant-lid tectonics.

5.2.2 The likelihood of plate tectonics on exoplanets

Even if plate tectonics does not constitute a strict necessity for microbial or complex life per se, we have seen that there are manifold positives originating from its existence. It is natural to ask ourselves what are the factors that influence the likelihood of plate tectonics on other worlds. There is, unfortunately, no universal answer due to the complex dynamical nature of plate tectonics and the large number of physical variables involved.

The first, and foremost, parameter that comes to mind is the mass (M) of the world. One can equivalently work with its radius (R) since they are correlated via the mass-radius relationship $M \propto R^{3.7}$ for approximately Earth-sized rocky planets (or moons). The issue of whether planets more massive than the Earth (super-Earths) have a higher or lower relative likelihood of plate tectonics has proven to be contentious. Several studies indicate that the combination of thinner plates and higher shear stresses are likely to boost the prospects for plate tectonics for super-Earths. On the other hand, other models have arrived at the opposite conclusion—namely, super-Earths are likely to have either stagnant-lid or episodic tectonics. The reason proposed for this behavior has to do with the decline in the ratio of driving to resisting stresses. Mantle convection is responsible for driving stresses that enable plate deformation, whereas the resistive stresses quantify the resistance to deformation offered by lithospheric plates. Typically, this ratio must exceed unity in order for plate tectonics to function effectively.

If the mass of the world is too high, the pressures in the mantle increase commensurately. As a consequence, the mantle viscosity is significantly escalated, thereby potentially suppressing plate tectonics. For worlds smaller than the Earth, it is conventionally presumed that their heat budgets are also comparatively lower, accordingly giving rise to early cooling that causes

the solidification of the crust and stagnant-lid tectonics. Thus, in balance, numerical models appear to suggest that worlds in the range $M \approx 1$ to $5\,M_{\oplus}$ are optimal for plate tectonics (Cockell et al. 2016). In what follows, we will describe the scalings for basic physical processes associated with plate tectonics by mirroring the approach of Valencia and O'Connell (2009). Our choice of this model is motivated by its comparative analytical simplicity, but the ensuing results that favor a higher likelihood of plate tectonics for super-Earths should not be viewed as being definitive.

The horizontal shear stress below the lithosphere arising from mantle convection, denoted by P_{xy}, has the following dependence:

$$P_{xy} \propto \left(\eta \rho g \mathcal{Q}_s\right)^{1/4}, \tag{5.15}$$

where η is the mantle viscosity, ρ and g are the average density and surface gravity of the planet, and $\mathcal{Q}_s$ signifies the radiogenic heat flux at the surface. Along the same lines, the thickness of the plate (δ_{pl}) obeys

$$\delta_{pl} \propto \left(\frac{\rho g \mathcal{Q}_s}{\eta}\right)^{-1/4}, \tag{5.16}$$

provided that the boundary condition used for calculating the thickness of the boundary layer is set by the heat flux. Next, the physical variables inherent in the above formulae scale with the mass M as follows, based on numerical simulations of the planetary interior structure: $\rho \propto M^{0.196}$, $g \propto M^{0.503}$, $\mathcal{Q}_s \propto M^{0.476}$, and $\eta \propto M^{-0.64}$. Substituting these relations into (5.15) and (5.16), we obtain $P_{xy} \propto M^{0.27}$ and $\delta_{pl} \propto M^{-0.45}$, leading us to

$$\frac{P_{xy}}{\delta_{pl}} \sim M^{0.72}. \tag{5.17}$$

Note that the power-law exponent on the right-hand side is positive. This feature matters more than the exact magnitude, which can vary depending on the ansatzen used for the viscosity and heat flux. For our purposes, it suffices to say that the ratio of driving to resistive stresses (denoted by SR) has the functional form

$$\mathrm{SR} = \frac{P_{xy}/\delta_{pl}}{C_1 + C_2 P_{xy}/\delta_{pl}}, \tag{5.18}$$

where C_1 and C_2 are positive mass-independent parameters; the details concerning the derivation are furnished in Valencia and O'Connell (2009). By combining (5.18) and (5.17), we see that the ratio increases monotonically with the mass. Thus, as per this model, worlds more massive than the Earth exhibit a higher propensity toward plate tectonics.

Looking beyond the planet's mass, a number of other factors have an equal or possibly greater influence on the likelihood of plate tectonics. The inventory of water in the crust and mantle is particularly important since it can make the lithosphere more susceptible to deformation. The age of the planet also matters because the radiogenic heat flux declines over time and regulates the ratio of driving to resistive stresses, as seen from the simplified analysis in the previous paragraph. Although the driving stresses due to mantle convection are typically dominant, tidal gravitational forces (explicated shortly hereafter) may generate lithospheric stresses that are comparable to the former in the case of planets in the HZ of late-type M-dwarfs with $M_\star \lesssim 0.2 M_\odot$ (cf. Zanazzi & Triaud 2019).

Other physical variables that play a crucial role in plate tectonics include the surface temperature as well as both the magnitude and functional form of the mantle viscosity, thermal conductivity, and expansivity. Stamenković and Seager (2016) presented a suite of numerical simulations that varied the mass, composition, radiogenic heat, and core size of the planet. Broadly speaking, they found that (1) increasing the iron content in the mantle decreased the chances for plate tectonics, (2) decreasing the initial concentration of radiogenic isotopes by a factor of $\lesssim 10$ relative to Earth enhanced the prospects for plate tectonics, (3) increasing the size of the metallic core improved the likelihood of plate tectonics, and (4) Earthlike planets with masses $\sim 1\, M_\oplus$ were more favorable for plate tectonics than those with higher masses, with all other factors held constant.

The difficulties in assessing tectonic activity on other worlds are only further compounded by the putative existence of multiple stable tectonic modes (Lenardic et al. 2016). In particular, stagnant-lid and plate tectonics are capable of operating on the same world at different stages during its temporal evolution. As a corollary, two bodies that are virtually similar in terms of their physical characteristics could nonetheless evolve along different paths. On account of all the complexities encountered until now, the question of whether plate tectonics is likely to manifest on a specific world

may not have a binary yes-no answer. Directly assessing the presence of plate tectonics on other worlds will be highly challenging, but measurements of the planet's magnetic field might enable us to indirectly determine whether that planet has plate tectonics.

5.3 TIDAL LOCKING AND ITS CONSEQUENCES

Ever since the pioneering studies by the likes of Su-Shu Huang and Stephen H. Dole in the 1950s and 1960s (Huang 1960; Dole 1964), it has been appreciated that tidal locking might play a crucial role in determining the long-term habitability of worlds. Before we discuss tidal locking, it is worth examining the tidal force first. The tidal force originates from the variation in gravitational force across the object under consideration. We will sketch a heuristic derivation for a planet with orbital radius a around a star with mass $M_\star$. In this event, the gravitational acceleration ($\vec{a}_{in}$) experienced by the side of the planet facing toward the star is

$$\vec{a}_{in} = -\hat{r}\frac{GM_\star}{(a-R)^2}, \tag{5.19}$$

where R is the planet's radius and $\hat{r}$ captures the direction along which the force acts; we do not include the tangential component in our analysis for simplicity. In the same spirit, the gravitational acceleration ($\vec{a}_{out}$) experienced by the side of the planet facing away from the star becomes

$$\vec{a}_{out} = -\hat{r}\frac{GM_\star}{(a+R)^2}, \tag{5.20}$$

and the difference in the two gravitational accelerations embodies the tidal force, or equivalently, its acceleration ($\vec{a}_{\rm tide}$). Thus, by taking the difference of the preceding two equations, we arrive at

$$|\vec{a}_{\rm tide}| = |\vec{a}_{in} - \vec{a}_{out}| \approx \frac{2GM_\star R}{a^3}, \tag{5.21}$$

implying that $a_{\rm tide}$ falls off as the inverse *cube* of the distance.

5.3.1 Timescales associated with tidal interactions

Now, let us turn our attention to tidal locking: the reader should peruse Murray and Dermott (1999) for a comprehensive overview of this subject. The stresses exerted by the gravitational field gradient (i.e., tidal force) distort the hitherto spherical planet into an ellipsoidal shape. If the spin and orbital angular velocities differ from each other, the planet is continuously deformed by friction from tidal oscillations and dissipates spin energy as a consequence. This spin-down continues, in principle, until the spin and orbital angular velocities become equal to each other (1:1 resonance): a state known as synchronous rotation.

There are, however, a couple of caveats that must be noted at this juncture. First, planets with moderately high orbital eccentricity may exist in other resonance states, such as Mercury with its 3:2 resonance; in other words, it rotates three times for every two orbits around the Sun. Second, a number of factors are currently being identified that can drive the planet into a state of asynchronous rotation. Some of the notable ones reviewed in Lingam and Loeb (2019d) include (1) the existence of a 1 bar atmosphere for stars with $M_\star \gtrsim 0.5\,M_\odot$, (2) sufficient triaxial deformation (i.e., variation in the three principal components of the planet's moment of inertia tensor), (3) semiliquid or semimolten planetary interiors unlike the Earth's solid mantle, and (4) the presence of a nearby companion that is close to orbital resonance with the planet.[3]

With the above caveats in mind, the characteristic timescale required for the planet to become tidally locked ($t_{\rm lock}$) is adopted from Murray and Dermott (1999) as follows:

$$t_{\rm lock} \sim 3 \times 10^{11}\ {\rm yrs} \left(\frac{Q_0/k_2}{1000}\right)\left(\frac{\rho}{\rho_\oplus}\right)\left(\frac{\Omega_i}{\Omega_\oplus}\right)\left(\frac{a}{1\ {\rm AU}}\right)^6 \left(\frac{M_\star}{M_\odot}\right)^{-2}, \tag{5.22}$$

where ρ and Ω_i represent the mean density and initial spin angular velocity of the planet, while $\rho_\oplus$ and $\Omega_\oplus$ are the corresponding values for present-day Earth. We can use $\rho/\rho_\oplus \sim (R/R_\oplus)^{0.7}$ for rocky Earth-sized worlds to rewrite (5.22) in terms of R. The parameters Q_0 and k_2 are subject to much more uncertainty, and we have opted to normalize the ratio Q_0/k_2 by

3. Two bodies are regarded as being in orbital resonance with each other when the ratio of their orbital periods is nearly a small integer.

its approximate value for Earth (Pierrehumbert & Hammond 2019). The tidal quality factor Q_0 is a measure of the energy dissipated and depends on the elastic properties of the planetary material, while k_2 is known as the Love number and quantifies the rigidity of the planet. If we specialize to rocky planets that receive stellar fluxes similar to our planet (Earth-analogs), substituting (4.5) into (5.22), we obtain

$$t_{\rm lock} \sim 3 \times 10^{11}\ {\rm yrs} \left(\frac{Q_0/k_2}{1000}\right) \left(\frac{R}{R_\oplus}\right)^{0.7} \left(\frac{\Omega_i}{\Omega_\oplus}\right) \left(\frac{M_\star}{M_\odot}\right)^{7}. \tag{5.23}$$

Hence, it is apparent that the tidal locking time is strongly dependent on the stellar mass $M_\star$. Suppose, for instance, that we seek to impose a tidal locking timescale of ≤ 1 Gyr for Earth-analogs. Using the above formula, we end up with $M_\star \lesssim 0.44\, M_\odot$. However, in view of parameters that are poorly constrained (namely, Ω_i, Q_0, and k_2), it is conceivable that stars with masses above this threshold can also host tidally locked planets.

Another crucial effect arising from tidal forces is that the orbital eccentricity decays over time due to energy dissipation inside the planet. The characteristic timescale for the orbit to circularize ($t_{\rm circ}$) and achieve an eccentricity close to zero is given by

$$t_{\rm circ} \sim 2.2 \times 10^{16}\ {\rm yrs} \left(\frac{Q_0}{100}\right) \left(\frac{R}{R_\oplus}\right)^{-5} \left(\frac{M}{M_\oplus}\right) \left(\frac{a}{1\ {\rm AU}}\right)^{13/2} \left(\frac{M_\star}{M_\odot}\right)^{-3/2}. \tag{5.24}$$

To simplify this timescale, we make use of the mass-radius relationship $M \propto R^{3.7}$ for rocky planets and focus on worlds that receive similar stellar fluxes as that of the Earth. Therefore, by making use of (4.5), we end up with

$$t_{\rm circ} \sim 2.2 \times 10^{16}\ {\rm yrs} \left(\frac{Q_0}{100}\right) \left(\frac{R}{R_\oplus}\right)^{-1.3} \left(\frac{M_\star}{M_\odot}\right)^{8.25}. \tag{5.25}$$

Holding all factors except the stellar mass fixed, we can determine the cut-off at which the circularization timescale drops below 1 Gyr. From (5.25), we find that $M_\star \lesssim 0.13\ M_\odot$ leads to sub-Gyr timescales for the orbit to circularize. This result is in agreement with state-of-the-art numerical models of tidal interactions (R. Barnes 2017) that also predict that the orbits of Earth-analogs around late-type M-dwarfs could have circularized within ~ 1 Gyr.

5.3.2 Implications of tidally locked planets

Our preceding analysis has illustrated that planets around M-dwarfs are likely to exist in a state of synchronous rotation in a span of 1 Gyr or thereabouts. In this event, along the same lines as what we witness for the Moon, one side of the planet will always face the star (dayside), whereas the other will perpetually face away from the star (nightside). In early studies of exoplanets, this feature was regarded as a critical impediment to habitability for a multitude of reasons.

It was initially presumed that the permanent dayside would be too hot for life whereas the cold temperatures at the permanent nightside would result in the freeze-out of both the atmosphere and the oceans. The existence of an atmosphere, among many other reasons, is necessary for exerting a finite pressure and thereby ensuring the presence of liquid water on the surface. Thus, in the absence of an atmosphere, the likelihood of surface-based life-as-we-know-it becomes nearly zero. In addition, worlds without atmospheres will also receive much higher doses of ultraviolet radiation and cosmic rays at the surface. Collectively, these putative reasons were taken to imply that tidally locked planets around M-dwarfs would lack atmospheres, evince high temperature contrasts, and therefore prove inimical to surface-based life.

Next, due to the near-total absence of starlight on the nightside of synchronous rotators, the prospects for photosynthesis-based biospheres also diminish accordingly. As photoautotrophs constitute the bedrock of food webs on Earth, it is plausible that the biomass densities on the nightside will be minimal. Thus, insofar as photosynthesis is concerned, the surface area suitable for hosting complex biospheres could be reduced by approximately 50 percent. When we tackle the issue of how landmasses and oceans influence the properties of biospheres in Section 5.5, we are therefore dealing with idealized worlds wherein most of the surface area is habitable in terms of reactants, temperature, and energy. Tidal locking also influences habitability indirectly by slowing down the planet's spin. As noted in Section 4.2.1, a lower value of the spin angular velocity may translate to a weaker planetary magnetic field, since most dynamo models predict that the latter increases monotonically with respect to the rotation rate. The strength of the magnetic field, in turn, influences the flux of cosmic rays reaching the surface as well as the atmospheric escape rate (Lingam & Loeb 2019d).

On the basis of the consequences highlighted above, it is apparent that the temperature difference is important from the standpoint of both the

thermal limits on life and the prevention of condensation of the atmosphere. Fortunately, state-of-the-art numerical models have revealed that even a modest atmosphere comprising $\gtrsim 0.1$ bar of CO_2 allows for the efficient redistribution of heat and reduction of the temperature contrast between the dayside and nightside (Joshi et al. 1997; Wordsworth 2015). Oceans provide yet another avenue by which heat is efficiently transported across the planet. Detailed overviews of the progress made in modeling the climate of tidally locked exoplanets can be found in Shields et al. (2016) and Pierrehumbert and Hammond (2019). Therefore, as per our current understanding of this subject, it appears as though tidal locking will not necessarily pose as much of a threat, in terms of generating large thermal gradients, as originally feared.

Most of the state-of-the-art climate models employed in the literature for studying tidally locked planets fall under the category of General Circulation Models (GCMs), whose results require some unpacking. Instead, we will reproduce the simple analytical model derived by Wordsworth (2015), which is consistent with GCM results. In this model, the atmosphere is transparent in the visible range, but is "gray" in the infrared—that is, it has a constant absorption coefficient. The stellar flux incident at the top of the atmosphere is $S_\star = L_\star / \left(4\pi a^2\right)$. The *local* energy flux balance at the planetary surface is given by

$$\sigma T^4 = \left(1 - A_p\right) S_\star \cos\theta_z + \mathrm{GLR}, \tag{5.26}$$

where θ_z denotes the zenith angle—that is, the angle between the zenith (the point directly overhead) at a given point and the line of sight to the center of the star. GLR quantifies the infrared radiative flux contribution from the atmosphere. In principle, a third term ought to be introduced in the right-hand side that accounts for the turbulent transport of heat by winds. The analogous atmospheric energy balance condition after vertical integration is expressible as

$$\mathrm{GLR} + \mathrm{OAR} = \mathcal{A}_{\mathrm{IR}} \sigma T^4, \tag{5.27}$$

with $\mathcal{A}_{\mathrm{IR}}$ representing the frequency-averaged absorbance of the atmosphere in the infrared. OAR signifies the infrared radiation emitted by the atmosphere into outer space. In the above expression, we have omitted an extra term on the right-hand side since it does not change our results after

horizontally averaging the equations. The role of wind transport has also been neglected, as stated above. We may now introduce the horizontally averaged quantities,

$$\mathcal{B}_{ds} = \frac{\int_{ds} \sigma T^4 \, dA'}{2\pi R^2} \tag{5.28}$$

and

$$\mathcal{B}_{ns} = \frac{\int_{ns} \sigma T^4 \, dA'}{2\pi R^2}, \tag{5.29}$$

where the subscripts *ds* and *ns* refer to the dayside and nightside, respectively; note that $\int dA'$ represents the surface integral over the appropriate region. Upon calculating the surface integral of (5.26) over the dayside, we end up with

$$\mathcal{B}_{ds} = \frac{1}{2}\left(1 - A_p\right) S_\star + \mathrm{GLR}, \tag{5.30}$$

where the geometric factor of 1/2 follows from the fact that the stellar flux intercepted over a cross-sectional area of πR^2 is distributed over the dayside hemisphere with area $2\pi R^2$. Similarly, repeating the above calculation for the nightside, we obtain

$$\mathcal{B}_{ns} = \mathrm{GLR}, \tag{5.31}$$

and there is no contribution from the stellar flux since the nightside does not receive any stellar radiation. We observe that the local and global values of the GLR as well as the OAR are the same since the atmosphere is assumed to be isothermal. Finally, we average (5.27) over the entire planet, thereby yielding

$$\mathrm{GLR} + \mathrm{OAR} = \frac{1}{2}\mathcal{A}_{\mathrm{IR}}\mathcal{B}_{ds} + \frac{1}{2}\mathcal{A}_{\mathrm{IR}}\mathcal{B}_{ns}. \tag{5.32}$$

In the limit where the atmosphere is optically thin—that is, photons can pass through the atmosphere with a low probability of being scattered or absorbed—the above equations are simplified considerably since $\mathcal{A}_{\mathrm{IR}} \approx \tau_{LW}$ and $\mathrm{OAR} \approx \mathrm{GLR} \approx \tau_{LW}\mathcal{B}_a$. As the derivation is somewhat intricate, we refer the reader to Pierrehumbert (2010) for more details. The total optical depth τ_{LW} is defined as follows:

$$\tau_{LW} \approx \frac{\kappa P_s}{g} \approx 200 \left(\frac{P_s}{1\,\mathrm{bar}}\right)\left(\frac{\kappa}{2 \times 10^{-2}\,\mathrm{m^2\,kg^{-1}}}\right)\left(\frac{g}{g_\oplus}\right)^{-1}. \tag{5.33}$$

Here, P_s and κ represent the surface pressure and the absorption coefficient, respectively. By employing the above simplifications, it is easy to verify that

$$\mathcal{B}_{ns} \approx \frac{\tau_{LW}\left(1-A_p\right) S_\star}{8}, \tag{5.34}$$

$$\mathcal{B}_{ds} \approx \frac{\left(1-A_p\right) S_\star}{2}, \tag{5.35}$$

and

$$\mathcal{B}_{a} \approx \frac{\left(1-A_p\right) S_\star}{8}, \tag{5.36}$$

where we have also used $\tau_{LW} \ll 1$. We can estimate the corresponding temperatures through the relations: $\mathcal{B}_{ns} \equiv \sigma\, T_{ns}^4$, $\mathcal{B}_{ds} \equiv \sigma\, T_{ds}^4$, and $\mathcal{B}_{a} \equiv \sigma\, T_{a}^4$. Note that T_{ns}, T_{ds}, and T_a are the nightside, dayside, and atmospheric temperatures, respectively. Next, we make use of (5.9) in conjunction with the definition of $S_\star$. Therefore, we end up with

$$T_{ns} \approx T_{\rm eq}\left(\frac{\tau_{LW}}{2}\right)^{1/4}, \tag{5.37}$$

$$T_{ds} \approx 2^{1/4} T_{\rm eq}, \tag{5.38}$$

and

$$T_{a} \approx 2^{-1/4} T_{\rm eq}. \tag{5.39}$$

By specifying $T_{\rm eq} \approx 255$ K for an Earthlike planet, it is easy to calculate the nightside temperature as a function of τ_{LW}, or equivalently the surface pressure from (5.33). However, the above formulae cannot be applied arbitrarily since they break down for $\tau_{LW} \gtrsim 1$. Koll and Abbot (2016) derived a theoretical model that is valid across the entire range of τ_{LW}, but we shall not provide the explicit expressions on account of their complexity. We have, instead, presented the dayside and nightside temperatures as functions of the optical depths for this model in Figure 5.4. If one supposes an Earthlike value of $T_{\rm eq}$, it is found that both the dayside and nightside temperatures exceed the lower thermal limit of ~ 250 K when the condition $\tau_{LW} \gtrsim 1$ is

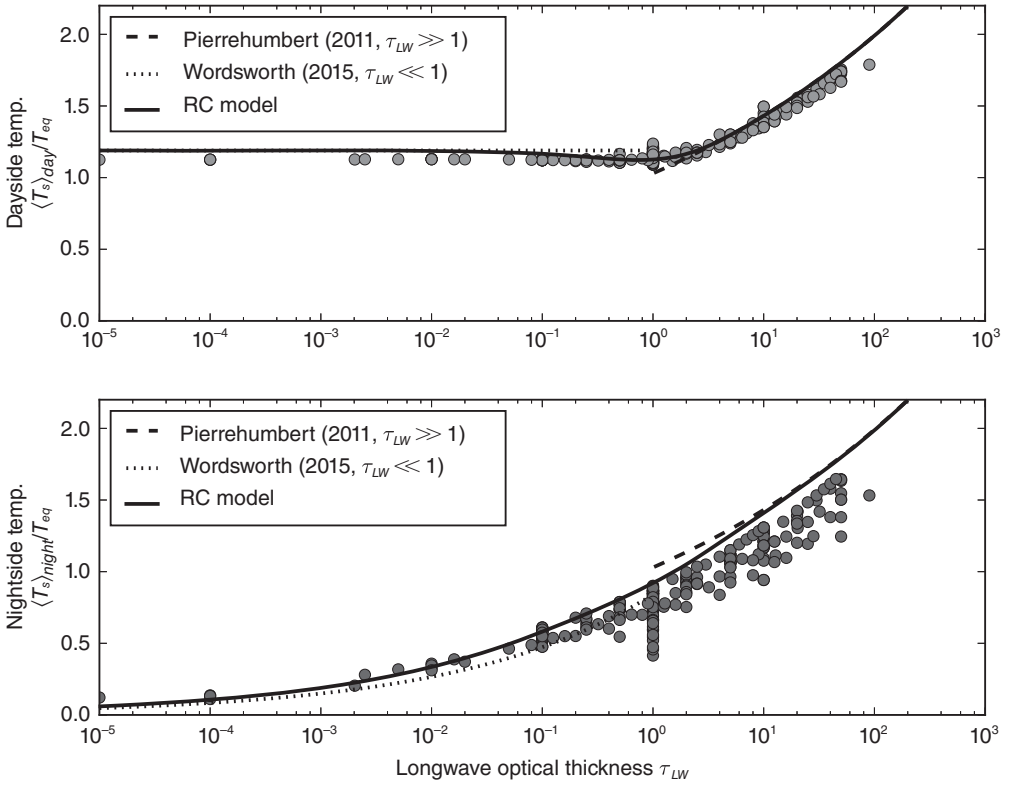


Figure 5.4 *Top:* Average dayside surface temperature normalized by the planet's effective temperature. *Bottom:* Average nightside temperature relative to the effective temperature of the planet. The circles correspond to temperatures computed from an ensemble of GCM simulations. The solid curve denotes the theoretical model, whereas the dashed and dotted curves are previously derived analytical models in the limits $\tau_{LW} \gg 1$ and $\tau_{LW} \ll 1$, respectively. (© The American Astronomical Society. *Source:* Daniel D. B. Koll and Dorian S. Abbot [2016], Temperature structure and atmospheric circulation of dry tidally locked rocky exoplanets, *Astrophysical Journal* 825[2]: 99, fig. 3.)

fulfilled. In this regime, the temperatures are also sufficiently high to ensure that the atmosphere does not condense on the nightside of the planet.

Hence, it is plausible that some tidally locked planets in the HZ will possess the necessary thermal conditions for habitability. We will return to the crucial issue of differentiating between the dayside-nightside temperature contrast on planets with and without an atmosphere or ocean via observations in Section 6.2.2.

5.4 ATMOSPHERIC COMPOSITION

It is self-evident that the atmospheric composition will influence the planet's habitability in a number of ways (Grenfell et al. 2020). For starters, let us contemplate the two most prominent gases in the Earth's atmosphere today: molecular nitrogen (N_2) and oxygen (O_2).

We covered the significance of O_2 in Section 3.4.1, owing to which a brief recap suffices. Oxygen levels influence the extent of the ozone layer that serves as a shield against ultraviolet radiation. The rise of oxygen also compelled organisms to evolve adaptations for mitigating the harm caused by reactive oxygen species (e.g., hydrogen peroxide). The partial pressure threshold for oxygen toxicity in humans is apparently on the order of $\lesssim 10$ bars (Crapo 1986). Most importantly, perhaps, the rise in O_2 levels may have permitted the diversification of complex multicellular organisms on Earth and probably elsewhere. The maximum size of aerobic organisms, sensu lato, is governed by the concentration of atmospheric O_2, as demonstrated in Section 3.4.1. Hence, it is plausible that worlds with elevated O_2 concentrations in the atmosphere and oceans compared to Earth, albeit only up to a certain threshold, might be "superhabitable" (Heller & Armstrong 2014).

Next, let us turn our attention to N_2. Nitrogen is one of the chief bioessential elements for life on Earth. For instance, it forms an essential component of amino acids through the amino group (-NH_2). Nitrogen fixation converts N_2 into compounds such as ammonia (NH_3), nitrites (NO_2^-), and nitrates (NO_3^-) that are suitable for uptake by organisms. In principle, the fixation of nitrogen can occur through both abiotic and biotic means. Setting aside the latter, abiotic channels include electrical discharge (lightning), stellar energetic particles, and asteroid or cometary impact events. In addition, we observe that nitrogen oxides formed via abiotic pathways are precipitated into oceans, where they might have served as electron acceptors for facilitating the origin of life near hydrothermal vents. Finally, nitrous oxide (N_2O), a potent greenhouse gas, is capable of raising the surface temperatures of planets that are otherwise too cold to host surficial liquid water.

Therefore, it is apparent that the abundance of atmospheric N_2 could influence the emergence and sustenance of extraterrestrial biospheres. As a matter of fact, it is worth remarking that the Earth's inventory of N_2 appears to have evolved significantly over time, from $\lesssim 0.5$ bar at $\gtrsim 2.7$ Ga to the present-day value of approximately 0.8 bar (Lammer et al. 2018). Note, however, that nitrogen is also sequestered in the planetary interior, primarily

in sedimentary rocks, and this repository must be taken into account when assessing biological habitability.

5.4.1 Carbon dioxide and complex life

Although carbon dioxide is a relatively minor component of Earth's atmosphere, it evidently plays an important role in the biosphere in multiple respects. Perhaps the best known among them is that it raises the surface temperature via the greenhouse effect. Of almost equal importance is the fact that photosynthesis relies on carbon dioxide as one of the reactants.

As we saw in Section 4.1, the concept of the classical HZ is founded on the premise that the atmosphere consists of CO_2 and H_2O as the sole greenhouse gases. Yet, we wish to highlight a couple of potential downsides associated with CO_2 in this context. In order to compensate for the lower stellar fluxes as one moves away from the star, it becomes necessary to raise the partial pressure of CO_2 to provide a higher degree of greenhouse warming, thereby ensuring that water can remain in the liquid phase. However, this trend toward higher CO_2 is dangerous for complex multicellular life on Earth comprising macroscopic organisms endowed with tissue differentiation and circulatory systems for the following reasons.

One of the most prominent among them is respiratory acidosis, which arises when the lungs become ineffective at expelling CO_2 from the body. In turn, this leads to high levels of carbon dioxide in the blood (known as hypercapnia), which makes it more acidic, thus lowering the pH; the same also applies to other bodily fluids. Consequently, increased acidity is expected to hamper respiration, transport of oxygen, and metabolic rates. In particular, elevated CO_2 concentrations are anticipated to adversely impact the hydrolysis of acetylcholine (a chemical used for interneuronal signaling), which regulates both cerebral circulation and respiration. Lastly, high atmospheric CO_2 levels translate to acidic oceans, due to the formation of carbonic acid that suppress the synthesis of protective shells and plates by aquatic organisms. In fact, the sum total of physiological and ecological stresses engendered by elevated CO_2 concentrations may have contributed to the devastating Permian-Triassic mass extinction (Knoll et al. 2007).

On Earth, the upper bound for the partial pressure of CO_2 for the sustenance of complex life is 5 to 50 mbar; note that 1 mbar is 10^{-3} bar. Apart from CO_2, we observe that carbon monoxide (CO) could also set

constraints on habitability. Carbon monoxide is especially dangerous for organisms with oxygen-carrying biomolecules such as myoglobin and hemoglobin because their affinity for CO is orders of magnitude higher than O_2. The ineffective transport of oxygen leads to metabolic breakdown in aerobic organisms that are dependent on steady oxygen supply. The critical CO threshold for humans is comparable to 10^{-4} bar. When it comes to planets around late-type M-dwarfs, two factors should be borne in mind regarding CO. First, the lifetime of CO molecules in the atmosphere is much longer due to the lower fluxes of UV-A radiation (315–400 nm). Second, under sufficiently oxic conditions, CO can accumulate in the atmosphere due to the burning of biomass, production by microbes, and photolysis of organic matter.

Hence, it is possible to combine the aforementioned constraints on complex life imposed by CO_2 and CO and compute new limits for the *complex life habitable zone* (CLHZ) that will be more stringent than those associated with the classical HZ. This analysis was carried out by Schwieterman, Reinhard, Olson, Harman, and Lyons (2019); the results are depicted in Figure 5.5. It is found that the CLHZ is merely 20 to 50 percent the width of the classical HZ for Solar-type stars. Second, the higher atmospheric CO concentrations stemming from the spectral properties of late-type M-dwarfs and close-in locations of their temperate planets seemingly suppress the prospects for complex life on these worlds. Notable examples in this category include Proxima Centauri b and the TRAPPIST-1 planets. If one adopts an upper bound of ~ 0.1 bar for the partial pressure of CO_2, the stellar flux (normalized by its value for the Earth) at the outer limit of CLHZ can be calculated from

$$S_{\rm eff} = S_{\rm eff,\odot} + \mathcal{A}\tilde{T}_\star + \mathcal{B}\tilde{T}_\star^2 + \mathcal{C}\tilde{T}_\star^3 + \mathcal{D}\tilde{T}_\star^4, \tag{5.40}$$

where $\tilde{T}_\star = T_\star - 5780$ K ($T_\star$ is the effective stellar temperature) and the appropriate parameters are as follows: $S_{\rm eff,\odot} = 0.6743$, $\mathcal{A} = 2.9876 \times 10^{-5}$, $\mathcal{B} = -4.6788 \times 10^{-9}$, $\mathcal{C} = -1.9425 \times 10^{-12}$, and $\mathcal{D} = -1.0263 \times 10^{-16}$. This expression has the same functional form as (4.8), but with different numerical values.

The preceding conclusions must be viewed with due caution since they are strongly anthropocentric—that is, based on the data for select macroscopic animals on Earth. Complex organisms on other worlds might have evolved suitable strategies to bypass CO_2 and CO poisoning. For example,

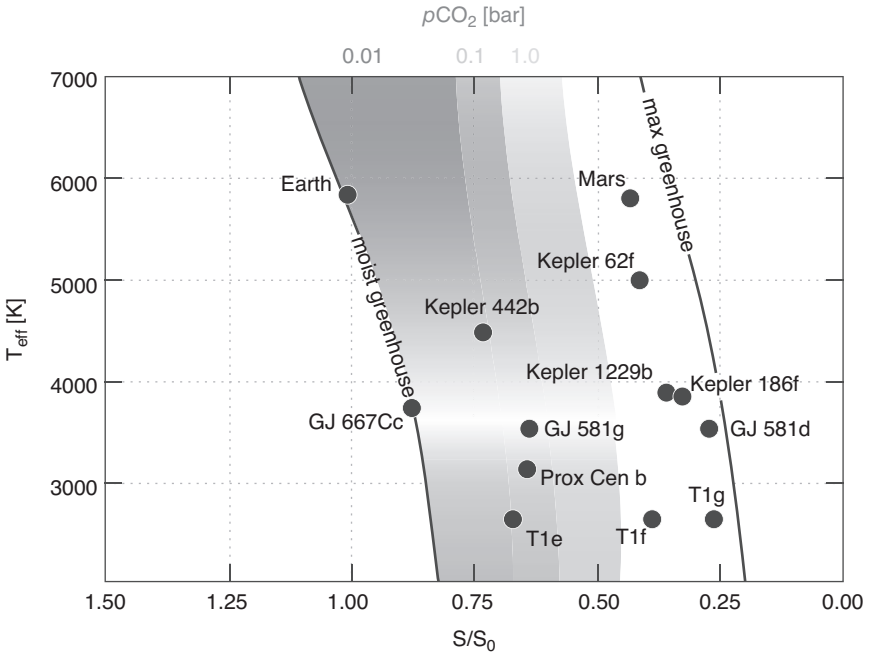


Figure 5.5 Boundaries for the HZ as functions of the normalized stellar flux incident on the planet and the stellar temperature. The moist and maximum greenhouse limits correspond to the limits of the classical HZ. In contrast, the dark-, intermediate-, and light-shaded regions represent the extent of the *complex life habitable zone,* assuming thresholds of 0.01 bar, 0.1 bar, and 1 bar, respectively. At effective stellar temperatures below 3200 K, the lifetime and concentrations of carbon monoxide are above the permissible limits for humans. Some of the famous terrestrial planets that lie within the classic HZ are also shown. (© The American Astronomical Society. CC-BY-3.0. *Source:* Edward W. Schwieterman, Christopher T. Reinhard, Stephanie L. Olson, Chester E. Harman, and Timothy W. Lyons [2019], A limited habitable zone for complex life, *Astrophysical Journal* 878[1]: 19, fig. 5.)

the algae *Cyanidium caldarium* have been shown to thrive in pure CO_2 atmospheres and highly acidic environments with pH as low as 0.5 (Rothschild & Mancinelli 2001). Similarly, neither algae nor plants apparently exhibit appreciable susceptibility to CO poisoning. Even if we specialize only to animals, oxygen-carrying proteins such as hemerythrin and hemocyanin (endowed with Fe and Cu atoms, respectively) found in some invertebrates do not manifest the high CO affinity characteristic of hemoglobin. Hence, it is not unreasonable to envision other worlds wherein complex organisms mirror the functionality of plants, fungi, algae, and certain animals.

5.4.2 The significance of hydrogen and methane

Although we shall focus our attention on molecular hydrogen (H_2), some of the ensuing statements are also applicable to methane (CH_4). Most notably, the inclusion of H_2 or CH_4 as atmospheric gases leads to an extension of the classical HZ, as seen in Section 4.1. The presence of rocky planets with hydrogen-rich atmospheres is plausible despite the fact that H_2 is capable of being rapidly depleted from the atmosphere on account of its low molecular weight; more precisely, the loss of H_2 can occur via Jeans escape, X-ray and UV photolysis, and stellar winds. Planets with hydrogen-dominated atmospheres could have accreted hydrogen during their formation from the stellar nebula or outgassed H_2 if they had formed from water-rich materials.

As mentioned earlier, the presence of H_2 permits greenhouse warming. Let us take a brief detour and consider the extreme limit wherein the planet is ejected into space. D. J. Stevenson (1999) argued that such worlds might host temperatures conducive to life in some instances. We will briefly sketch his reasoning below. First, the total radiogenic heating rate (in W) is conventionally held proportional to the mass of the planet; the Earth's radiogenic heating rate is around 2.2×10^{13} W. For this ansatz, the radiogenic heat flux at the surface is $\mathcal{Q}_s \propto M/R^2 \propto R^{1.7}$, with the last scaling following from the mass-radius relationship. The effective temperature of the planet can be calculated by equating $\sigma T_{\rm eq}^4$ with $\mathcal{Q}_s$. After simplification, we end up with

$$T_{\rm eq} \approx 29\,{\rm K} \left(\frac{R}{R_\oplus}\right)^{0.43}. \tag{5.41}$$

Next, the surface pressure P_s can be expressed in terms of the atmospheric mass $M_{\rm atm}$ as

$$P_s = \frac{M_{\rm atm} g}{4\pi R^2} \approx 1.15 \times 10^3\,{\rm bar} \left(\frac{M_{\rm atm}}{10^{-3} M_\oplus}\right) \left(\frac{R}{R_\oplus}\right)^{-0.3}, \tag{5.42}$$

where $M_\oplus$ is the mass of Earth. In order to calculate the surface temperature (T_s), Stevenson proposed that the adiabatic pressure-temperature relationship $T \propto P^{0.36}$ is an adequate approximation if the pressure falls within 1 bar and P_s. The phase diagram of H_2 reveals that the maximum pressure at which molecular hydrogen remains in the gaseous phase at (5.41) is ~ 1 bar. Moreover, for the atmospheric pressure (5.42), an optical depth close to unity is achieved when the pressure is around 1 bar. Thus, utilizing the

adiabatic relationship with the normalization roughly specified to be $P \sim 1$ bar and $T \sim T_{\rm eq}$, the surface temperature becomes

$$T_s \sim 366\,{\rm K} \left(\frac{R}{R_\oplus}\right)^{0.32} \left(\frac{M_{\rm atm}}{10^{-3} M_\oplus}\right)^{0.36}. \tag{5.43}$$

Hence, for worlds with pressures on the order of $\lesssim 10^3$ bar, we see that temperatures amenable to the existence of liquid water on the surface might be maintained. Since the model was constructed on the basis of radiogenic heating, it should be noted that the surface temperature will decline slowly over time. Although the typical surface pressure seems very high, Earth-based microbes have been documented to survive at transient pressures as high as $\sim 10^6$ bar, albeit with low survival fractions typically on the order of 10^{-2} to 10^{-3}.

Now, let us return to analyzing worlds with relatively thin ($\lesssim 30$ bar) hydrogen-rich (> 70 percent) atmospheres orbiting stars. In the past couple of chapters, we have seen the importance of photosynthesis from the standpoint of sustaining complex biospheres. It is therefore natural to ask ourselves if some variant of photosynthesis could function on worlds with hydrogen-dominated atmospheres. We will adopt the approach outlined in Bains et al. (2014) to address this matter. The analog of oxygenic photosynthesis on worlds with hydrogen-rich atmosphere was posited to be

$$CH_4 + H_2O + h\nu \xrightarrow{\text{pigments}} CH_2O + 2H_2, \tag{5.44}$$

with CH_4 constituting the equivalent of CO_2 in oxygenic photosynthesis; CH_2O is a stand-in for organic compounds. The product of (5.44) is molecular hydrogen, in the same fashion as oxygenic photosynthesis generates molecular oxygen. On account of this similarity, the above reaction was christened *hydrogenic photosynthesis*. Apart from hydrogenic photosynthesis, certain versions of anoxygenic photosynthesis on Earth may operate on these worlds. One of the most notable among them entails the use of Fe^{2+} as the reducing agent (electron donor) to synthesize organic compounds, with Fe^{3+} being the product. It is also conceivable that other photosynthetic pathways can exist that do not entail the formation of H_2. One such example is based on Cu^+ serving as an electron acceptor and being reduced to copper (Cu) as follows:

$$CH_4 + H_2O + 4Cu^+ + h\nu \xrightarrow{\text{pigments}} CH_2O + 4Cu + 4H^+. \quad (5.45)$$

Note that both (5.44) and (5.45) require methane as one of the reactants. There are two pathways by which the formation of methane can potentially occur. The first set of processes are abiotic in nature (Etiope & Sherwood Lollar 2013). Methane is a minor component of volcanic gases, but its abundance depends on the oxidation state of the mantle. Laboratory experiments and in situ studies indicate that abiotic methane is generated at substantial concentrations (up to 10^{-3} M in aqueous solutions) via high-temperature reactions involving magmatic vapors at $\gtrsim 700$ K. It has also been derived from low-temperature ($\lesssim 400$ K) gas-water-rock reactions cognate to the Fischer-Tropsch synthesis that yields hydrocarbons. The second avenue is biological in nature. Microbes analogous to methanogens on Earth are adept at reducing CO_2 to produce CH_4 because the net chemical reaction is highly favorable on thermodynamic grounds:

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O \quad (5.46)$$

Thus, this process could continue until the entire inventory of atmospheric CO_2 has been nearly exhausted and converted into methane.

There are a couple of advantages associated with hydrogenic photosynthesis. First, the energy required to synthesize a given quantity of biomass through this pathway is about five to ten times lower than that needed for generating an equivalent amount of biomass via oxygenic photosynthesis. Second, hydrogenic photosynthesis can be decomposed into two half-reactions, of which the oxidation of methane is expressible as

$$CH_4 + H_2O \rightarrow CH_2O + 4H^+ + 4e^-. \quad (5.47)$$

The energy required for this reaction to occur, after taking photosynthetic efficiency into account, is ~ 0.82 eV. It is easy to determine that the wavelength of a photon endowed with this energy is ~ 1.5 μm. In principle, photons with wavelengths smaller than this threshold should be capable of powering hydrogenic photosynthesis. In contrast, the corresponding photon wavelength for oxygenic photosynthesis, by adopting this line of reasoning, is ~ 1 μm, as demonstrated in Section 4.3.5. Hence, hydrogenic photosynthesis might function at longer wavelengths with respect to oxygenic

photosynthesis, thereby making it an attractive prospect for planets in the HZ of low-mass stars that receive more radiation in the infrared regime.

Two basic components are involved in hydrogenic photosynthesis: (1) capturing electromagnetic radiation to generate H_2 and (2) oxidation of methane to produce H_2. It so happens that these criteria are simultaneously fulfilled by certain microbes on Earth, although hydrogenic photosynthesis has not been documented. In the previous chapters, we highlighted the significance of O_2 for complex life. Ipso facto, worlds with hydrogen-rich atmospheres might have a much lower likelihood of evolving complex multicellular organisms that are motile and macroscopic ($\gtrsim 0.01$ m in size), but this conclusion is not definitive for the following reasons. It was recently established that some species from the animal phyla *Loricifera* and *Cnidaria* ostensibly inhabit permanently anoxic environments, although they are both minuscule ($\lesssim 10^{-3}$ m) and relatively inert. Moreover, organic compounds that release sufficient energy for metabolic activities after reduction with hydrogen may exist, analogous to the roles of carbohydrates and fats in aerobic respiration.

Evidently, not enough research has been undertaken in assessing the prospects for life in atmospheres dominated by H_2. Hence, a combination of laboratory experiments on microbial metabolism and theoretical analyses drawing on quantum chemistry are necessary in order to advance our understanding of this topic. While in vitro studies of methanogens in H_2-centric environments have been performed for decades, Seager et al. (2020) carried out experiments in 100 percent H_2 atmospheres. The authors found that, ceteris paribus, the steady-state concentration of the bacterium *Escherichia coli* in this setting was merely ~ 2 times lower than its value in air (i.e., the atmosphere of modern Earth). In contrast, the more complex organism *Saccharomyces cerevisiae* (a species of yeast) attained a maximal cell concentration that was more than two orders of magnitude smaller than the control experiments in air. Thus, while the study demonstrated the feasibility of Earth-based life surviving in H_2 atmospheres, provided that sufficient nutrients are available, it also underscored the necessity for further studies along similar lines.

5.4.3 Atmospheric hazes

Saturn's moon, Titan, is notable for its thick and hazy atmosphere. The *Cassini-Huygens* mission has enhanced our understanding of how these hazes are formed on Titan. The irradiation of the methane-nitrogen (N_2)

atmosphere with ultraviolet radiation and energetic particles leads to the formation of aerosol particles that agglomerate together and generate thick hazes. Theoretical models allied to laboratory experiments and geochemical proxies suggest that organic hazes also existed on early Earth, with the total aerosol production rate potentially being on the order of 10^{11} kg / yr (Trainer et al. 2006).

The extent of haze formation depends on a number of factors. The atmospheric composition is arguably the most dominant factor. Substantial atmospheric methane fluxes, on the order of 10^{15} molecules m^{-2} s^{-1}, are possibly required for the production of thick hazes. In addition, simulations indicate that a CH_4 to CO_2 ratio of approximately 0.2 might be necessary to initiate the synthesis of hazes. As most of the CH_4 fluxes on early or modern Earth were presumably generated via biological activity, hazes may perhaps require the preexistence of abundant microbial life. The inclusion of sulfur gases is predicted to lower the stringent requirements imposed on the CH_4 to CO_2 ratio. At higher concentrations of O_2, the production rate of hazes declines as per numerical models.

It is, however, not only planetary attributes that influence hazes but also stellar factors (Arney et al. 2017). The formation of hazes is dependent on the flux of UV radiation at wavelengths in the range $\sim$115–400 nm. As noted in Section 4.3, the total UV fluxes received by Earth-analogs vary considerably as a function of the stellar mass. F-type stars ($1 < M_\star/M_\odot < 1.4$) have high far-UV fluxes that yield oxygen species from photolysis. These species are responsible for the destruction of hydrocarbons, owing to which the likelihood of haze formation is lowered on planets in the HZ of F-type stars. A similar argument also applies to active M-dwarfs since stellar flares emit substantial fluxes of far-UV radiation. However, when it comes to quiescent M-dwarfs, haze formation may be initiated at CH_4 to CO_2 ratios that are smaller than 0.2.

Hazes are anticipated to impact the prospects for habitability in many ways. First, hazes absorb and scatter incoming stellar radiation. As a result, the lower regions of the atmosphere could undergo heating while the planetary surface is subject to cooling. The extent of cooling depends on the optical properties, size, and shape of the particles as well as the spectral classification of the host star. Let us consider M-dwarfs to illustrate our point further. Most of the radiation emitted by these stars is within 0.7 to 2.5 μm, and organic hazes are relatively transparent in this range. Therefore, ceteris paribus, one may expect the degree of cooling to be less pronounced on exoplanets around M-dwarfs as opposed to those orbiting G-type stars. The

second consequence of note originates from the fact that the fractal structure of aerosols comprising the haze renders them optically thick to ultraviolet radiation with respect to visible light; the ratio of the two optical depths is ~ 20 (Wolf & Toon 2010). In other words, the flux of UV radiation reaching the surface is attenuated because hazes constitute an effective UV shield. Thus, hazes can protect surface-based organisms from the harmful effects of UV radiation that were elucidated in Section 4.3.

5.5 THE EXTENT OF LANDMASSES AND OCEANS ON THE SURFACE

Earth's marbled surface, characterized by the juxtaposition of continents and oceans, has attracted copious artistic attention on account of its beauty. Our planet's surface topography is notable for an unusual coincidence: the fraction covered by landmasses (29 percent) is close to that spanned by water bodies, predominantly oceans (71 percent). Furthermore, these two fractions may have been comparable to one another for most of Earth's geological history (from ~ 3 Ga), as seen in Figure 5.6. This statement should, however, not be regarded as being definitive in light of the numerous uncertainties arising from the scarcity of robust data (Korenaga 2018). This coincidence is likely to be unusual since most worlds are anticipated to be largely dominated by deep oceans or arid landmasses (deserts) for reasons that will be delineated later.

The pivotal question that arises here is whether the coincidence between land and water fractions is merely random chance or attributable to observation-selection effects—that is, whether this particular feature optimized the emergence of intelligent and conscious observers (humans) in some respects (F. Simpson 2017). Several other coincidences have been subsequently interpreted as vital players in the emergence of hominins. Hence, it behooves us to investigate how the global properties of biospheres and their evolutionary trajectories are dependent on the fraction of the surface covered by land and water. In what follows, we will not tackle subsurface biospheres because they will be addressed in Chapter 7. We shall concern ourselves only with worlds where the surface of the ocean floor remains in direct contact with rocky material from the oceanic crust as they generally have a higher likelihood of being habitable (Noack et al. 2016).

Moreover, we shall restrict ourselves to analyzing planets with oceans and continents that are virtually identical to Earth in most functional

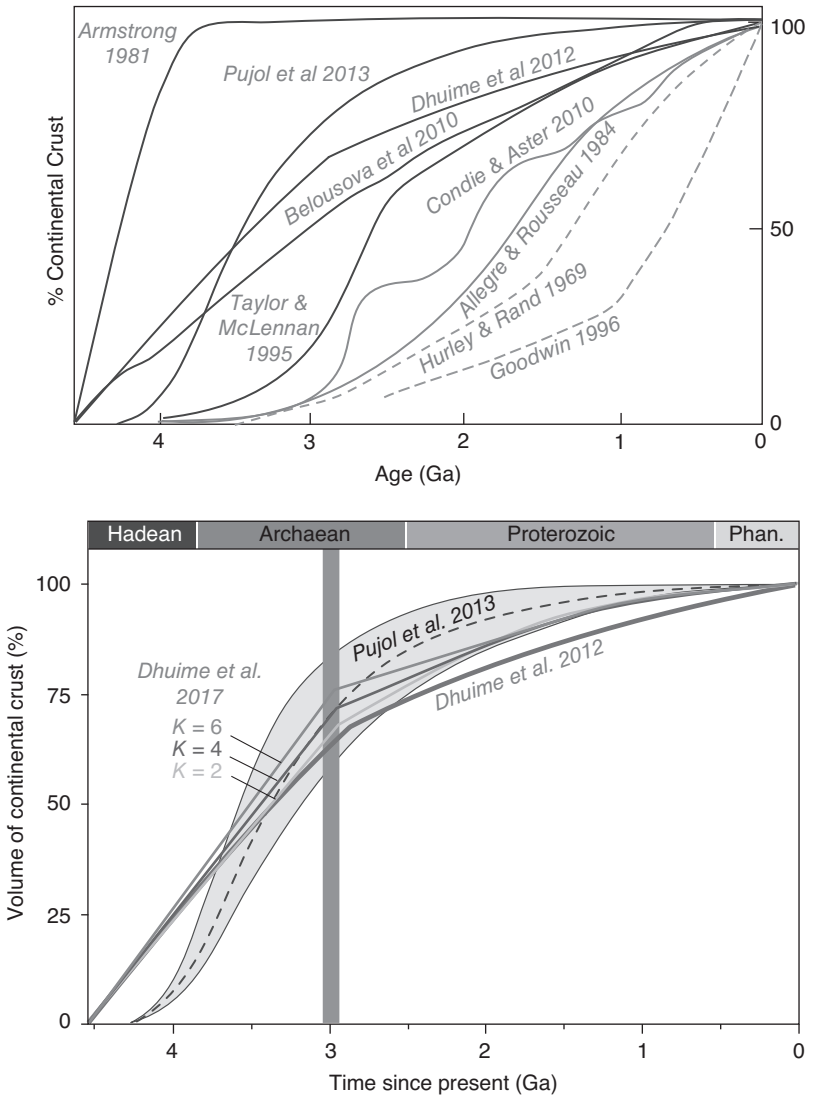


Figure 5.6 *Top:* Models depicting the volume of the continental crust relative to the modern value over time. Dashed curves extrapolate from the age distribution of rocks currently preserved on our planet. Light-shaded solid curves are based on the premise that the present-day age distribution of rocks reflects the volume of continental crust during different epochs. Dark-shaded solid curves rely on geochemical arguments and constraints, mostly by way of isotope ratios of trace elements. *Bottom:* Continental crust growth models as a function of time derived from the analysis of hafnium (Hf), neodymium (Nd) and argon (Ar). The two graphs reveal that ~70 percent of the continental crust could have formed by ~3 Ga (Gyr ago), according to some recent models, although this result varies significantly from study to study. (© China University of Geoscience [Beijing] and Peking University. CC BY-NC-ND 4.0. *Source:* Chris Hawkesworth, Peter A. Cawood, and Bruno Dhuime [2019], Rates of generation and growth of the continental crust, *Geoscience Frontiers* 10[1]: 165–173, figs. 5–6.)

respects. Needless to say, in moving beyond this idealized treatment, our results may undergo major changes when certain planetary characteristics are altered. For example, replacing oxygenated deep oceans analogous to Earth with their anoxic counterparts could result in ocean planets (i.e., sans continents) becoming habitable (Syverson et al. 2020), contra upcoming results in Sections 5.5.1 and 5.5.2. To continue in this vein, worlds with different photosynthetic photon fluxes, surface pressures, rotation rates, and ambient ocean temperatures are expected to have lower or higher rates of biomass production (Olson et al. 2020; Salazar et al. 2020; Lingam & Loeb 2020a). Lastly, we will operate under the assumption that most of the planetary surface is habitable in principle. This simplification is conceivably not valid for tidally locked planets with synchronous rotation—consequently leading to one side of the world permanently facing away from the star—which must confront potential issues such as temperature extremes and minimal photon fluxes on the nightside.

When we invoke the word *water,* it refers to liquid H_2O unless explicitly stated otherwise (permanent icy regions are not included); as per our convention, *land* encompasses all landmasses. We will denote the surface land and water fractions by f_ℓ and f_w, respectively. The ratio (δ_w) of these two fractions is defined as

$$\delta_w = \frac{f_\ell}{f_w} = \frac{1 - f_w}{f_w}. \tag{5.48}$$

Note that $f_\ell = 1 - f_w$ is only valid when ice coverage is minimal and has thus been employed to simplify (5.48). The water fraction for the Earth is $f_\oplus \approx 0.7$, while the land-water ratio is $\delta_\oplus \approx 0.4$. Although the range of δ_w spans 0 and ∞, we see that it is of order unity for the Earth. Our notation and methodology mirror those specified in Lingam and Loeb (2019b). A parallel exploration of kindred topics was conducted by Glaser et al. (2020), who arrived at qualitatively similar results.

5.5.1 How productive are extraterrestrial biospheres?

Before embarking on a more quantitative analysis, it is valuable to consider the two extreme limits to gain a qualitative picture. For starters, let us suppose that $f_w \to 0$, implying that the world is dominated by continents. Due to the minimal availability of water, the majority of landmasses will ostensibly resemble deserts on Earth that are characterized by low biomass

densities. Such worlds have earned the moniker *desert planets,* although the terminology Dune-*like planets* is preferred by some authors; the latter pays homage to the classic *Dune* series by Frank Herbert. The habitability of these desert planets has attracted much attention (Abe et al. 2011; Leconte et al. 2013). In view of the extensive deserts anticipated to exist in the limit $\delta_w \gg 1$, it seems reasonable to hypothesize that water availability constitutes the limiting factor for putative biospheres. In consequence, we would expect the net primary productivity (NPP)—the net rate of organic compounds synthesized by organisms (mostly via photosynthesis)—to be regulated by the access to water (Churkina & Running 1998).

Next, let us consider the diametrically opposite scenario, in which $\delta_w \ll 1$, wherein the surface is almost entirely covered by oceans. These planets are conventionally known as *ocean planets* or *water worlds* in the literature, and their habitability has been subjected to numerous analyses (Léger et al. 2004; Kite & Ford 2018; Ramirez & Levi 2018). In Section 2.7, we pointed out the fact that life-as-we-know-it requires bioessential elements such as phosphorus. Phosphorus (P), in the form of phosphates, is vital for metabolism, replication, and other biological functions; see Section 2.2.5 for additional details. Because of the low solubility of phosphates in sea water, several lines of evidence imply that phosphorus serves as the limiting nutrient for oceanic productivity (Tyrrell 1999; Filippelli 2008). Therefore, during the Precambrian period, it seems plausible that the NPP and ecosystem structure in the oceans was modulated by the availability of dissolved P (Reinhard, Planavsky, et al. 2017; Laakso & Schrag 2018). We will thus operate under the premise that the concentration of dissolved P controls NPP in the limit $\delta_w \ll 1$. It is, however, essential to recognize that other constraints—such as the fluxes of electron donors (Ward et al. 2019)—could restrict NPP further, owing to which our analysis is by no means definitive.

Yet, there is an important distinction between the aforementioned limits that deserves to be highlighted. In the absence of *any* surface water, one would expect the NPP of the world to approach zero since water is seen as a prerequisite for life-as-we-know-it. In contrast, when we consider the case $f_w \to 1$, the NPP ought to be nonzero. This expectation stems from the datum that the weathering of the ocean floor (submarine weathering) provides a nonzero supply of nutrients, including phosphorus. To sum up, one limit is regulated by the abundance of water, whereas the other is governed by the access to dissolved P. Hence, for biospheres that fall between these two

extremes, we will model their associated NPP through the superposition of these two factors.

5.5.1.1 Land fraction and NPP

In accordance with our preceding discussion, the NPP is presumed to be dictated by the access to water in the limit $f_w \ll 1$. Although we do not know the specifics of the hydrological cycle on other worlds, it is nonetheless reasonable to assume that the evaporation of oceans constitutes the major source for precipitation on land. The total evaporation rate $\dot{\mathcal{R}}_E$ from the oceans (in kg / yr) can be expressed as

$$\dot{\mathcal{R}}_E \sim f_w \Phi_E \left(4\pi R^2\right), \tag{5.49}$$

where Φ_E represents the evaporation flux (in kg m^{-2} yr^{-1}). It is not easy to specify Φ_E since it depends on various environmental parameters like the wind speed and ambient temperature. There are various empirical and theoretical prescriptions, such as the Penman equation and its extensions, that could be used for this purpose. We opt to hold the external parameters constant, thus effectively ensuring that Φ_E is equal to that of the Earth.

The evaporated water is, to leading order, precipitated over land and oceans in proportion to their respective areas. Hence, the total amount of H_2O precipitated over land is $\dot{\mathcal{R}}_L = f_l \cdot \dot{\mathcal{R}}_E$, where $f_l = 1 - f_w$ is the land fraction. Needless to say, the precipitation will be distributed in a nonuniform fashion. To account for this feature, consider a toy model wherein a fraction f_h of the land receives precipitation with the remaining fraction ($f_d \equiv 1 - f_h$) being completely arid; in other words, the latter refers to the fraction of land covered by deserts. Next, the precipitation flux in the non-arid regions is given by $\Phi_{\mathcal{P}}$ (in kg m^{-2} yr^{-1}). Using this information, the global precipitation rate $\dot{\mathcal{R}}_{\mathcal{P}}$ (in kg / yr) over the land is estimated to be

$$\dot{\mathcal{R}}_{\mathcal{P}} \sim f_h f_l \left(\Phi_{\mathcal{P}}\right) \left(4\pi R^2\right) + \left(1 - f_h\right) f_l \cdot (0) \cdot \left(4\pi R^2\right). \tag{5.50}$$

It is worth reiterating that f_h signifies the fraction of landmasses that receive precipitation and are therefore habitable insofar as water availability is concerned.

Let us further assume that $\Phi_{\mathcal{P}}$ and Φ_E are comparable in magnitude ($\Phi_{\mathcal{P}} \sim \Phi_E$). By definition, the total precipitation on land per unit time ($\dot{\mathcal{R}}_{\mathcal{P}}$)

should equal the fraction of total water evaporated from the oceans per unit time that is deposited on land ($\dot{\mathcal{R}}_L$). Upon making use of the preceding relations, we end up with

$$f_h \approx f_w. \tag{5.51}$$

From (5.51), it is apparent that $f_d \approx 1 - f_w$ or, equivalently, $f_d \approx f_\ell$. For our planet, it is well-known that $f_d \approx 0.33$ and $1 - f_w \approx 0.3$. Hence, in spite of the many idealizations that were invoked, this model is surprisingly accurate for f_h on Earth. One can easily generalize (5.51) by modeling f_h as a power-law function of f_w with a positive exponent. We will not pursue this path here but will instead work with (5.51) henceforth.

During the course of this derivation, we modeled deserts as being entirely devoid of precipitation, therefore implying that their NPP ought to approach zero. Although this represents a clear idealization, it is nonetheless fairly accurate since the NPP of deserts is much smaller compared to most other ecosystems (Hadley & Szarek 1981). The NPP per unit area—which is measured in units of kg m^{-2} yr^{-1} of carbon (C) fixed—for deserts is approximately an order of magnitude smaller than savannas on average; the NPP for the latter is ~ 0.7 kg m^{-2} yr^{-1} (Field et al. 1998). Given the heterogeneity of deserts on Earth, it is not surprising that the NPP spans two orders of magnitude: the most extreme deserts have NPP of $\sim 3 \times 10^{-3}$ kg m^{-2} yr^{-1}, whereas the NPP of semideserts is ~ 0.2 kg m^{-2} yr^{-1} (Dokulil 2019).

If we wish to calculate the total NPP or biomass of other worlds, it requires not only an in-depth knowledge of the energy and reactant fluxes but also detailed understanding of metabolic networks and pathways associated with the putative extraterrestrial organisms. It goes without saying that we are hampered by the absence of information concerning these aspects. Hence, we shall assume that the average NPP per unit area on land and oceans is commensurate with the corresponding values for the Earth. On the basis of this simplification, the NPP on land, denoted by $\mathcal{B}_\ell$ (in kg / yr), is estimated to be

$$\mathcal{B}_\ell \sim 5.6 \times 10^{13}\,\mathrm{kg\,/\,yr} \left(\frac{f_w}{f_\oplus}\right) \left(\frac{1-f_w}{1-f_\oplus}\right) \left(\frac{R}{R_\oplus}\right)^2, \tag{5.52}$$

with the normalization factor on the right-hand side adopted from Field et al. (1998). Along the same lines, provided that the mean biological

turnover time of extraterrestrial primary producers (autotrophs) is comparable to the Earth's, their total biomass on land (M_ℓ) equals

$$M_\ell \sim 4.5 \times 10^{14}\,\mathrm{kg} \left(\frac{f_w}{f_\oplus}\right)\left(\frac{1-f_w}{1-f_\oplus}\right)\left(\frac{R}{R_\oplus}\right)^2, \tag{5.53}$$

where the normalization factor is the biomass of land-based primary producers on Earth (Bar-On et al. 2018). By inspecting (5.52) and (5.53), we see that they are vanishingly small in the limits $f_w \to 0$ and $f_w \to 1$. In the first regime, the absence of liquid water effectively rules out the prospects for surface-based life-as-we-know-it. On the other hand, only oceans would exist in the latter case, therefore automatically ruling out land-based ecosystems.

5.5.1.2 Water fraction and NPP

Recall that the biological productivity of the oceans has been posited to depend on the abundance of dissolved phosphorus (viz., the limiting nutrient). As per this hypothesis, the NPP of the oceans ($\mathcal{B}_w$), can be expressed as

$$\mathcal{B}_w \propto \phi_P f_w R^2, \tag{5.54}$$

where ϕ_P denotes the average steady-state concentration of dissolved P and the factor $f_w R^2$ is proportional to the area spanned by oceans. In reality, the above equation has a subtle caveat: it becomes valid only when the concentration of P is sufficiently below its saturation limit. In other words, ϕ_P must be replaced by $\phi_P/\left(\mathcal{K}_P + \phi_P\right)$, where $\mathcal{K}_P$ represents the Monod constant and depends on the biological organisms under question. The concentration ϕ_P is determined from

$$\frac{d\mathcal{C}_P}{dt} = \sum \mathcal{S}_P - \sum \mathcal{L}_P \mathcal{C}_P, \tag{5.55}$$

where $\mathcal{C}_P$ (units of mol) is the total amount of dissolved P in the oceans. Note that $\mathcal{S}_P$ signifies a given source of P (in mol / yr) and $\mathcal{L}_P$ denotes a particular sink of P (in yr^{-1}). Thus, the steady-state value of $\mathcal{C}_P$ is obtained by setting the left-hand side to zero. The corresponding estimate for ϕ_P is determined via $\phi_P \approx \mathcal{C}_P / M_{\mathrm{oc}}$, where M_{oc} is the mass of the ocean. Thus,

we end up with

$$\phi_P = \frac{\sum \mathcal{S}_P}{M_{\rm oc} \sum \mathcal{L}_P}. \tag{5.56}$$

Solving for ϕ_P is not an easy task since all the sources and sinks evolve with time and are contingent on the abundances of other chemical elements such as iron. In addition, there exist multiple sources and sinks of P, of which many are poorly constrained. Hence, we shall focus only on the primary mechanisms herein.

One of the most prominent inputs of soluble P into the oceans is driven by continental weathering, with the transport facilitated via rivers. We argued in Section 5.5.1.1 that only a fraction f_h of the land actually receives precipitation. Therefore, it seems reasonable to assume that the rivers originate only in such regions. Hence, the riverine source of P is given by

$$\mathcal{S}_{\rm P} \sim 3 \times 10^{10}\,{\rm mol\,/\,yr} \left(\frac{f_w}{f_\oplus}\right)\left(\frac{1-f_w}{1-f_\oplus}\right)\left(\frac{R}{R_\oplus}\right)^2, \tag{5.57}$$

with the normalization factor for Earth adopted from Wallmann (2010). Clearly, this equation vanishes in the limiting cases when (1) oceans do not exist (zero precipitation and river runoff) and (2) land is absent (zero continental weathering). The factor $f_w\left(1-f_w\right)$ enters (5.57) because it encapsulates the fraction of non-arid land—namely, $f_h \cdot f_l$. The latter result yields the former after making use of (5.51) as well as the definition of f_l. Before moving ahead, we note that the above equation is truly applicable only if the riverine flux of P is comparable to modern Earth. In actuality, this flux is not guaranteed to be roughly constant across space and time—even on Earth, it may have been higher by a factor of a few during the late Archean era with respect to its current value (Hao et al. 2020a).

Another primary source of phosphorus to the oceans entails the deposition of atmospheric dust, aerosols, and volcanic ash. It goes without saying that this source depends on the fraction of available land (f_ℓ). We shall posit that the particulate matter liberated from erosion is dispersed across the oceans in proportion to their area. As per this line of reasoning, the magnitude of the source should be proportional to the factor $f_w \cdot f_l = f_w\left(1-f_w\right)$. This ansatz vanishes when $f_w \to 0$ and $f_w \to 1$, which is to be expected, because the existence of only oceans or landmasses would negate this source. A more accurate treatment can be undertaken by replacing f_ℓ with $f_d \cdot f_\ell$

because most of the aeolian erosion takes place in the deserts (Shao et al. 2011), but the final results are mostly unchanged. The magnitude of the influx from this source is

$$\mathcal{S}_{\rm P} \sim 1 \times 10^{10}\,{\rm mol\,/\,yr} \left(\frac{f_w}{f_\oplus}\right)\left(\frac{1-f_w}{1-f_\oplus}\right)\left(\frac{R}{R_\oplus}\right)^2, \tag{5.58}$$

where the prefactor on the right-hand side quantifies the amount of soluble reactive P supplied by atmospheric sources on Earth (Benitez-Nelson 2000).

The third source that we shall consider is the submarine weathering of the ocean floor by seawater to generate soluble P. In actuality, this mechanism serves as a minor source on our planet but is very important nevertheless for reasons that shall soon become apparent. The reason behind the comparatively small magnitude of this source is that the flux of dissolved P released exhibits an exponential dependence on the pH. Therefore, it is crucial to recognize that the pH of rainwater is around 5.6, whereas seawater in the current epoch has a pH of approximately 8.0; in contrast, the pH of the oceans at ~ 4 Ga might have been ~ 6.6 (see Section 2.2.5). The details underlying the derivation of $\mathcal{S}_{\rm P}$ for submarine weathering are presented in Section 7.6.2.1, owing to which we shall not repeat them here. It suffices to say that the present-day magnitude of this source is about two orders of magnitude smaller than the riverine influx, thus yielding

$$\mathcal{S}_{\rm P} \sim 1.3 \times 10^{8}\,{\rm mol\,/\,yr} \left(\frac{f_w}{f_\oplus}\right)\left(\frac{R}{R_\oplus}\right)^2. \tag{5.59}$$

In contrast, if we seek to estimate $\mathcal{S}_{\rm P}$ due to submarine weathering at ~ 4 Ga, we may have to multiply the above equation by a factor of ~ 20 because of the lower pH in this period. The vital point to appreciate here is that (5.59) vanishes when $f_w \to 0$, but it does *not* vanish when in the limit $f_w \to 1$ (i.e., when land is absent) since it does not depend on the presence of continental landmasses.

Extensive submarine weathering is expected to be functional only when seawater is in contact with the oceanic crust. On worlds with deep oceans, the pressure at the seafloor will become so high as to result in the formation of high-pressure ices that suppress direct rock-water interactions. By inspecting the phase diagram of H_2O, one can verify that the pressure at which ice formation occurs is ~ 2 GPa, provided that the upper limit for

the temperature is 373 K. Hence, if the average ocean depth is denoted by $\mathcal{H}$, we require $\rho_w g\mathcal{H} \lesssim 2$ GPa for the existence of liquid H_2O; note that ρ_w and g represent the density of water and the surface gravity of the planet, respectively. By solving for $\mathcal{H}$ and using $M_{\rm oc} \approx 4\pi f_w \rho_w R^2 \mathcal{H}$, we obtain

$$M_{\rm oc} \lesssim 52\, M_{\rm oc,\oplus} \left(\frac{f_w}{f_\oplus}\right) \left(\frac{R}{R_\oplus}\right)^{0.3}, \tag{5.60}$$

where $M_{\rm oc,\oplus} \approx 1.4 \times 10^{21}$ kg is the mass of Earth's oceans and we have made use of $g \propto R^{1.7}$.

Three additional sources of P are not incorporated into our analysis. The first involves volcanic activity, which is not considered because it is highly localized and unlikely to play a major role on global scales (Benitez-Nelson 2000). The second entails the exogenous delivery of soluble reactive P in the form of the mineral schreibersite via meteorites. This source is proportional to the flux of impactors and therefore drops sharply once the frequency of impacts declines. Even if we consider the young Earth, which was subject to recurring impacts, the upper bound on soluble phosphorus delivered via this channel has been estimated to be $\sim 10^8$ kg / yr (Pasek et al. 2017), which translates to $\sim 3.3 \times 10^9$ mol / yr, thereby making it smaller than the riverine and atmospheric sources. The last source of P is based on weathering by glaciers, but the magnitude of this source is expected to fluctuate greatly over time and thus remains poorly constrained. In the current epoch, $\mathcal{S}_{\rm P}$ for glacial weathering is, at most, comparable to (5.57) and (5.58) on our planet. Hence, at least insofar as Earth is concerned, our results are expected to vary only by a factor of order unity even if these mechanisms are taken into account.

Next, we turn our attention to the sinks of P. They can be divided into two broad categories: (1) burial of marine sediments and (2) precipitation in the vicinity of hydrothermal vents (Paytan & McLaughlin 2007). The latter must be properly accounted for since its magnitude, in the case of present-day Earth, is potentially only ~ 3 times smaller than the former, based on chemical analyses (Wheat et al. 2003). For reasons that are explicated in Section 7.6.2.1, the sinks are characterized by the common scaling:

$$\sum \mathcal{L}_P \propto \left(\frac{M}{M_{\rm oc}}\right). \tag{5.61}$$

The characteristic depletion timescale for phosphorus ($\tau_P \equiv 1/\sum \mathcal{L}_P$) would vanish in the limit $M_{\rm oc} \to 0$ (i.e., when oceans and their phosphorus inventories are absent). The constant of proportionality in (5.61) must be chosen to maintain consistency with that of the Earth, but its exact value is subject to some variability.

By making use of the preceding results concerning the sinks and sources along with (5.56), we end up with

$$\frac{\phi_P}{\phi_\oplus} \sim \left(\frac{f_w}{f_\oplus}\right)\left[\left(\frac{1-f_w}{1-f_\oplus}\right) + 3.3 \times 10^{-3}\right]\left(\frac{R}{R_\oplus}\right)^{-1.7}, \tag{5.62}$$

where we have made use of the mass-radius relationship $M \propto R^{3.7}$ applicable to rocky worlds with $R \gtrsim R_\oplus$. Note that $\phi_\oplus \approx 2.1\ \mu$M encapsulates the average concentration of dissolved phosphorus in our oceans. This result follows from the fact that the total inventory of dissolved P is approximately 3×10^{15} mol (Paytan & McLaughlin 2007), while the mass of Earth's oceans is 1.4×10^{21} kg. As a consistency check, suppose that the fraction of continental crust in the Archean around 3 Ga was ~ 5 percent of the coverage today, based on B. W. Johnson and Wing (2020) and a few of the continental growth models depicted in Figure 5.6. After using (5.62), we obtain $\phi_P \sim 0.16\ \mu$M, which is not far removed from the estimate $\phi_P \approx 0.04$–$0.13\ \mu$M, calculated using a semiempirical ocean chemistry model in C. Jones et al. (2015).

Upon substituting (5.62) in (5.54), the NPP of the oceans is given by

$$\mathcal{B}_w \sim 4.9 \times 10^{13}\,\mathrm{kg/yr}\left[\left(\frac{1-f_w}{1-f_\oplus}\right) + 3.3 \times 10^{-3}\right]\left(\frac{f_w}{f_\oplus}\right)^2\left(\frac{R}{R_\oplus}\right)^{0.3}, \tag{5.63}$$

where the normalization for the Earth has been adopted from Field et al. (1998). Inspecting this equation, we see that $\mathcal{B}_w \to 0$ when $f_w \to 0$ along expected lines. On the other hand, in the limit $f_w \to 1$, we observe $\mathcal{B}_w$ remains finite. This trend is attributable to the datum that the inventory of dissolved P is nonzero, albeit very small compared to present-day Earth, even in the total absence of landmasses.

Let us consider Archean Earth to evaluate the accuracy of this model, whose land fraction is modeled as ~ 5 percent of the modern value in accordance with a recent study (B. W. Johnson & Wing 2020). Therefore,

upon substituting $f_\ell \approx 1.5 \times 10^{-2}$ in (5.63), we end up with $\mathcal{B}_w \sim 5.2 \times 10^{12}\,\mathrm{kg\,/\,yr}$. This value falls within the oceanic NPP range of $\mathcal{B}_w \approx 2.4$–$6.4 \times 10^{12}\,\mathrm{kg\,/\,yr}$ predicted by Canfield et al. (2006), albeit derived from electron donor constraints not included herein. It also exhibits good agreement with the upper bound of $\mathcal{B}_w \sim 3.6 \times 10^{12}\,\mathrm{kg\,/\,yr}$ obtained by Hao et al. (2020b) via modeling the biogeochemical cycling of P on Archean Earth. Finally, our result is compatible with the upper limits derived by Laakso & Schrag (2019) and Hao et al. (2020a) for Earth's oceans in the Proterozoic and Archean eons.

Next, by supposing that the typical biological turnover time in extraterrestrial oceans is comparable to the Earth, the total biomass of primary producers in the oceans (M_w) can be estimated accordingly, thereby yielding

$$M_w \sim 10^{12}\,\mathrm{kg}\left(\frac{f_w}{f_\oplus}\right)^2\left[\left(\frac{1-f_w}{1-f_\oplus}\right)+3.3\times10^{-3}\right]\left(\frac{R}{R_\oplus}\right)^{0.3}. \tag{5.64}$$

The biomass of producers in Earth's oceans has been adopted from Bar-On et al. (2018). Upon comparing (5.64) with (5.53), it is evident that the latter is around two orders of magnitude higher with respect to the former for worlds with Earthlike water fractions. The underlying reason for the much higher biomass of primary producers on Earth's landmasses relative to the oceans is inextricably linked with the evolution of land plants. The latter, as pointed out in Section 3.9.2, arguably represents one of the most significant developments in our planet's evolutionary history.

5.5.1.3 Total NPP and biomass

We are now in a position to calculate the total net primary productivity $\mathcal{B}_t = \mathcal{B}_\ell + \mathcal{B}_w$ and the total biomass of primary producers $M_t = M_\ell + M_w$ as a function of the water fraction and the radius of a given world. It is more instructive, instead, to calculate their values relative to the Earth. Therefore, we shall introduce the following ratios:

$$\Delta_{\mathcal{B}} = \frac{\mathcal{B}_t}{\mathcal{B}_{t,\oplus}} \sim f_w\left(\frac{R}{R_\oplus}\right)^2\left[2.52\left(1-f_w\right)+3.2f_w\left[\left(1-f_w\right)+10^{-3}\right]\left(\frac{R}{R_\oplus}\right)^{-1.7}\right] \tag{5.65}$$

$$\Delta_M = \frac{M_t}{M_{t,\oplus}} \sim f_w \left(\frac{R}{R_\oplus}\right)^2 \Bigg[4.76 \left(1 - f_w\right) + 1.51 \times 10^{-2} f_w \left[\left(1 - f_w\right) + 10^{-3}\right] \left(\frac{R}{R_\oplus}\right)^{-1.7} \Bigg] \tag{5.66}$$

$\mathcal{B}_{t,\oplus}$ and $M_{t,\oplus}$ are the total NPP and biomass of primary producers for our planet. It is straightforward to compute the f_w's at which the maximum values of the Δ's occur. By setting $d\Delta_{\mathcal{B}}/df_w = 0$, we find

$$f_{\mathcal{B}} \approx \frac{1}{3}\Bigg[1 - 0.79 \left(\frac{R}{R_\oplus}\right)^{1.7} + \sqrt{1 + 0.79 \left(\frac{R}{R_\oplus}\right)^{1.7} + 0.62 \left(\frac{R}{R_\oplus}\right)^{3.4}} \Bigg], \tag{5.67}$$

where $f_{\mathcal{B}}$ signifies the water fraction at which $\Delta_{\mathcal{B}}$ reaches a maximum. Likewise, we can compute f_M by means of $d\Delta_M/df_w = 0$, thus ending up with

$$f_M \approx \frac{1}{3}\Bigg[1 - 315 \left(\frac{R}{R_\oplus}\right)^{1.7} + \sqrt{1 + 315 \left(\frac{R}{R_\oplus}\right)^{1.7} + 99173 \left(\frac{R}{R_\oplus}\right)^{3.4}} \Bigg]. \tag{5.68}$$

If we specialize to $R = R_\oplus$, we obtain $f_{\mathcal{B}} \approx 0.59$ and $f_M \approx 0.5$. The corresponding Δ's are given by $\Delta_{\mathcal{B}} \approx 1.07$ and $\Delta_M \approx 1.19$. Thus, according to this toy model, the net primary productivity and biomass of Earth's biosphere are close to being optimal. By using (5.48), we see that $\delta_w\left(f_{\mathcal{B}}\right) \approx 0.7$ and $\delta_w\left(f_M\right) \approx 1$, implying that the ratio of land and water fractions is close to unity in both instances.

Recall that the variable δ_w was introduced because of its ability to span many orders of magnitude in principle. Hence, it is more instructive to plot (5.65) and (5.66) as a function of δ_w. The results are depicted in Figure 5.7 for three different values of $R/R_\oplus$. If we consider $R = R_\oplus$, we find that $\Delta_{\mathcal{B}} > 0.1$ is satisfied when $\delta_w \in (0.02, 20)$, implying that δ_w spans three orders of magnitude. Repeating the same analysis for $\Delta_M > 0.1$,

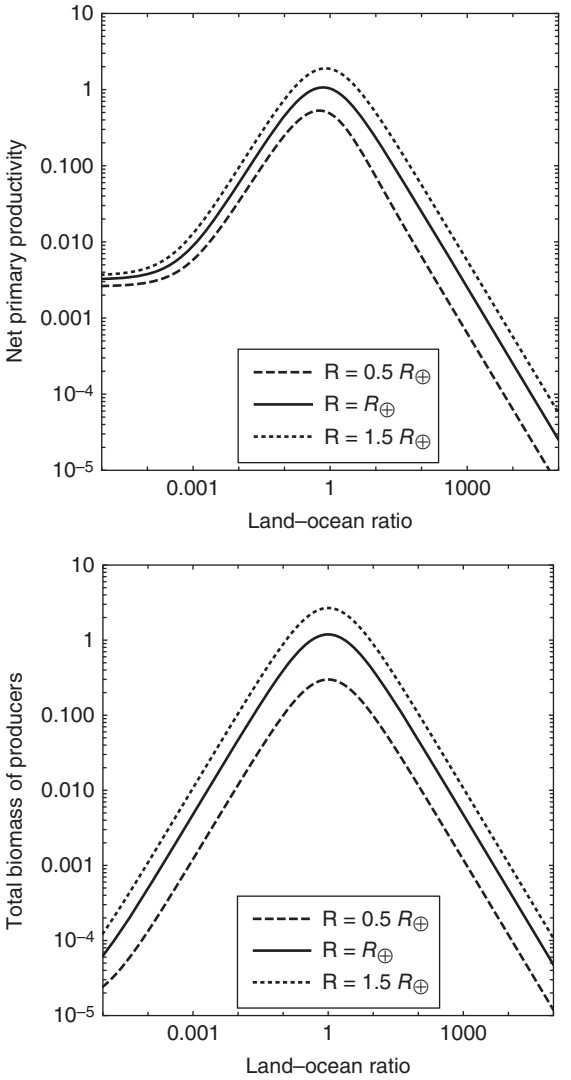


Figure 5.7 *Top:* Net primary productivity (NPP) normalized by its value for the Earth. *Bottom:* Total biomass of primary producers (mostly photosynthetic organisms) normalized by that of the Earth's biomass. In both graphs, the horizontal axes signify the ratio of the world's surface covered by landmasses to that spanned by oceans. The dashed, solid, and dotted curves correspond to worlds with radii of $0.5\,R_\oplus$, $R_\oplus$, and $1.5\,R_\oplus$, respectively. (© Manasvi Lingam and Avi Loeb.)

the corresponding domain turns out to be similar, i.e., we end up with $\delta_w \in (0.02, 50)$. In the top panel of Figure 5.7, the curves are found to flatten as $\delta_w \to 0$. This behavior stems from the fact that (5.65) is nonvanishing in the limit $f_w \to 1$ (equivalent to $\delta_w \to 0$) because of the supply of soluble phosphorus through the submarine weathering of the ocean floor. This flattening is also applicable to (5.66), but this trend is not manifested in the bottom panel of Figure 5.7 since δ_w does not become sufficiently small.

5.5.2 Consequences for the buildup of atmospheric oxygen

The total NPP ($\mathcal{B}_t$), as specified earlier, embodies the rate of organic carbon fixation primarily via photosynthesis. An important point worth bearing in mind is that photosynthesis should not be exclusively conflated with *oxygenic* photosynthesis, given the existence of several anoxygenic photosynthetic pathways based on electron donors like Fe^{2+} and H_2S. Nevertheless, there is no doubt that the evolution of oxygenic photosynthesis was responsible for dramatic changes in the evolutionary landscape, as outlined in Section 3.3. The most significant contributor, by far, to the total NPP on modern Earth is oxygenic photosynthesis. Hence, in dealing with Earthlike worlds, it appears reasonable to suppose that oxygenic photosynthesis constitutes the dominant channel for the biological production of organic compounds.

On Earth, only a small fraction of the total organic carbon synthesized via oxygenic photosynthesis undergoes burial. As this carbon is not otherwise consumed, it effectively serves as an indirect source of atmospheric O_2. Thus, by combining the NPP with the efficiency of carbon burial, we can calculate the net production rate of O_2 (denoted by $\mathcal{S}_{O_2}$). Precise estimates of the average burial efficiency are difficult, even when it comes to our planet. Hence, we shall make the simplifying assumption that the burial efficiency is weakly dependent on the land and water fractions of the planet, i.e., it is held constant and equal to that of present-day Earth. In this event, by mirroring our analysis in Section 4.3.5, we obtain

$$\mathcal{S}_{O_2} \sim 10^{13}\ \text{mol / yr}\ \Delta_{\mathcal{B}}, \tag{5.69}$$

with $\Delta_{\mathcal{B}}$ defined in (5.65). Next, when it comes to oxygen sinks, the chief one involves rapid reactions between O_2 and reducing gases released from volcanism and submarine weathering. Following Section 4.3.5, we treat the average flux of these reducing gases (units of mol m^{-2} yr^{-1}) as being roughly

constant—that is, independent of the surface water fraction. By doing so, the depletion rate of atmospheric O_2 (represented by $\mathcal{L}_{O_2}$) is given by

$$\mathcal{L}_{O_2} \sim 5.7 \times 10^{12}\,\mathrm{mol\,/\,yr} \left(\frac{R}{R_\oplus}\right)^2. \tag{5.70}$$

With regard to our notation ($\mathcal{L}_{O_2}$) in the above equation, it must be recognized that it has different units compared to $\mathcal{L}_P$ in Section 5.5.1.2. In order for atmospheric O_2 to accumulate in the atmosphere, we require $\mathcal{S}_{O_2} > \mathcal{L}_{O_2}$. Hence, it is helpful to reintroduce the oxygen source-to-sink ratio, $\Delta_{O_2} = \mathcal{S}_{O_2}/\mathcal{L}_{O_2}$, which is expressible as

$$\Delta_{O_2} \sim 4.42 f_w \left(1 - f_w\right) + 5.61 f_w^2 \left[\left(1 - f_w\right) + 10^{-3}\right] \left(\frac{R}{R_\oplus}\right)^{-1.7}. \tag{5.71}$$

In the parameter space where $\Delta_{O_2} > 1$ is valid, it could make the buildup of atmospheric O_2 feasible. The regions where this inequality holds true can be determined from an inspection of Figure 5.8. The above criterion is fulfilled when the plots cross upward of the horizontal dash-dotted line corresponding to $\Delta_{O_2} = 1$. When we restrict ourselves to $R = R_\oplus$, we see that $\Delta_{O_2} > 1$ necessitates $f_w \in (0.23, 0.88)$, with the corresponding interval for δ_w becoming $\delta_w \in (0.14, 3.35)$. In other words, it is evident that δ_w ought to fall within a fairly narrow range in order to permit the accumulation of atmospheric O_2. Another result worth highlighting in Figure 5.8 is that the allowed range for δ_w broadens as one considers smaller worlds. To put it differently, it is conceivable that the constraints on surface water fractions are less stringent for smaller worlds when it comes to permitting the buildup of O_2 in the atmosphere. We will elaborate on the consequences for the detection of biosignatures in Section 5.5.5.

5.5.3 Major evolutionary events and the land-water ratio

One of the central themes in Chapter 3 was that the evolution of modern humans (i.e., technological intelligence) appears to have necessitated a small number (< 10) of crucial evolutionary steps. These transitions occurred in sequential order and were broadly characterized by increasing biological complexity. Furthermore, we beheld in Section 3.9 that several theoretical analyses indicate that the total number of major evolutionary events on Earth was perhaps five or six.

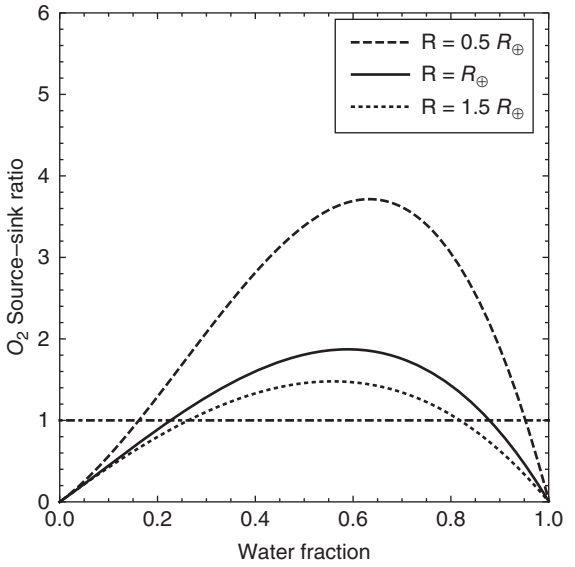


Figure 5.8 The ratio of oxygen sources to sinks is plotted as a function of the water fraction f_w for different values of the planetary radius R. When the curves cross the horizontal dash-dotted line, it may be possible for atmospheric O_2 to build up. The dashed, solid, and dotted curves correspond to worlds with radii of 0.5 $R_\oplus$, $R_\oplus$, and 1.5 $R_\oplus$, respectively. (© Manasvi Lingam and Avi Loeb.)

At this stage, two important caveats are worth highlighting. First, theoretical models that explain the emergence of complex life and technologically sophisticated life on our planet (and elsewhere) through a succession of critical steps do not provide the only viable explanation. Instead, it is conceivable that each major evolutionary event has a relatively high likelihood of occurrence, given the right environmental conditions, because of many paths culminating in the same outcome. Second, even if extraterrestrial technological intelligence did arise on other worlds, the major evolutionary events that led to its emergence may not necessarily parallel those that occurred on Earth. Nevertheless, as argued in Section 3.9.4, there are fairly plausible reasons for contending that at least a subset of such worlds would be distinguished by the actualization of major evolutionary innovations congruent to those identified on Earth.

For the time being, let us focus on a particular evolutionary breakthrough—namely, the origin of life (abiogenesis)—to illustrate our approach. Many theoretical models tend to envision abiogenesis as the successful

outcome of innumerable independent, random trials (de Duve 2005). The total number of such trials is clearly governed by both temporal and spatial aspects, but the majority of studies have concerned themselves solely with the former—to wit, the timescale(s) for abiogenesis to be initiated. However, the spatial component merits equal attention, especially if one presumes that life originated and evolved in localized environments (Deamer 1997). Needless to say, we lack sufficient knowledge of either the characteristic area of this putative microenvironment or its geochemical properties, but we shall presume that the former equals $\mathcal{A}_\ell$ (land) or $\mathcal{A}_w$ (ocean).

In the case of microenvironments situated in the ocean, the number of microenvironments ($\mathcal{N}_m$) is expressible as $\mathcal{N}_m \propto f_w R^2/\mathcal{A}_w$. Along the same lines, for land-based microenvironments, we have $\mathcal{N}_m \propto f_w \left(1 - f_w\right) R^2/\mathcal{A}_\ell$. Note that the extra factor of f_w has been introduced since we chose to focus on the land fraction that is "habitable" (non-arid), and we have utilized (5.51) accordingly. An implicit assumption invoked herein is that the *filling fractions* of these microenvironments, either on landmasses or in oceans, do not vary significantly across worlds. Moreover, if we further conjecture that the areas of the relevant microenvironments are similar across all worlds, we obtain $\mathcal{N}_m \propto f_w R^2$ and $\mathcal{N}_m \propto f_w \left(1 - f_w\right) R^2$ for oceans and land, respectively.

Hitherto, our analysis was framed in the context of abiogenesis, but it can be readily generalized to encompass other major evolutionary events. Next, it seems reasonable to propose that $\mathcal{P}_j \propto \mathcal{N}_j$, where $\mathcal{P}_j$ represents the likelihood of occurrence for the j-th critical step. In essence, this relation implies that the probability of a successful evolutionary breakthrough is directly proportional to the total number of spatial "trials" feasible (de Duve 2005)—namely, the maximum number of accessible microenvironments as per our setup. In Section 3.9, we saw that the number and choice of major evolutionary events differs from framework to framework. Hence, in what follows, we will investigate the ramifications of one such model—the six megatrajectories identified by Knoll and Bambach (2000) and covered in Section 3.9.2. In principle, the same analysis is easily repeatable for other frameworks and is left as an exercise for the reader.

(1) From abiogenesis to the Last Universal Common Ancestor: We do not currently have a clear picture of when and where life originated on Earth, but the various hypotheses concerning the sites of abiogenesis can be classified into ocean- and land-based environments; see Section 2.7 for more details. Another crucial detail that remains poorly understood is the water

(or land) fraction at the time of abiogenesis on our planet. In light of these systemic uncertainties, we will hypothesize that the temporal evolution of water fraction on other worlds is akin to that of the Earth. In this scenario, for ocean-based sites of abiogenesis, we find

$$\mathcal{P}_1 \propto f_w R^2, \tag{5.72}$$

whereas assuming the origin of life on land-based microenvironments leads to

$$\mathcal{P}_1 \propto f_w \left(1 - f_w\right) R^2. \tag{5.73}$$

Due to the scarcity of empirical evidence from fossils and biomarkers, determining whether the next three steps unfolded on land or oceans is difficult. It appears more plausible that the oceans served as the sites for these evolutionary events since they occurred during the Precambrian period (Vermeij 2017). Hence, we shall operate under this premise henceforth.

(2) Metabolic diversification of bacteria and archaea: By 3.4 Ga, most of the metabolic pathways documented in present-day bacteria and archaea had probably arisen. Since we have supposed that this step occurred in the oceans, we have

$$\mathcal{P}_2 \propto f_w R^2. \tag{5.74}$$

(3) Evolution of the eukaryotic cell: The causes and timing of eukaryogenesis have been debated extensively, but fossil evidence seemingly indicates that they had evolved by 1.8 Ga or even earlier. The likelihood for this step is expressible as

$$\mathcal{P}_3 \propto f_w R^2. \tag{5.75}$$

(4) Complex multicellularity: Recall that *complex multicellularity* essentially refers to organisms that are characterized by their sophisticated cell differentiation. Animals, plants, and fungi are the chief examples in this category. Along the lines of the previous two steps, the likelihood is given by

$$\mathcal{P}_4 \propto f_w R^2. \tag{5.76}$$

The final two megatrajectories were contingent on the existence of landmasses on our planet, which necessitates an adjustment of the likelihood functions accordingly. A vital point worth reiterating is that the evolutionary events specified herein are not automatically guaranteed to arise in the

same temporal sequence, or in similar environments, on other habitable worlds.

(5) Invasion of the land: In this context, the so-called invasion alludes to the expeditious expansion of land plants during the Paleozoic era, whose likelihood is

$$\mathcal{P}_5 \propto f_w \left(1 - f_w\right) R^2. \tag{5.77}$$

(6) Intelligence and technology: As we shall explore in Section 7.7, there exist compelling grounds for believing that cetaceans exhibit cultural signatures as well as high intelligence. Yet, at the same time, it appears relatively unlikely for purely aquatic species to develop sophisticated technology capable of reshaping the entire biosphere, as noted in Section 7.7. Hence, under the assumption that this megatrajectory is always land-based, we have

$$\mathcal{P}_6 \propto f_w \left(1 - f_w\right) R^2. \tag{5.78}$$

On the basis of the above steps, we define the function $\mathcal{P}_t = \prod_{i=1}^{6} \mathcal{P}_j$ that encapsulates the cumulative likelihood for the emergence of technological intelligence on a particular world. Since there are several proportionality constants involved, it is more instructive to construct the likelihood relative to the Earth. We will, therefore, introduce

$$\Delta_{\text{Int}} = \frac{\mathcal{P}_t}{\mathcal{P}_{t,\oplus}}, \tag{5.79}$$

where it should be recognized that two different functions arise because the microenvironments associated with abiogenesis are unknown. The superscripts ℓ and w are used to differentiate between the cases where life originated on land and in the ocean, respectively. Given the above considerations and notation, we obtain

$$\Delta_{\text{Int}}^{(\ell)} = \left(\frac{f_w}{f_\oplus}\right)^6 \left(\frac{1 - f_w}{1 - f_\oplus}\right)^3 \left(\frac{R}{R_\oplus}\right)^{12} \tag{5.80}$$

$$\Delta_{\text{Int}}^{(w)} = \left(\frac{f_w}{f_\oplus}\right)^6 \left(\frac{1 - f_w}{1 - f_\oplus}\right)^2 \left(\frac{R}{R_\oplus}\right)^{12}. \tag{5.81}$$

By inspecting (5.80) and (5.81), several points stand out immediately. First, the likelihood functions are highly sensitive to the size of the world because of the very strong dependence on R (namely $\propto R^{12}$). Using (5.80), it is found that $\Delta_{\text{Int}}^{(\ell)}$ attains a maximum when $f_w = 2/3$ ($\delta_w = 1/2$), with the value of this quantity becoming $\Delta_{\text{Int}}^{(\ell)} \approx 1.02$ at $R = R_\oplus$. In the same vein, using (5.81), we see that the maximum occurs at $f_w = 3/4$ ($\delta_w = 1/3$) corresponding to $\Delta_{\text{Int}}^{(w)} \approx 1.05$ at $R = R_\oplus$. Therefore, as per our toy model, the Earth's current land-water ratio might be nearly optimal insofar as the evolution of technological intelligence is concerned. However, due to the R^{12} dependence, worlds larger than the Earth with the same land-water ratio are potentially more habitable. An example of a rocky planet bigger than the Earth is Kepler-20b, at a distance of 285 pc from our planet, whose radius and mass are approximately $1.87\,R_\oplus$ and $9.70\,M_\oplus$, respectively.

As before, taking advantage of the greater range spanned by δ_w, (5.80) and (5.81) have been plotted as a function of this variable in Figure 5.9. Let us consider $R = R_\oplus$ for the sake of simplicity. In the top panel, we observe that $\Delta_{\text{Int}}^{(\ell)} > 0.1$ is fulfilled when $\delta_w \in (0.09, 2.17)$, or equivalently $f_w \in (0.32, 0.92)$. Similarly, in the bottom panel, it can be verified that $\Delta_{\text{Int}}^{(w)} > 0.1$ is attainable provided that $\delta_w \in (0.04, 1.68)$—that is, when $f_w \in (0.37, 0.96)$. Thus, in both instances, we find that δ_w spans only about an order of magnitude before the overall relative likelihood drops below 10 percent. In other words, the prospects for the emergence of technological intelligence may become < 10 percent with respect to our planet when $f_w < 0.3$.

In view of the above items, our model implies that extensive water bodies are required to boost the chances for the evolution of technological intelligence. This result stems from the fact that all of the critical steps, regardless of whether they are land- or ocean-based, depend on the availability of surface liquid water. Thus, insofar as the land and water fractions are concerned, the latter is more pertinent than the former. With that being said, the presence of land was probably crucial for subsequent high-performance evolutionary innovations on Earth (Vermeij 2017), and conceivably would be on other worlds as well. Consequently, upon taking the limit $f_w \to 1$, one can easily verify that the functions (5.80) and (5.81) approach zero. Our result differs from Section 5.5.1.3 in this regard, because the latter dealt with nontechnological life.

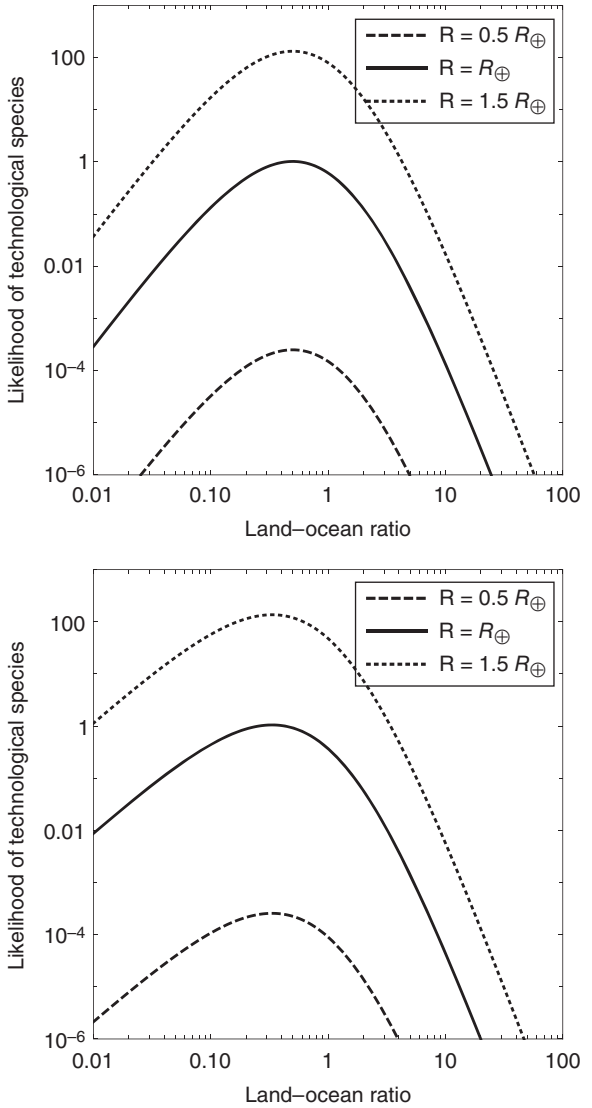


Figure 5.9 *Top:* Cumulative likelihood of technological intelligence relative to Earth, under the assumption that the origin of life necessitated a land-based environment. *Bottom:* The same function, assuming that life originated in an oceanic environment. In both graphs, the *x*-axis signifies the ratio of the planetary surface covered by landmasses to that spanned by oceans. The dashed, solid, and dotted curves correspond to worlds with radii of 0.5 $R_\oplus$, $R_\oplus$, and 1.5 $R_\oplus$, respectively. (© Manasvi Lingam and Avi Loeb.)

5.5.4 How common are desert worlds and water worlds?

The first, and possibly the most important, point to bear in mind is that the surface water fraction of the planet evolves over time. The initial water inventory of the planet is dictated by a number of dynamical considerations, including the location at which the planet was formed as well as its final location (if the two differ) and the delivery of volatiles via asteroids and comets (O'Brien et al. 2018). Theoretical models indicate that water-rich worlds are quite common and that H_2O inventories of planets vary widely. Over time, surface water is either depleted due to photolysis involving X-rays and ultraviolet radiation (XUV) during the pre-main-sequence phase (Section 4.3.1) or transferred to the mantle via the deep-water cycle (Monteux et al. 2018). The former depends on stellar properties (e.g., mass and age) as well as the size of the planet, whereas the latter is governed by geological properties such as temperature, pressure, interior composition, and plate tectonics. Hence, when viewed in toto, determining the probability distribution of water fraction remains a challenging endeavor.

Instead, we will consider two different candidates for the distribution function and explore how the results vary accordingly. We define the water fraction as $\mathcal{F}_w = M_W/M$, where M_W is the total mass of water present, not only in the oceans but also in the mantle.[4] For the sake of convenience, we opt to focus on worlds with $R \approx R_\oplus$ endowed with an interior structure resembling that of Earth (see Doyle et al. 2019). In both models, the upper bound on $\mathcal{F}_w$ is fixed at 0.5, motivated by the premise that some of the TRAPPIST-1 planets might be characterized by similar values (Unterborn et al. 2018), although other studies indicate $\mathcal{F}_w \lesssim 0.05$ for the TRAPPIST-1 system (Grimm et al. 2018; Turbet et al. 2020). In principle, the lower bound on $\mathcal{F}_w$ can become arbitrarily small, but we shall truncate it at 10^{-12} on the basis of the numerical simulations by Tian and Ida (2015).

Since the water is partitioned between the hydrosphere and mantle, we introduce $M_W = M_{\rm oc} + M_u$, where M_u signifies the mass of water located within the mantle. Even if we restrict ourselves to Earth, calculating M_u has proven to be challenging, owing to which current estimates vary widely. Using the available evidence, we choose the fiducial value of $M_u \approx 4M_{\rm oc,\oplus}$

4. In addition, "water" is present in the form of hydrogen in the core, but the magnitude of this repository is poorly constrained (Peslier et al. 2017), and therefore we choose to exclude it from our analysis.

(Peslier et al. 2017). Of the water present in the mantle, a sizable amount is believed to exist in the mantle transition zone (MTZ), which separates the upper and lower mantle. Interestingly, recent experiments and modeling suggest that the Earth's MTZ is close to water saturation (Fei et al. 2017). Next, we note that the tallest landmass above sea level on our planet is Mount Everest at a height of nearly 9 km. Hence, if we seek to submerge all topographical features above sea level, the mass of water required is $\sim \rho_w \left(4\pi R_\oplus^2\right)(9\,\mathrm{km})$. Adding this extra mass to the water already present in Earth's hydrosphere yields $\sim 4.3\, M_{\mathrm{oc},\oplus}$.

As per our preceding analysis, the Earth might be transformed into a water world when the inequality $M_W > M_u + 4.3 M_{\mathrm{oc},\oplus}$ is satisfied. Using the expression for M_u defined in the above paragraph, this amounts to stating that $\mathcal{F}_w > 1.9 \times 10^{-3}$ must be valid. This result is surprisingly close to the estimate 2×10^{-3} that was obtained by means of a sophisticated theoretical model (Cowan & Abbot 2014). Likewise, the condition for desert worlds can be derived by setting $M_{\mathrm{oc}} = 0$ and assuming that the M_u is close to its saturation value. Hence, we find that the criterion $\mathcal{F}_w < 9.4 \times 10^{-4}$ may suffice to ensure that an Earth-analog is almost completely devoid of oceans eventually. In other words, when $9.4 \times 10^{-4} < \mathcal{F}_w < 1.9 \times 10^{-3}$ holds true, it seems plausible that such worlds would comprise both oceans and landmasses. Equipped with this data, we are now in a position to investigate two different probability distribution functions for $\mathcal{F}_w$.

The first model we consider is one where $\mathcal{F}_w$ is uniformly distributed between 10^{-12} and 0.5. Our results remain mostly unchanged if the lower bound is adjusted from 10^{-12} to any value lying between zero and 10^{-4}. Using the data presented earlier, we arrive at the following conclusions. The fraction of Earth-analog water worlds (i.e., sans any landmasses on the surface) equals $\sim (0.5 - 1.9 \times 10^{-3})/(0.5) \sim 0.996$. The fraction of worlds with a mixture of landmasses and oceans on the surface is $\sim (1.9 \times 10^{-3} - 9.4 \times 10^{-4})/(0.5) \sim 2 \times 10^{-3}$. Finally, the fraction of Earth-analog desert worlds with no oceans on the surface equals $\sim (9.4 \times 10^{-4})/(0.5) \sim 2 \times 10^{-3}$. Hence, the uniform probability distribution function predicts that the overwhelming majority of worlds (99.6 percent) would be ocean worlds. However, it must be recognized that this outcome arises because the uniform probability distribution is strongly biased to favor worlds with higher water fractions.

The second model that we analyze is a log-uniform distribution, wherein $\log \mathcal{F}_w$ (instead of $\mathcal{F}_w$) is distributed uniformly. The chief advantage

with employing this distribution function is that it mitigates the aforementioned bias toward water worlds that was inherent in the uniform distribution. On the other hand, the lower bound must be nonzero to avoid singular behavior. The fraction of Earth-analog desert worlds is given by $\sim \left(\log(9.4 \times 10^{-4}) - \log(10^{-12})\right) / \left(\log(0.5) - \log(10^{-12})\right)$, which translates to ~ 0.77. Repeating this procedure, the fraction of worlds with both landmasses and oceans equals 0.03 while the fraction of ocean worlds without landmasses is ~ 0.2. Hence, as per this model, we see that the distribution is roughly bimodal, in the sense that the fractions of water worlds and desert worlds are roughly comparable to one another, whereas the number of Earth-analogs with both landmasses and oceans is lower (~ 3 percent).

Now we shall attempt to calculate the fraction of worlds that have $0.3 < f_w < 0.9$, since they are anticipated to have the highest likelihood of Earth-like biospheres, based on the analysis undertaken previously. Noting that $f_w \approx 0.7$ yields $M_{\rm oc} \approx M_{\rm oc,\oplus}$ and hypothesizing that the average water depth of the oceans remains unaltered, we require $4.4 M_{\rm oc,\oplus} < M_W < 5.3 M_{\rm oc,\oplus}$. In terms of $\mathcal{F}_w$ for Earth-analogs, this range can be expressed as $10^{-3} < \mathcal{F}_w < 1.2 \times 10^{-3}$. For a uniform probability distribution, we find that the fraction of such worlds is $\sim 4 \times 10^{-4}$. On the other hand, using a log-uniform probability distribution function for $\mathcal{F}_w$, the corresponding fraction is $\sim 7 \times 10^{-3}$. Hence, at least insofar as these two models are concerned, it might be plausible that true Earthlike biospheres are uncommon, with abundances on the order of 0.01 to 0.1 percent.

Tian and Ida (2015) developed a detailed numerical model that accounts for the initial water inventories as well as the depletion of water due to XUV radiation during the pre-main-sequence phase. Their results are presented in Figure 5.10. For stars with $0.3 M_\odot$, it is clear that the distribution of water fractions is highly bimodal: water worlds make up ~ 56 percent of the population, whereas the fraction of desert planets is approximately 42 percent. In other words, worlds with Earthlike water fractions ($10^{-3} \lesssim \mathcal{F}_w \lesssim 10^{-4}$) are uncommon ($\lesssim 2$ percent). This bimodal behavior is also observed for planets in the HZ of K-type stars, whose mass is $0.5 M_\odot$. We find that the fraction of desert worlds is ~ 77 percent whereas the corresponding value for ocean planets is ~ 21 percent. Thus, once again, we see that Earthlike water fractions are rare ($\lesssim 2$ percent). On the other hand, when we consider planets in the HZ of Sunlike stars, the fraction of worlds with Earthlike water fractions is somewhat higher (~ 6 percent). The majority of planets are characterized by $10^{-7} \lesssim \mathcal{F}_w \lesssim 10^{-5}$. If this were to represent the total

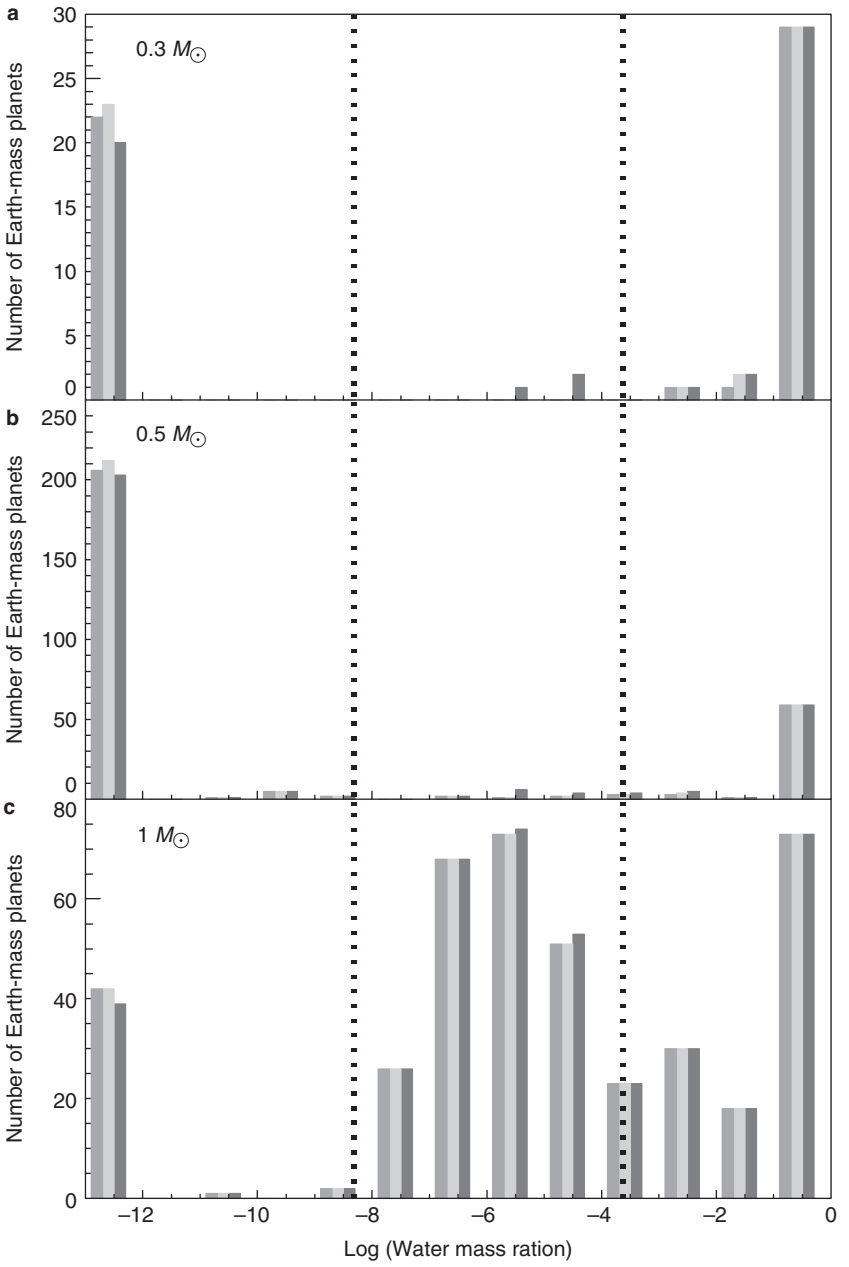


Figure 5.10 Distribution of water fractions for Earth-mass planets in the habitable zone 90 Myr after the star has entered the pre-main-sequence phase. The three bars correspond to different XUV luminosities and photolysis efficiencies. The rightward and leftward vertical dotted lines correspond to the surface water fraction of our planet and the detected water fraction in the atmosphere of Venus, respectively. (© Macmillan Publishers Limited. *Source:* Feng Tian and Shigeru Ida [2015], Water contents of Earth-mass planets around M dwarfs, *Nature Geoscience* 8[3]: 177–180, fig. 2.)

water fraction (not just on the surface), it would imply once again that desert worlds dominate.

Hence, based on the fact that K- and M-type stars are more common than G-type stars by roughly an order of magnitude, it seems plausible that most planets will fall into the category of either desert or water worlds. It would therefore appear that the log-uniform distribution for the water fraction is reasonable to a certain degree. Lastly, both theoretical and numerical models discussed herein suggest that the fraction of worlds with Earthlike water inventories are relatively uncommon in the Universe.

5.5.5 Summary

We have examined the characteristics of putative extraterrestrial biospheres by supposing that the availability of water limits biological productivity on land and that the access to dissolved phosphorus compounds plays an equivalent role in the oceans. We arrived at the following conclusions:

- The net primary productivity (NPP) and total biomass approach zero in the limit $f_w \to 0$ but remain nonzero (albeit small) when $f_w \to 1$. This behavior is manifested because surface water is essential for life, whereas ocean planets are theoretically capable of sustaining oligotrophic (nutrient-poor) biospheres.
- Neither the NPP nor total biomass decline significantly across a broad range of δ_w (ratio of surface land and ocean fractions). Their maxima occur at values of δ_w close to that of the Earth, ostensibly implying that our planet's topography is nearly optimal in this respect. These two functions are also found to have a moderate dependence on the radius R of the world.
- NPP regulates the buildup of atmospheric O_2, provided that the majority of carbon fixation takes place through oxygenic photosynthesis. For such worlds that are Earth-sized, we found that only a narrow range of surface water fractions (23 to 88 percent) might enable the buildup of molecular oxygen in the atmosphere to levels that are eventually detectable.
- The likelihood of technological intelligence was conceivably highly sensitive to the size of the world, with larger worlds being much more advantageous in this regard. This function is also strongly dependent on δ_w and vanishes in the limits $f_w \to 0$ and

$f_w \to 1$ because some evolutionary steps require the existence of landmasses whereas others necessitate oceans.

- When it comes to Earth-sized worlds, surface water fractions in the range of 30 to 90 percent may, perhaps, ensure that the evolution of technological intelligence has a reasonable likelihood of success compared to the Earth. The Earth's value of δ_w is apparently close to optimum insofar as the emergence of technological species is concerned, although super-Earths with similar surface water fractions are potentially superhabitable.
- Worlds with a mixture of landmasses and oceans on the surface are possibly rare, although not exceptionally uncommon, in agreement with previous studies.

Our results are expected to be more reliable in the regimes $\delta_w \ll 1$ and $\delta_w \gg 1$ compared to the scenario where $\delta_w \sim 1$.

Before rounding off our analysis, some of the associated caveats are worth mentioning. Our model comprised only two parameters: the surface water fraction (f_w) and the size of the world (R). Needless to say, multitudinous abiotic (e.g., ocean pH and photon flux) and biotic (e.g., biomass density and turnover timescales) factors were not taken into consideration. Incorporating all of these variables and quantifying the ensuing results self-consistently is a highly challenging endeavor, given that we must solve a system of coupled nonlinear partial differential equations. To offer one specific example, we have presupposed that the phosphorus sinks and sources can be treated in isolation. In reality, the oceanic P inventory depends on the concentration and oxidation state of iron, which in turn depends on O_2 abundance in the atmosphere and oceans.

In deriving our formulae, the majority of the planet's surface area was implicitly regarded as being habitable, modulo the existence of liquid water, but the actual extent of habitable area depends on several stellar and planetary factors—for example, tidal locking and temperature. Furthermore, this analysis was primarily centered on the spatial element and therefore did not investigate temporal aspects. It was thus presumed that the worlds studied herein are habitable for sufficiently long intervals to permit microbial life and technological intelligence to evolve. Yet, as we have seen in Chapter 4, a number of potential issues arise when it comes to low-mass stars ranging from the fluxes of ultraviolet radiation and energetic particles to atmospheric erosion driven by stellar winds and flare-related phenomena.

Another noteworthy time-dependent feature is that f_w evolves temporally due to atmospheric escape and the transport of water from oceans to the underlying mantle. As a result, the duration over which $\delta_w \sim 0.1$–1 holds true for a given world may influence its propensity for building up atmospheric O_2 and evolving technological intelligence significantly. Our planet was probably lucky in this regard, because this criterion could have been fulfilled since at least ~ 3 Ga (see Figure 5.6; see also Korenaga 2018). In fact, even worlds with Earthlike masses and water inventories are not guaranteed a priori to evolve toward the Earth's present-day ocean-to-land ratio as per certain numerical models (Höning et al. 2019).

Although the preceding issues cannot be neglected sensu stricto, the chief advantage inherent in our models and their accompanying predictions is that they are falsifiable in the decades to follow. For starters, the ocean and land fractions are, at least in principle, quantifiable via spectrophotometric observations, as described in Section 6.2.1. Likewise, much progress has been made in identifying atmospheric signatures of biological O_2 and distinguishing it from abiotic O_2, as we shall see in the upcoming chapter. Thus, in the event that biogenic O_2 is detected primarily on worlds with $f_w \in (0.23, 0.88)$, it may serve to validate our model. A number of avenues are subsequently adumbrated for identifying signatures of technological intelligence (to wit, technosignatures) in Chapter 9, which are harder to misidentify prima facie. Hence, if worlds with such signatures were to be characterized by $f_w \in (0.3, 0.9)$, it might accord some credibility to our model. It is, however, equally important to recognize that technological species and their concomitant technosignatures can exist on outwardly uninhabitable worlds, especially if they are post-biological in nature.

To conclude, there are multiple grounds for contending that $\delta_w \sim 1$ on Earth is no mere coincidence, because it optimizes many facets of our biosphere, including the buildup of O_2 in the atmosphere and the advent of technological intelligence. In Chapter 3, we saw that both these events drove major evolutionary changes on Earth, and one can therefore conjecture that they would have similar effects on other worlds. Due to the fact that worlds with $\delta_w \sim 1$ are ostensibly uncommon ($\lesssim 0.1$–1 percent), one might be tempted to interpret this result as bolstering the *Rare Earth* hypothesis, which posits low numbers of technological species in the Galaxy. Yet, we emphasize that this inference is not robust in light of the manifold unknowns involved. It is worth remembering at this juncture that the total number of stars in the Milky Way is very high ($\sim 10^{11}$). Hence, even in case the fraction

of stellar systems hosting complex biospheres with high productivity was small, the overall numbers for the latter may still be significant.

5.6 THE DISTRIBUTION OF LANDMASSES AND OCEANS

Until now, we have restricted ourselves to examining how the partitioning of the surface into landmasses and oceans governs key biological traits. Our description was, by necessity, coarse-grained since the only quantities of interest to us were the areas spanned by oceans and landmasses, corresponding to $4\pi f_w R^2$ and $4\pi(1-f_w)R^2$, respectively; strictly speaking, we were primarily interested in non-arid regions, whose area was $4\pi f_w(1-f_w)R^2$ using (5.51).

It becomes apparent that this approach yields promising results, but only to leading order. To understand why crucial higher-order effects are not captured, let us focus only on landmasses henceforth since this treatment can be appropriately generalized to study oceans. The total surface area comprising land will be denoted by $A_{\rm land} = 4\pi(1-f_w)R^2$. Now, one is free to envision the landmasses being distributed in any number of ways. We consider two cases to illustrate our point. In the first, nearly all of the available land could be aggregated to form a single supercontinent; this setup is reminiscent of Pangaea, which is known to have existed between ~ 200 and 300 Ma. Alternatively, as opposed to just a few continents, we may imagine a large number of fragmented landmasses analogous to our planet's archipelagos. These two scenarios ineluctably raise the question: What distribution of landmasses is more optimal from the standpoint of biological habitability?

A brief historical digression is worth pursuing here. The above question, mutatis mutandis, was the subject of intense debate in the 1970s and 1980s, after which it gradually receded into the background. In the context of conservation biology, it came to be reframed in the following fashion: Is a single large patch or several small patches (with the same total area) more effective at preserving greater biodiversity? The ensuing discourse came to be termed the SLOSS (single large or several small) debate. Early studies, taking their cue from Jared Diamond's influential paper on biological conservation (Diamond 1975), argued in favor of the single large patch approach (i.e., SL > SS). However, a distinct majority of empirical studies in this realm ($\gtrsim 70$ percent) appear to support the opposite conclusion (viz., SS > SL; Fahrig 2020). We note, however, this convoluted issue is far from being definitively settled, owing to which our subsequent discussion should be

treated with due caution. Needless to say, unearthing rigorous answers to this conundrum has acquired real significance in our era, with its ongoing mass extinction driven to a substantive degree by massive habitat losses (E. O. Wilson 2016).

Before attempting to answer our original question more quantitatively, a few qualitative observations are in order. First, we observe that it may be relatively problematic for precipitation to occur in the central regions of supercontinents due to the increased distance from the oceans. Climate models predict that large swathes of the Pangaean interior were arid or semiarid, although the extent tends to vary from one geological period to another (Parrish 1993). However, it does not automatically follow that the interior regions of supercontinents are always arid since rainfall depends on a number of other factors including the wind speed, temperature, topography (e.g., mountain ranges), and even the type of vegetation. Second, islands are characterized by their high proportion of endemic species (unique to a given location) and therefore their "endemic richness" has been estimated to be nearly an order of magnitude higher compared to mainland regions (Kier et al. 2009). Third, fragmented landmasses yield a longer coastline compared to supercontinents for the same area. To see why this is the case, suppose that the total land area ($A_{\rm land}$) is broken into n_0 disjointed circular islands of equal radius r_c. In consequence, we obtain

$$r_c = 2R\sqrt{\frac{1-f_w}{n_0}}. \tag{5.82}$$

For a single supercontinent of area $A_{\rm land}$ that is roughly circular, the length of the coastline is proportional to R. In contrast, for the n_0 islands considered above, the total coastline length should be proportional to $n_0 r_c$. After substituting (5.82), we find that this length is proportional to $\sqrt{n_0}R$. Hence, depending on the value of n_0, the total coastline may be expanded considerably relative to worlds with only one or more supercontinents. The importance of this simple calculation stems from the fact that coastal regions on Earth are typically characterized by their high biodiversity on account of their capacity to host multiple habitats (J. S. Gray 1997).

In light of the preceding discussion, there appear to be sufficient grounds for hypothesizing that worlds with fragmented landmasses could manifest a higher biological potential in comparison to their counterparts dominated

by supercontinents. To quantify this hypothesis further, we must turn our attention to the species-area relationship (SAR), one of the cornerstones of biodiversity studies. Although it is widely accepted that a larger area is qualitatively correlated with a higher number of species, the precise mathematical representation of this relationship varies from one study to another. The most widely employed SAR takes on the form

$$S_r = C_0 A_r^z, \tag{5.83}$$

where S_r and A_r represent the number of species and area of a given region, respectively, while C_0 and z serve as free parameters in this model. The most important point to recognize here is that the power-law exponent z is *not* constant since it depends on the spatial scale of the region(s) under consideration. Excluding some rare cases, the value of z typically lies between 0 and 1 (Rosenzweig 1995). In theory, $z \to 1$ is predicted to occur when the sample area approaches infinity (Hubbell 2001). From a more realistic standpoint, z appears to be close to unity for biological provinces, i.e., self-contained regions wherein species arise almost exclusively due to local speciation. While dealing with such landmasses, using the fiducial value $z \approx 0.8$ is fairly reasonable (Rosenzweig 1995), but we shall leave z unspecified for the time being.

Let us denote the total species richness for a single supercontinent by S_{SC}. We can calculate S_{SC} from (5.83) by specifying $A_r = A_{\text{land}}$. We end up with $S_{SC} \propto R^{2z}$, implying that the scaling with respect to the size of the world is not particularly noteworthy owing to $z \lesssim 1$. Next, consider the scenario in which the total area comprises n_0 provinces of equal size. The corresponding species richness S_I is estimated from (5.83), thereby yielding

$$\frac{S_I}{S_{SC}} = n_0^{1-z}, \tag{5.84}$$

implying that, as long as $0 < z < 1$, the number of species will be higher for worlds with fragmented landmasses. From a qualitative standpoint, this finding is supported by the available data for the genus richness of Phanerozoic (0–0.54 Ga) marine invertebrates (Zaffos et al. 2017). The limiting case of $z = 1$ is special because the above ratio remains constant regardless of the number of provinces. Now, if we select $n_0 = 30$ and $z \approx 0.8$, we end up with $S_I/S_{SC} \approx 2$. In principle, this ratio can become $\gtrsim 10$, but it will require an

unrealistically high number of provinces to exist. In reality, the number of fragments will be constrained by the geological properties (e.g., plate tectonics) of the planet, and it should also be remembered that the exponent z will change in magnitude as the area of each province is decreased.

Hitherto, we have concerned ourselves only with fragments of the same size, but this idealization can be generalized as follows. Suppose that the total number of provinces (N_p), total area of landmasses, and total number of species (S_{tot}) are all held fixed on a given world. We introduce the notation n_i that quantifies the number of fragments with area A_i. It is more transparent to work with $p_i = n_i/N_p$, which represents the probability associated with randomly selecting a province with area A_i from the entire collection. Our aim is to determine the *least biased estimate* for the probability (p_i), based on the minimal information provided above. In order to do so, we draw on the famous Principle of Maximum Entropy (MaxEnt) to infer p_i accordingly. MaxEnt was founded on the radical, but simple, idea that the Shannon entropy $H = -\sum_i p_i \ln p_i$ should be maximized subject to holding the relevant constraints fixed (Jaynes 2003). In our case, we have

$$\sum_i p_i = 1, \tag{5.85}$$

$$\sum_i p_i A_i = \frac{A_{\text{land}}}{N_p}, \tag{5.86}$$

and

$$\sum_i p_i A_i^z = \frac{S_{\text{tot}}}{\mathcal{C}_0 N_p}, \tag{5.87}$$

which embody the constraints on the total (1) number of fragments, (2) area, and (3) number of species. In order to obtain (5.87), we made use of (5.83), but this is easily replaceable with any other SAR. Therefore, the function that we seek to maximize is

$$-\sum_i p_i \ln p_i - \Lambda_P \left(\sum_i p_i \right) - \Lambda_A \left(\sum_i p_i A_i \right) - \Lambda_S \left(\sum_i p_i A_i^z \right), \tag{5.88}$$

where the Λ's denote the appropriate Lagrange multipliers. Taking the derivative of this function with respect to p_i yields

$$p_i = \exp\left[-\left(\Lambda_P + 1 + \Lambda_A A_i + \Lambda_S A_i^{\tilde{z}}\right)\right]. \tag{5.89}$$

It is straightforward to repeat the same analysis by taking the continuous limit of these variables and solving for the probability distribution function (PDF) instead. The only notable differences are $A_i \to A$ and $p_i \to p(A)$, with A and $p(A)$ signifying the province area and the PDF, respectively. An interesting point about this ansatz is that p_i may become nonmonotonic with respect to the area A_i if the parameters Λ_A and Λ_S are opposite in sign; the same result also applies to $p(A)$ with regard to A. In other words, depending on the Lagrange multipliers, prescribed in turn by the specific constraints and parameters of the chosen system, the PDF could attain an extremum at some optimal area. An alternative approach is to calculate the mean time for the extinction of this network of fragments and maximize it, thus yielding an optimal area and number of fragments (Ovaskainen 2002).

To sum up, our central conclusion is that the biodiversity of habitable worlds ought to increase, until a certain limit, commensurately with the number of fragmented landmasses, with the total land area held constant. While upcoming telescopes are capable of resolving the ocean and land fractions, it will nevertheless be much harder to empirically investigate the correlation between biological potential and the distribution of landmasses (or oceans).

5.7 LIFE IN THE ATMOSPHERE

Until this point in our discussion, we have taken for granted that life exists in terrestrial or aquatic habitats. Yet, as we know, the Earth has not only landmasses and oceans but also its atmosphere. Is it conceivable, perhaps, that life may exist in the atmospheres of other worlds? This question acquires an added relevance when the planet either lacks a well-defined surface (e.g., Jupiter) or is otherwise uninhabitable (e.g., Venus). There has been a general tendency to dismiss the atmosphere as a possible abode for life because it does not have a solid substrate for biochemical phenomena to freely operate, experiences extreme temperatures, and is concomitantly subject to higher fluxes of ultraviolet radiation, X-rays, and energetic particles; many of these impediments are believed to be particularly pronounced for the stratosphere and layers above it (Smith et al. 2011).

However, writing off atmospheres altogether seems premature. At favorable altitudes, physicochemical conditions could permit the existence

of liquid water, moderate temperatures and pressures, and nutrients. Observations have revealed that the cloud layer in the Venusian atmosphere at the height of 47.5 to 50.5 km has a temperature of ~ 323 K, a pressure of ~ 1 atm, and the existence of μm-sized sulfuric acid aerosols. Hence, ever since the pioneering proposal by Harold Morowitz and Carl Sagan more than five decades ago (Morowitz & Sagan 1967), the Venusian atmosphere has been considered by scientists to constitute a potential site for life; we refer the reader to Cockell (1999), Limaye et al. (2018) and Seager et al. (2021) for meticulous reviews of this subject.

The marginalized field of Venusian astrobiology received an unexpected impetus in September 2020 due to the contentious detection of phosphine gas at a concentration of ~ 1 ppb in the temperate cloud decks (Greaves et al. 2020); the spectroscopic evidence remains equivocal at the time of writing this monograph. As phosphine is produced by biological activity on Earth, it has been theorized that this gas might be indicative of life in the atmospheres of Venus and anoxic terrestrial exoplanets, although the existence of hitherto unknown abiotic mechanisms cannot and ought not be dismissed without extensive research. The most unambiguous means of settling the question of whether our sister planet hosts life in its atmosphere is to carry out in situ analyses or return samples to Earth. One such low-cost astrobiology mission to Venus, endowed with the capacity to seek biosignatures of prospective Venusian lifeforms, was delineated in Hein et al. (2020).

Aside from Venus, it is possible in theory that Jupiter may host conditions suitable for exotic life in its atmosphere. This unorthodox hypothesis was mooted around five decades ago by distinguished astronomers such as Harlow Shapley, Carl Sagan, and Edwin Salpeter. The latter duo, in particular, formulated theoretical models to describe the motion of putative microbes and macroscopic organisms in the Jovian atmosphere as well as the synthesis of organic molecules (Sagan & Salpeter 1976). Laboratory experiments indicate that the synthesis of a number of valuable prebiotic compounds such as amino acids, formaldehyde, and hydrogen cyanide is feasible in Jovian atmospheres under the action of ultraviolet radiation.

Looking beyond giant planets akin to Jupiter, we enter the domain of brown dwarfs. The latter are often referred to as *failed stars,* but the dividing line between brown dwarfs, low-mass stars, and giant planets is fuzzy. Brown

dwarfs cannot burn hydrogen via the so-called proton-proton reaction to create helium along the lines of normal stars. The conventional definition of brown dwarfs, bearing the above caveat in mind, is objects that burn deuterium (d) to yield helium-3 via the following schematic reaction: d + p $\rightarrow$ ^{3}He, where "p" refers to a proton. By this definition, brown dwarfs span the mass range of $\sim 10M_J$ to $\sim 75M_J$, where $M_J \sim 9.5 \times 10^{-4} M_\odot$ is the mass of Jupiter. Objects with masses below $10M_J$ are often referred to as sub–brown dwarfs by some authors and as giant planets by others. The reason we have touched on brown dwarfs is because their atmospheres are known to share some commonalities with those of giant planets (Marley & Robinson 2015).

Over the past two decades, the detection of cool brown dwarfs has proceeded apace. In particular, a number of Y-dwarfs—that is, cool brown dwarfs at temperatures of < 500 K—have been detected. A widely studied Y-dwarf is WISE J085510.83-071442.5, which has an effective temperature of $T_{\rm eq} \sim 250$ K and is situated at a distance of 2 pc from the Sun (Luhman 2014). In the spirit of the above discussion, we can speculate about whether brown dwarfs could also host life in their atmospheres. This notion was briefly broached by Harlow Shapley in his visionary book *Beyond the Observatory* (1967), but the first quantitative treatment of this topic was presented in Yates et al. (2017).

We will now explore some of the constraints on putative life in the atmospheres of brown dwarfs. We opt to focus on brown dwarfs instead of giant planets or Venus-analogs for two reasons. First, some of the details outlined herein are also applicable to planets. Second, the atmospheric habitable zones of brown dwarfs are larger and more long-lived. It was recently demonstrated by Lingam and Loeb (2019a) that the total spatiotemporal habitable volume encompassed by brown dwarf atmospheres might be two orders of magnitude higher than the corresponding volume for Earthlike planets in the habitable zones of main-sequence stars; we will not delve into the derivation as it involves many empirical and theoretical details. Hence, if this model is accurate, it is conceivable that the most abundant sites for life in the Universe could turn out to be the atmospheres of brown dwarfs.

As this part of the chapter is mostly independent of the rest, those who wish to skip this discussion may jump directly to Section 5.8. Much of our discussion is drawn from Lingam and Loeb (2019a), and readers should consult this work for additional references and further details.

5.7.1 The prospects for life in brown dwarf atmospheres

We will now examine some of the salient features relating to habitability and hypothetical lifeforms in brown dwarf atmospheres. We use the subscript "BD" to denote quantities pertaining to brown dwarfs.

5.7.1.1 Sizes of putative organisms

When an object is falling through the atmosphere, it experiences two forces to leading order: gravity and drag. By balancing these two forces, it is easy to solve for the terminal velocity v_t. The gravitational force is $m_0 g_{\rm BD}$, where m_0 is the organismal mass and $g_{\rm BD}$ is the brown dwarf's gravitational acceleration. The drag force is computed from Stokes' law (named after its discoverer, the eminent nineteenth-century physicist Sir George Stokes) and equals $6\pi\rho_f \nu_f r_0 v_t$, where ρ_f and ν_f are the density and kinematic viscosity of the fluid, respectively, while r_0 is the size of the organism. Thus, after equating the two forces, we have

$$r_0^2 \rho_0 \sim \frac{9 v_t \nu_f \rho_f}{2 g_{\rm BD}}, \tag{5.90}$$

where ρ_0 is the density of the organism; we have modeled the organism as a spherical particle. We can solve for v_t from the above formula. To leading order, we may suppose that this downward velocity associated with gravitational settling is balanced by convection, which is known to play a key role in the atmospheres of giant planets and brown dwarfs; in other words, we work with $v_c \sim v_t$, where v_c signifies the convective velocity. Thus, after rearranging this expression, we find

$$r_0 \sim 135.5\,\mu{\rm m} \left(\frac{v_c}{1\,{\rm m/s}}\right)^{\frac{1}{2}} \left(\frac{\rho_0/\rho_f}{10^3}\right)^{-\frac{1}{2}} \left(\frac{M_{\rm BD}}{M_J}\right)^{-\frac{5}{6}} \times \left(\frac{\nu_f}{10^{-5}\,{\rm m^2\,s^{-1}}}\right)^{\frac{1}{2}}, \tag{5.91}$$

with $M_{\rm BD}$ representing the mass of the brown dwarf and the expression for $g_{\rm BD}$ adopted from Burrows and Liebert (1993). Holding all other parameters fixed, even if we work with an upper bound of $M_{\rm BD} \approx 75 M_J$ for brown dwarfs, we obtain $r_0 \sim 3.7\ \mu$m. In contrast, the sizes of most microorganisms

on Earth are $\lesssim 1\ \mu$m. Hence, as per this simple theoretical model, microbes up to sizes that are ~ 10 times larger than those commonly found on Earth may be able to survive in the atmospheres of brown dwarfs (Yates et al. 2017).

5.7.1.2 Biomass in atmospheric habitable zones

Next, we can attempt to estimate the upper bound on the potential biomass in brown dwarf atmospheres. Needless to say, the lower bound will be zero in the event that life is excluded in aerial settings.

The total atmospheric biomass ($M_{\rm bio}$) is estimated to be

$$M_{\rm bio} \sim 4\pi \rho_m R_{\rm BD}^2 \mathcal{H}_{\rm BD}, \tag{5.92}$$

where ρ_m is the typical density of microbes in the atmosphere, $R_{\rm BD}$ is the radius of the brown dwarf, and $\mathcal{H}_{\rm BD}$ is the "width" of the atmospheric habitable zone where the microbes may exist. The brown dwarf radius derived from Burrows and Liebert (1993) is given by

$$R_{\rm BD} \approx 35.6 R_\oplus \left(\frac{M_{\rm BD}}{M_J}\right)^{-1/3}, \tag{5.93}$$

whereas the width of the atmospheric habitable zone, based on Burrows and Liebert and Yates et al. (2017), is expressible as

$$\mathcal{H}_{\rm BD} \sim 2.6 \times 10^3\ {\rm km} \left(\frac{M_{\rm BD}}{M_J}\right)^{-5/3}. \tag{5.94}$$

The remaining variable in (5.92) is ρ_m, which is much harder to determine. After a detailed analysis of the density of aerosols in different planets and moons of our Solar system, the fraction that could host microbes, and the masses of hypothetical microbes, Lingam and Loeb (2019a) argued that the values for ρ_m span a very wide range. The lower bound proposed in this study was $\rho_{\rm min} \sim 2.5 \times 10^{-11}\ {\rm kg/m^3}$, while the upper bound was $\rho_{\rm max} \sim 10^{-4}\ {\rm kg/m^3}$. From the geometric mean of these two quantities, we have $\bar{\rho} \sim 5 \times 10^{-8}\ {\rm kg/m^3}$. For these densities, we can calculate (5.92) accordingly. It makes more sense, however, to normalize the total biomass

with respect to the Earth, i.e., we introduce the ratio

$$\delta_{\rm bio} \equiv \frac{M_{\rm bio}}{M_{\rm bio,\oplus}}, \tag{5.95}$$

where $M_{\rm bio,\oplus} \sim 1.1 \times 10^{15}$ kg after converting the estimate for biogenic carbon on Earth into the total biomass (Bar-On et al. 2018).

The ratio (5.95) is plotted as a function of $M_{\rm BD}$ in Figure 5.11 for the three biomass densities described earlier. The maximum biomass declines with $M_{\rm BD}$ because the radius and width of the habitable layer decline with the brown dwarf mass. By inspecting Figure 5.11, we see that $\delta_{\rm bio} \gg 1$ for all values of $M_{\rm BD}$ when the optimistic estimate for the biomass density ($\rho_{\rm max}$) is utilized. Instead, if we work with the lower bound ($\rho_{\rm min}$), we see that $\delta_{\rm bio} \ll 1$, although the total biomass is still significant when viewed in absolute terms. When we invoke the mean biomass density ($\bar{\rho}$), it is found that an interesting transition from $\delta_{\rm bio} > 1$ to $\delta_{\rm bio} < 1$ occurs at $M_{\rm BD} \approx 6.5 M_J$.

5.7.1.3 Aerosols and the origin of life

We spent a considerable amount of time on the origin-of-life question in Chapter 2, owing to which we shall only sketch some of the possible benefits accruing from aerosols serving as the sites for abiogenesis.

Some of the major qualitative advantages inherent to aerosols, in their capacity as prebiotic reactors, include the following:

- Observations and laboratory experiments have confirmed the emergence of aerosols with inverted micelle structures near water-air interfaces on Earth. Aerosols belonging to this category are composed of liquid water, minerals, and small organic molecules collectively enclosed inside an organic film made up of fatty acids.
- These structures resemble vesicles, although the lipid bilayers affiliated with the latter exhibit greater functionality. Vesicles may have played a vital role in the origin of protocells, by serving as compartments that permitted the replication of biopolymers in their interiors.
- As the aerosols are transported through different regions of the atmosphere, they would experience fluctuations in the ambient

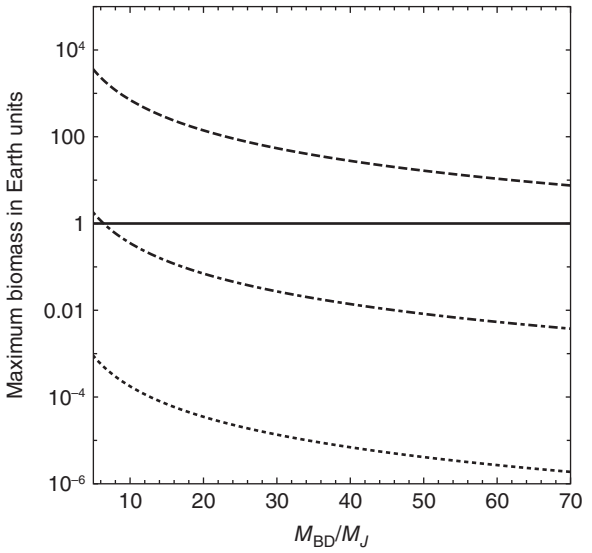


Figure 5.11 The potential upper bound on the biomass that may exist in the atmospheric habitable zones of cool brown dwarfs with $T_{eq} \sim 300$ K (normalized by Earth's biomass) as a function of the brown dwarf's mass (M_{BD}) measured in units of Jupiter's mass (M_J). The dashed, dotted-dashed, and dotted curves yield the biomass obtained from a biomass density of ρ_{max}, $\bar{\rho}$, and ρ_{min}, respectively. The solid horizontal line represents the limit where the atmospheric biomass is equal to the Earth's biomass. (© Manasvi Lingam and Avi Loeb.)

humidity. In particular, brown dwarf atmospheres with patchy cloud cover could give rise to such effects. This heterogeneity in the relative humidity conceivably enables the activation of wet-dry cycles that are ideal for the selection, concentration, and oligomerization of prebiotic molecules, as explained in Section 2.4.1.

- In conventional aqueous environments, the formation of peptide bonds (necessary for protein synthesis) is disfavored on thermodynamic and kinetic grounds. However, the same process is favorable at the air-water interfaces created by atmospheric aerosols (Griffith & Vaida 2012).
- Chemical reactions in aerosols resembling aqueous microdroplets might permit the synthesis of ribonucleosides (i.e., the ingredients of RNA) and sugar phosphates (Nam et al. 2018).

In view of the preceding points, aerosols with the inverted micelle structure might serve as prototypes for protocells. Given that cells have the capability to divide, we must ask ourselves whether aerosols can mimic protocells in this respect and undergo fission. However, as explained in Donaldson et al. (2001), the fission of "pure" aerosols is disfavored on thermodynamic grounds because the free energy is at a minimum. However, when it comes to "mixed" aerosols with organic films coating the surface, fission is theoretically possible. Let us suppose that the parent aerosol has a radius of r_P and a surface tension of γ_P. It splits into two daughter aerosols with radii r_1 and $r_2 \equiv \zeta r_1$, with surface tensions of γ_1 and γ_2, respectively; note that $\zeta \in (0, 1)$ by definition. As per the physics of liquids, the free energy is proportional to the product of the surface tension and the area, owing to which we require

$$\gamma_1 \left(4\pi r_1^2\right) + \gamma_2 \left(4\pi r_2^2\right) < \gamma_P \left(4\pi r_P^2\right) \tag{5.96}$$

in order for fission to become favorable on thermodynamic grounds. However, as the total mass must be conserved during this process, we impose the constraint

$$\frac{4\pi}{3} r_1^3 + \frac{4\pi}{3} r_2^3 = \frac{4\pi}{3} r_P^3, \tag{5.97}$$

where we have assumed that the aerosol density stays constant. Thus, by combining the above equations, we end up with

$$\gamma_1 + \gamma_2 \zeta^2 < \gamma_P \left(1 + \zeta^3\right)^{2/3}. \tag{5.98}$$

Now, suppose that we consider "pure" aerosols with $\gamma_1 = \gamma_2 = \gamma_P$. This leads us to $1 + \zeta^2 < \left(1 + \zeta^3\right)^{2/3}$, which cannot be fulfilled across the domain spanned by ζ. On the other hand, if we allow for variations in γ, satisfying the above inequality becomes viable. Let us begin by setting $\zeta \approx 1$. After applying (5.98), we have $\gamma_1 + \gamma_2 < 1.59\gamma_P$. For the sake of simplicity, suppose that $\gamma_1 \approx \gamma_2$ as the two droplets have approximately the same size. In other words, we must have $\gamma_1 < 0.8\gamma_P$, implying that the surface tension of the daughter droplets must undergo a sizable reduction. Instead, if we adopt $\zeta = 0.1$, we find that (5.98) is transformed into $\gamma_1 + 0.01\gamma_2 < 1.001\gamma_P$. This condition is relatively easy to fulfill, thus implying that the fission may exhibit a tendency toward asymmetry.

Next, we direct our attention toward the total number of trials feasible in atmospheric aerosols. As pointed out in Section 5.5.3, the likelihood of abiogenesis is sensitive to the total number of trials (N_T). We must take both the spatial and temporal aspects into consideration, although we focused only on the former in Section 5.5.3. Let us commence our exploration of N_T by tackling the spatial component first. In Section 2.6.1, we elucidated a famous model for the origin of protocells developed by Freeman Dyson. In this model, the emergence of an ordered low-entropy state (akin to life) from a disordered collection of monomers may necessitate around 10^{10} of them, i.e., a single trial may require the interplay of $\sim 10^{10}$ droplets.

The number density of aerosols (n_a) in the atmosphere is not only subject to variations in space and time for a given brown dwarf but also diverges from one world to another. However, for the sake of simplicity, we select a fiducial value of $\sim 10^7$ m^{-3} because it is consistent with observations of Jupiter's atmosphere. Therefore, the number of trials feasible in the spatial realm is $V_{\rm BD} n_a / 10^{10}$, where $V_{\rm BD} = 4\pi R_{\rm BD}^2 \mathcal{H}_{\rm BD}$. Now, let us turn our attention to the temporal aspect. The length of a single "cycle" to transition from a disordered to an ordered state is denoted by τ_C. In actuality, we have no knowledge whatsoever of τ_C, but we employ the normalization factor of ~ 100 s because this rough estimate was indirectly employed for the Earth in Tuck (2002). As a consequence, the total number of trials over the entire habitability interval is $\tau_{\rm BD}/\tau_C$, where the habitable duration ($\tau_{\rm BD}$) is (Lingam & Loeb 2019a)

$$\tau_{\rm BD} \approx 7.5\,{\rm Myr} \left(\frac{M_{\rm BD}}{M_J} \right)^{2.55} . \tag{5.99}$$

By combining all of these factors together, we obtain

$$N_T \sim 4 \times 10^{33} \left(\frac{M_{\rm BD}}{M_J} \right)^{0.22} \left(\frac{n_a}{10^7\,{\rm m}^{-3}} \right) \left(\frac{\tau_C}{10^2\,{\rm s}} \right)^{-1} . \tag{5.100}$$

In comparison, at the time when the first traces of life were preserved in the geological record of Earth, it has been estimated that $N_T \lesssim 10^{31}$ (Tuck 2002). Thus, as per (5.100), after adopting the fiducial values for n_a and τ_C, the maximum number of trials that are feasible in atmospheric aerosols of cool brown dwarfs is possibly ~ 3 orders of magnitude higher relative to our planet at the time of life's emergence.

5.7.1.4 Bioessential elements

In the earlier chapters of our book, we saw that the major bioessential elements are carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur. The availability of the first three elements ought not pose issues because of the relatively abundant availability of atmospheric methane and water vapor. Thus, we will focus on the remaining three bioessential elements for the most part, with a particular emphasis on phosphorus.

The majority of sulfur in the atmospheres of brown dwarfs is locked up in the form of hydrogen sulfide. Elemental sulfur may exist in the form of sulfur aerosols in the atmospheres of certain planets, with Venus being one example. It has been theorized that the existence of microbes analogous to *Acidithiobacillus ferrooxidans* is possible. In the absence of oxygen, *A. ferrooxidans* utilizes Fe^{3+} as an electron acceptor and oxidizes elemental sulfur, thereby giving rise to products like sulfuric acid. On worlds where sulfide is the dominant contributor, it would make sense for microbial metabolic pathways to employ sulfide anions. One of the best known among them is anoxygenic photosynthesis, which we encountered in Section 3.2. *Chlorobiaceae* (green sulfur bacteria) offer a classic illustration of this reaction network, in which hydrogen sulfide is oxidized to yield elemental sulfur. However, a potential drawback with regard to this pathway is the availability of CO_2; while CO_2 can exist in the lower layers of brown dwarf atmospheres, it is probably not abundant as one moves toward higher altitudes.

The next bioessential element we consider is nitrogen. Ammonia is one of the major atmospheric gases in cool brown dwarfs. It represents a positive aspect of brown dwarf atmospheres because it obviates the need for biological nitrogen fixation (biological synthesis of ammonia) by diazotrophs. On Earth, NH_3 is used by aerobic or anaerobic bacteria to generate nitrite, which is subsequently converted into nitrate and consumed by organisms as a nutrient. It is conceivable that analogous metabolic pathways could be functional in habitable atmospheres, but an essential requirement is the access to suitable oxidants like oxygen or nitrite. As the atmospheres of brown dwarfs are strongly reducing, any oxidants present therein may be rapidly consumed and thus undergo depletion.

The last, but by no means the least, bioessential element is phosphorus, whose importance was already underscored in Section 2.2.5. The major drawback concerning phosphorus on early Earth is that it was mostly locked up in the form of phosphate minerals, which are virtually insoluble in water.

For instance, fluorapatite ($Ca_5(PO_4)_3F$) is the standard example of a phosphate mineral. It has a solubility of 3×10^{-3} g / L in pure water at pH of 7 and temperature of 307 K. Let us now pivot to phosphorus content in cool brown dwarfs. At effective temperatures of ~ 300 K, under the assumption of equilibrium chemistry, most of the phosphorus is predicted to exist in the form of tetraphosphorus hexaoxide (P_4O_6) as per theoretical models.

The compound ammonium dihydrogen phosphate ($NH_4H_2PO_4$) is, however, of much more interest for reasons that will become apparent shortly. It is formed via the condensation of P_4O_6 and NH_3 as per the following reaction:

$$P_4O_6 + 4NH_3 + 10H_2O \rightarrow 4NH_4H_2PO_4 + 4H_2. \tag{5.101}$$

The temperature at which the condensation of $NH_4H_2PO_4$ occurs was estimated by Visscher et al. (2006) to be

$$T_c \approx \frac{10^4\,\mathrm{K}}{30 - 0.2\left(11 \log P_T + 15[X/H]\right)}, \tag{5.102}$$

where T_c is the condensation temperature, P_T is the pressure (in units of bar), and $[X/H]$ quantifies the metallicity. A brown dwarf with solar metallicity and a pressure of ~ 1 bar at a given altitude has a condensation temperature of $T_c \sim 333$ K. It is therefore apparent that cool brown dwarfs might have clouds comprising $NH_4H_2PO_4$. After analyzing the spectra of the Y-dwarf WISE J085510.83-071442.5 and detecting an obscuration of photon flux in the near-infrared, Morley et al. (2018) suggested that this feature could be explained by the prevalence of $NH_4H_2PO_4$ clouds at a pressure of ~ 10 bar.

One of the chief reasons we have highlighted $NH_4H_2PO_4$ is because it has a high solubility in water, unlike the standard phosphate minerals found on Earth. At room temperature (~ 300 K) and neutral pH, the solubility of ammonium dihydrogen phosphate is $\sim 4 \times 10^2$ g / L, implying that it is $\sim 10^5$ times more soluble than fluorapatite. In addition, it has been widely employed in studies of prebiotic chemistry; a few select examples are listed below.

- In the 1960s and 1970s, several experiments synthesized nucleotides and their oligomers (precursors of nucleic acids) by adding ammonium dihydrogen phosphate and heating the mixtures.

- Glycerol phosphates are notable precursors of complex lipids found in cell membranes. These compounds can be synthesized by heating a mixture of glycerol and ammonium dihydrogen phosphate. In a similar vein, the synthesis of phosphate amphiphiles, which are analogous to phospholipids in cell membranes, is feasible by using $NH_4H_2PO_4$.
- Adenosine triphosphate (ATP) is ubiquitous in cellular biology on Earth as it plays the role of energy "currency." Broadly speaking, polyphosphates (of which ATP is an example) are vital for Earth-based life in several respects. They can be generated by heating $NH_4H_2PO_4$ at temperatures of ~ 333 to 373 K.

Although all of these experiments made use of $NH_4H_2PO_4$, they were unable to identify plausible sources of this compound on early Earth. In contrast, ammonium dihydrogen phosphate clouds expected to occur in the atmospheres of cool brown dwarfs at pressures of ~ 1–10 bar constitute tenable sources of this valuable compound (Morley et al. 2018).

5.7.1.5 Photosynthesis in cool brown dwarfs

It will hardly come as a surprise that photosynthesis represents the dominant contributor to carbon fixation on our planet. We will thus explore the feasibility of photosynthesis in the atmospheres of cool brown dwarfs. Before proceeding further, it is important to distinguish between conventional oxygenic photosynthesis and hydrogenic photosynthesis, introduced in Section 5.4.2. We will tackle the photon requirements for both pathways herein.

Let us first consider free-floating brown dwarfs that do not receive any radiation from their companion star(s). In this case, the only source of photons is the brown dwarfs themselves. As the atmospheric habitable zones are situated at altitudes much smaller than R_{BD}, to leading order we can model the PAR flux by the blackbody flux emitted from the "surface" over a suitable range of wavelengths. The photon flux (Φ_{BD}) for the blackbody is

$$\Phi_{BD} = \int_{\lambda_{min}}^{\lambda_{max}} \frac{2c}{\lambda^4} \left[\exp\left(\frac{hc}{\lambda k_B T_{eq}} \right) - 1 \right]^{-1} d\lambda, \tag{5.103}$$

where we specify $T_{\rm eq} \approx 300$ K, as we are interested in cool brown dwarfs with effective temperatures in the range ~ 250–350 K. Both the maximum ($\lambda_{\rm max}$) and minimum ($\lambda_{\rm min}$) wavelengths viable for photosynthesis are difficult to assess for other worlds.

For now, let us adopt the limits $\lambda_{\rm min} \approx 0.35$ μm and $\lambda_{\rm max} \approx 1.1$ μm because these wavelengths are used by anoxygenic photoautotrophs on Earth. Subsequently, we will hold $\lambda_{\rm min}$ fixed as this choice is compatible with the inhibition of photosynthesis by UV radiation. After substituting these numbers into (5.103), we obtain $\Phi_{\rm BD} \approx 1.2 \times 10^6$ m^{-2} s^{-1}. In contrast, the minimum photon flux required for photosynthetic organisms on Earth is $\Phi_c \approx 1.2 \times 10^{16}$ m^{-2} s^{-1} (Wolstencroft & Raven 2002). This bound is believed to be fairly robust as it is based on biophysical factors as H^+ leakage and protein turnover. Hence, at first glimpse, it seems impossible for (Earthlike) anoxygenic photosynthesis to function in the atmospheres of free-floating brown dwarfs.

For the sake of completeness, the total photon flux ($\Phi_{\rm tot}$) emitted by the brown dwarf is

$$\Phi_{\rm tot} \approx 4.8c \left(\frac{k_B T_{\rm eq}}{hc} \right)^3 . \tag{5.104}$$

By evaluating $\Phi_{\rm tot}$ for $T_{\rm eq} \approx 300$ K, we end up with $\Phi_{\rm tot} \approx 1.3 \times 10^{22}$ m^{-2} s^{-1}. Interestingly, this value is actually ~ 6.5 times higher than the total solar photon flux incident on our planet. Hence, as far as total photon flux is concerned, the lower atmospheres of brown dwarfs may be awash in a higher photon flux in comparison to the Earth. Of course, the chief difference is that most of the emitted photons have wavelengths of ~ 10 μm.

Now, let us explore the prospects for unorthodox versions of oxygenic photosynthesis not found on Earth. Wolstencroft and Raven (2002) suggested that oxygenic photosynthesis could exploit longer wavelengths by harnessing more photons to carry out the fixation of carbon dioxide; this is equivalent to the production of O_2. The basic relationship between the photon number and the maximum wavelength was delineated in (4.58), in Section 4.3.5. It can be readily rewritten as

$$\chi \sim 2 \left(\frac{\lambda_{\rm max}}{0.7\,\mu{\rm m}} \right), \tag{5.105}$$

where χ effectively represents the number of photons required for fixing each CO_2 molecule in oxygenic photosynthesis; on Earth, we have $\chi = 2$.

If we allow for multiphoton schemes, the minimum photon flux required is transformed into $(\chi/2)\,\Phi_c$. By demanding that $\Phi_{\rm BD}$ should equal this lower limit and subsequently invoking (5.103), we find $\lambda_{\rm max} \approx 2.7\ \mu$m and $\chi \approx 7.8$. However, as χ must be an integer, we end up with $\chi = 8$ and $\lambda_{\rm max} = 2.8\ \mu$m. Therefore, for oxygenic photosynthesis to operate, a unique photosystem based on eight photons per electron transfer is presumably necessary.

At this juncture, some crucial caveats are worth highlighting. First, our analysis was founded on the premise that all photons in the range $\lambda_{\rm min} < \lambda < \lambda_{\rm max}$ reach the atmospheric layer where the photosynthetic organisms are present. Second, the minimum photon flux employed above is an ideal limit because it assumes 100 percent absorption by the photosystem(s). Third, if eight photons are truly required, the absorption of stray photons at higher energies may cause overheating and disrupt the photosynthetic apparatus. Finally, thermodynamics will impose strict constraints on the *efficiency* at which PAR is utilizable. The Carnot efficiency ($\mathcal{W}_C$) encountered in Section 4.3.7 is roughly an upper bound in most, albeit not all, instances. At the atmospheric temperature of T_a, the conventional expression for $\mathcal{W}_C$ is

$$\mathcal{W}_C = 1 - \frac{T_a}{T_{\rm eq}}. \tag{5.106}$$

For $T_a \approx 280K$ and $T_{\rm eq} \approx 310$ K, we find that the Carnot efficiency is only around 10 percent. In fact, the above formula predicts that the extraction of work is impossible when $T_a > T_{\rm eq}$. A more realistic treatment requires the use of exergy introduced in Section 4.3.7, but we shall not address it herein. The thermodynamic constraint enforced by the Carnot bound is rendered relatively unimportant when it comes to cool brown dwarfs that are companions of stars.

Due to the above caveats, it is very plausible that the values of $\lambda_{\rm max}$ and χ calculated previously are lower bounds. On the other hand, extraterrestrial photosynthetic machinery might exhibit a higher efficiency and functionality relative to Earth-based organisms as a result of having evolved in low-light conditions. Before proceeding further, we note that $\lambda_{\rm max}$ and χ can be duly estimated for hydrogenic photosynthesis by repeating the same procedure if the equivalent of Φ_c is known. However, in light of our ignorance, we opt to maintain consistency with Section 5.4.2 and suppose that (1) the minimum flux is the same as oxygenic photosynthesis and (2) two

photons of 1.5 μm are necessary. In other words, the only change is that we must replace 0.7 μm with 1.5 μm in (5.105). After working through the calculations, we find $\lambda_{\max} \approx 2.6$ μm and $\chi \approx 3.5$ for free-floating brown dwarfs. For the integral values of $\chi = 3$ and $\chi = 4$, we obtain wavelengths of approximately 2.25 μm and 3.0 μm, respectively.

We will now consider brown dwarfs that are situated near stars. By treating the star as a blackbody, the critical orbital radius (a_c) at which the PAR flux becomes equal to the minimum photon flux can be estimated for a given photosynthetic pathway and number of photons involved per electron. To avoid confusion, we will only work with oxygenic photosynthesis, whose $\lambda_{\max}$ and χ are related via (5.105). In order to find the critical orbital radius, we determine the photon flux at a_c and then equate it to $(\chi/2)\,\Phi_c$. After simplification of the resultant expression, we have

$$a_c \sim 387 \text{ AU} \left(\frac{L_\star}{L_\odot}\right)^{1/2} \left(\frac{T_\star}{T_\odot}\right)^{-1/2} \sqrt{\mathcal{I}(T_\star)}, \tag{5.107}$$

where the function $\mathcal{I}(T_\star)$ is defined to be

$$\mathcal{I}(T_\star) \approx \frac{2}{\chi} \int_{2\ell_1(T_\star)/\chi}^{\ell_2(T_\star)} \frac{x'^2\, dx'}{\exp(x') - 1}, \tag{5.108}$$

with the integration limits given by $\ell_1(T_\star) \approx 3.32\,(T_\star/T_\odot)^{-1}$ and $\ell_2(T_\star) \approx 7.12\,(T_\star/T_\odot)^{-1}$. It is, however, essential to recognize that this estimate for a_c in (5.107) fails to take into account the opacity of the brown dwarf atmosphere to the incoming stellar radiation. Hence, it is plausible that the actual value of a_c might be orders of magnitude smaller due to the attenuation introduced by clouds, hazes, and other species.

In Figure 5.12, the critical orbital radius a_c is plotted as a function of the stellar mass ($M_\star$) and the number of photons (χ) involved in photosynthesis; mass-luminosity and mass-temperature scalings from Section 4.3.5 were used to generate the plot. A scrutiny of Figure 5.12 reveals that χ does not alter our results substantially for stars with $M_\star > M_\odot$. However, upon considering $M_\star \sim 0.1 M_\odot$, we find $a_c \approx 1.9$ AU for conventional photosynthesis ($\chi = 2$), whereas we arrive at $a_c \approx 4.9$ AU for $\chi = 4$. Hence, multiphoton schemes may facilitate an expansion of the photosynthesis zone around low-mass stars.

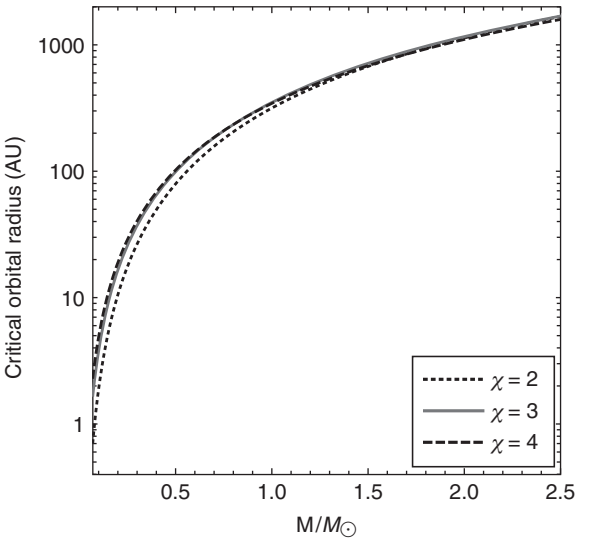


Figure 5.12 The upper bound on the orbital radius at which photosynthesis might be possible as a function of stellar mass ($M_\star$) measured in units of solar mass ($M_\odot$). The three curves represent the number of photons (χ) employed per electron transfer, with $\chi = 2$ for conventional oxygenic photosynthesis. (© Manasvi Lingam and Avi Loeb.)

5.7.2 Detecting life in brown dwarf atmospheres

In order to determine whether brown dwarf or planetary atmospheres host life, it is necessary to identify biological signatures that are discernible by remote sensing. This endeavor constitutes the subject of Chapter 6, due to which we will restrict ourselves to merely summarizing the basic details herein.

Let us first consider the presence of dead and decomposing organisms. In this scenario, it makes sense to search for the biomolecular building blocks of organisms. On Earth, it is well-known that many biomolecules exhibit peak absorption in the UV and visible regions. Nucleic acids and proteins display maximum absorption at wavelengths of 260 nm and 280 nm, respectively. Iron-sulfur clusters (Chapter 2) are vital actors in the cellular metabolic theater with peak absorbance at wavelengths of 280–450 nm. Biological pigments such as chlorophylls, carotenoids, and pterins are strongly absorbing at wavelengths < 500 nm. Thus, from the perspective of biomolecules, elevated reflectance (i.e., diminished absorption) is expected at wavelengths $\gtrsim 500$ nm. This behavior is consistent with Venus's spectra,

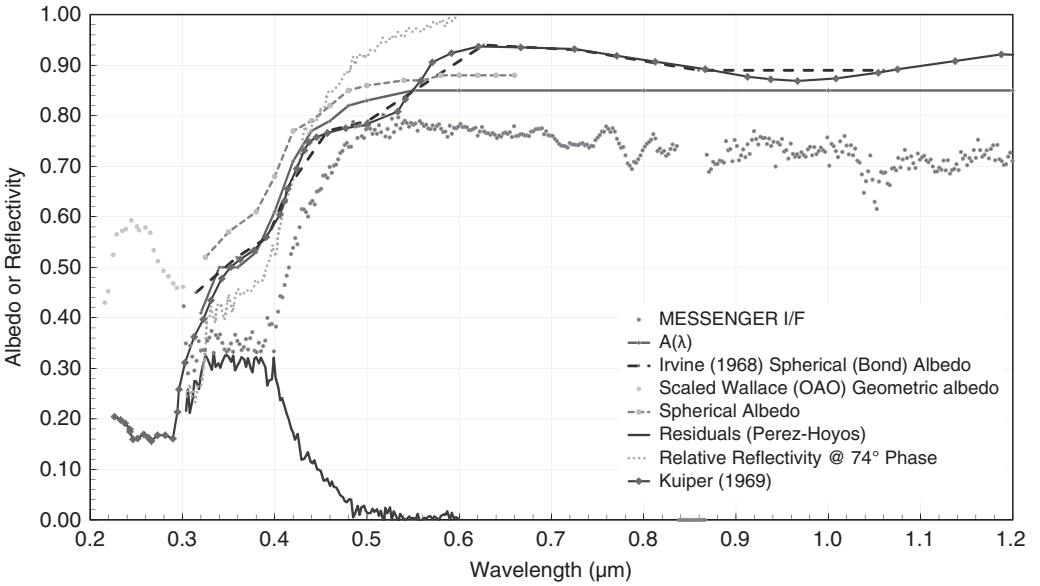


Figure 5.13 The spectra of the Venusian atmosphere, based on various observations. The bottom curve is meant to illustrate the unexplained absorption at UV and visible wavelengths. (© The Authors. CC-BY-NC-4.0. *Source:* Sanjay S. Limaye, Rakesh Mogul, David J. Smith, Arif H. Ansari, Grzegroz P. Słowik, and Parag Vaishampayan [2018], Venus's spectral signatures and the potential for life in the clouds, *Astrobiology* 18[9]: 1181–1198, fig. 2.)

which is depicted in Figure 5.13, consequently sparking suggestions that the planet's clouds might harbor microbes (Limaye et al. 2018).

Next, let us turn our attention to alive organisms. As we saw earlier, photosynthesis offers a potential avenue for biomass production and sustaining biospheres. In the case of free-floating brown dwarfs, the maximum wavelength at which photons are absorbed is $\sim$2.3–3.0 μm, based on our preceding calculations. Hence, discovering a distinct increase in reflectance (arising from a sharp decline in absorbance) at these wavelengths might be indicative of biogenic activity; this feature is called a *spectral edge*. When it comes to brown dwarfs near stars, as long as $a < a_c$ holds true, photosynthesis is feasible. The wavelengths at which spectral edges occur are dependent not only on the actual metabolic pathway (e.g., oxygenic vs. hydrogenic photosynthesis) but also on stellar properties. In spite of this variability, it is typically anticipated that photosynthetic spectral edges would be manifested at visible or near-infrared wavelengths.

Oxygen produced in the course of oxygenic photosynthesis will rapidly react with reduced gases like hydrogen and methane that are abundant in brown dwarf atmospheres, owing to which its detection is rendered challenging. Ammonia represents a potential product of hydrogenic photosynthesis, but it is difficult to distinguish between abiotic and biological NH_3 in isolation. However, a potential means of bypassing this conundrum is to look for gases that are depleted (i.e., underabundant) because they serve as reactants in microbial metabolisms. For example, by inspecting (5.44), it is apparent that methane is consumed during hydrogenic photosynthesis. Hence, if a clear mismatch between theory (involving only abiotic mechanisms) and observations of atmospheric gases is confirmed, it might point toward anomalies instigated through biotic activity. This discrepancy is analogous to the hypothesis that phosphine seemingly identified in the Venusian atmosphere could not have arisen from canonical abiotic processes, and is therefore possibly biogenic in origin (Greaves et al. 2020).

It must, however, be appreciated that even apparently anomalous signals may have abiotic roots. For instance, compounds such as sulfur dioxide, iron chloride, and salts were historically invoked in tandem to explain the distinct transition in Venusian albedo observed at ~ 500 nm (Zasova et al. 1981). Minerals like cinnabar (HgS) also give rise to spectral edges, albeit not at the same wavelengths as photosynthetic organisms on Earth (Seager et al. 2005). Moreover, from a pragmatic observational standpoint, the presence of thick clouds will obscure biosignatures even if they do exist. This obstacle could be surmounted, in principle, on those worlds with patchy cloud cover by extracting time-resolved spectra.

Finally, it is helpful to gain an understanding of how many nearby brown dwarfs are situated around stars. Out to a distance of $d_\star \sim 100$ pc from our planet, it is known that $\sim 5 \times 10^5$ stars exist, of which ~ 50 percent are M-dwarfs with $M_\star \lesssim 0.2 M_\odot$. If we are interested in brown dwarfs within an orbital radius of ~ 1 to 30 AU (with a geometric mean of ~ 5.5 AU), it is reasonable to assume that the fraction of stars hosting brown dwarfs within this orbital range is ~ 1 percent (Dieterich et al. 2012). As we are primarily interested in cool brown dwarfs, we consider objects with masses $< 20 M_J$. For this mass range, the yield is further diminished by a factor of ~ 3, based on the latest surveys (Kirkpatrick et al. 2019). Thus, by synthesizing these factors together, we find that $\sim 10^3$ cool brown dwarfs might be appropriate targets for biosignature searches by upcoming telescopes. An observation time of a few hours per target may suffice to permit the

detection of biosignatures (should they exist) with the James Webb Space Telescope (Lingam & Loeb 2019a).

5.8 CONCLUSION

Stretched beneath the heavens blue
Carpet-like magnificent,
In the sun the snow is sparkling;
Dark alone is the wood transparent,
And thro' the hoar gleams green the fir,
And under the ice the rivulet sparkles.

—Alexander Pushkin, *Winter Morning*

Are planets akin to the Earth, with its dense and complex biosphere, highly uncommon in the Universe? This question has been, and will probably be, the subject of countless debates in the past, present, and future. Adherents to the *Rare Earth* hypothesis (P. Ward & Brownlee 2000) tend to invoke the numerous features that are unique to the Earth in our Solar system, ranging from the presence of a large moon to plate tectonics. While this question can only be resolved by observations, theoretical models help us in sharpening our understanding of the general principles underlying habitability and in identifying suitable target stars and planets for empirical studies. Hence, we sought to examine how various planetary factors influence habitability. Needless to say, our analysis is subject to certain limitations, which were delineated at the beginning of this chapter.

The parameters we investigated are classifiable into two broad categories. The first comprises phenomena that do not appear to be strictly necessary for habitability, such as the presence of plate tectonics and the absence of tidal locking. Moreover, including or excluding them may increase or decrease the prospects for life only by a factor of order unity. In contrast, the second category is possibly more fundamental since these factors might raise or lower the likelihood of complex biospheres by orders of magnitude. One of the foremost among them is the fraction of landmasses and oceans on the surface. If the world has no liquid water on the surface, it becomes uninhabitable for life-as-we-know-it. In the other limit, where there are only oceans, the availability of bioessential elements such as phosphorus could be severely suppressed, thus resulting in oligotrophic (nutrient-poor) biospheres with low levels of atmospheric O_2. The average surface temperature is another physical parameter that can substantially

influence the distribution of complex life because molecular, physiological, and ecological phenomena are presumably sensitive to the temperature.

Collectively, our analysis yields a few qualitative conclusions that may be falsifiable over the next few decades. First, not all characteristics associated with our planet are equally important for habitability. For instance, worlds with stagnant-lid tectonics are theoretically capable of sustaining clement temperatures over Gyr timescales. Second, some of Earth's properties might be uncommon and optimized to favor the emergence and sustenance of complex life, although this issue is complicated by observation-selection effects. A notable example is the present-day fraction of Earth's surface covered by landmasses being comparable to that spanned by oceans. Third, our analysis does not preclude, and in fact favors, the existence of some planets that have a higher propensity toward habitability relative to the Earth; such worlds were dubbed "superhabitable" by Heller and Armstrong (2014). A potential candidate for a superhabitable planet is one that retains all the basic features of the Earth, except for slightly larger size and elevated atmospheric O_2 levels. Hence, contra the German polymath Gottfried Wilhelm Leibniz,[5] our planet is not necessarily "the best of all possible worlds."

Obviously, the above conclusions should not be construed as definitive, because they are inevitably subject to a certain degree of anthropocentrism. The ultimate validation or falsification of theoretical models and hypotheses will necessitate a thorough and systematic search for biosignatures and technosignatures, which form the subject of Chapters 6 and 9, respectively. It is only by approaching the data thus garnered from an unbiased standpoint that we can hope to determine the commonality of microbial life, complex multicellular life, and sapient life in the Universe.

---

5. Although Leibniz is most renowned in the scientific world today for his formulation of differential and integral calculus independently of Newton, he was also a celebrated polymath—an accomplished philosopher, engineer, and diplomat, to name a few.

Chapter 6

THE QUEST FOR BIOSIGNATURES

And science tells me that each twinkling star
That smiles above us is a peopled sphere,
Or central sun, diffusing light afar;
. . . . . .
This earth is one vast mystery to man.
First find the secrets of this planet out,
Then other planets, other systems scan!

—Mikhail Lomonosov, *Evening Meditations on Seeing the Aurora Borealis*

One of the most strikingly universal, yet partly underappreciated, characteristics of life on Earth is its remarkable propensity for actively modifying its abiotic environment (Lewontin 2000). Thanks to the epochal work undertaken by Charles Darwin, we are familiar with the oft-misused maxim "survival of the fittest" and the innate role of the environment in driving evolution, but it is quite essential to recognize that the converse is also true. While it remains challenging to definitively pinpoint one individual who originated the notion that biomes actively regulate their environments, this thesis was arguably espoused for the first time in its modern form by Vladimir Vernadsky in his pioneering monograph, *The Biosphere* (1926). A lucid exposition of this theme is presented in Levins and Lewontin (1985, p. 106):

> So the organism influences its own evolution, by being both the object of natural selection and the creator of the conditions of that selection. . . . Darwinism cannot be carried to completion unless the organism is reintegrated with the inner and outer forces, of which it is both the subject and the object.

Through the alteration of their environment, organisms facilitate adaptation and often (albeit not always) enhance their prospects for survival.

This phenomenon of environmental modification by lifeforms, known as niche construction, plays out to varying degrees across multiple scales ranging from the microscopic to the macroscopic. Niche construction has acquired greater prominence in the twenty-first century (Odling-Smee et al. 2003; Sultan 2015; Laland et al. 2016), and this development has been viewed in some quarters as a welcome change after many decades of desuetude, during which period it epitomized a "neglected process in evolution" (cf. Gupta et al. 2017; Futuyma 2017). In light of recent attempts to place niche construction on a credible mechanistic footing, motivated by information-theoretic principles (Constant et al. 2018), it is conceivable that this mechanism may possess a universal basis. Let us contemplate a few examples, as its inherent significance will become obvious shortly hereafter.

It is instructive to commence our analysis at the microscopic level. Microbes known as methanogens, which we encountered previously and will revisit later, are the dominant source of atmospheric methane as a consequence of metabolism. Likewise, the ancestors of modern cyanobacteria pumped out molecular oxygen as a by-product of oxygenic photosynthesis. As we saw in Chapter 3, the rise of molecular oxygen on Earth transformed the biosphere in myriad ways. The diversification of minerals, the construction of new biomolecules, the formation of new ecological niches, and perhaps even the emergence of complex life were some of the potential consequences wrought by the rise in atmospheric and oceanic oxygen concentrations after the evolution of oxygenic photosynthesis.

We segue to macroscopic organisms and choose land plants as our next example. Naturally, one of the most striking consequences of their emergence was biomass synthesis at unprecedented levels and the subsequent surge in the diversity and density of food webs. In addition, plants influence the biosphere in subtler, but equally intriguing, ways. The Amazonian rainforest, in a manner of speaking, regulates the amount of rainfall that it receives through multiple avenues. Fungi and trees in the rainforest have been documented to release particles rich in potassium salts that subsequently influence the density of cloud condensation nuclei (Pöhlker et al. 2012), with the latter playing a key role in the formation of clouds. In a similar vein, rainforest transpiration (depletion of water from aerial segments of plants) alters the convection patterns in the region and plays a vital role in facilitating the dry-to-wet seasonal transition (J. S. Wright et al. 2017).

Next, we turn our attention to animals. Animals are, of course, famously effective "ecosystem engineers" (Butterfield 2011); they are anticipated to have driven (1) diversification of plankton, (2) reworking of sedimentary deposits (bioturbation), (3) feedback mechanisms in nutrient cycling, (4) increase in body size, and (5) formation of minerals by lifeforms (biomineralization). From an evolutionary standpoint, it is tenable that animals contributed to a potential elevation of the oceanic oxygen levels around 600 Myr ago (Lenton et al. 2014), through the evolution of body size and new modes of food consumption (e.g., benthic filter feeding). The last example we wish to highlight is *Homo sapiens*. Our large-scale engineering of the environment has, among many other things, engendered global warming, ocean acidification, global biodiversity losses, rapid deforestation, and the construction of cities (Steffen et al. 2015).

The reader may wonder why we have brought up the topic of niche construction in a chapter dealing with biosignatures. For the purposes of this chapter, a biosignature is defined as an "entity" (e.g., substance or pattern) whose origin necessitates a biological agent. The connection between biosignatures and niche construction is as follows. If life did not modify its environment in manifestly evident ways (e.g., O_2 buildup indirectly via photosynthesis), it would prove to be very difficult to distinguish between worlds with and without life through remote sensing. Therefore, at the risk of oversimplifying an intricate subject, we might contend that niche construction is an essential prerequisite for generating tangible biosignatures. In this context, note that niche construction serves, at best, as a necessary but not sufficient condition for generating detectable biosignatures. For instance, one can envision "cryptic biospheres" that are otherwise endowed with biological activity and niche construction but do not yield observable biosignatures (Cockell 2014); biospheres in subsurface oceans beneath ice envelopes are quintessential examples of this category (Chapter 7).

An important point worth bearing in mind while proceeding further is that nearly all biosignatures specified herein could originate from abiotic (nonbiological) sources. One of the exemplars in this regard is molecular oxygen, which may be generated by oxygenic photosynthesis but also through UV-mediated photolysis of water or carbon dioxide. Hence, when confronted with potential biosignatures, it is necessary to evaluate their likelihood of being *false positives,* i.e., originating from abiotic processes. In our

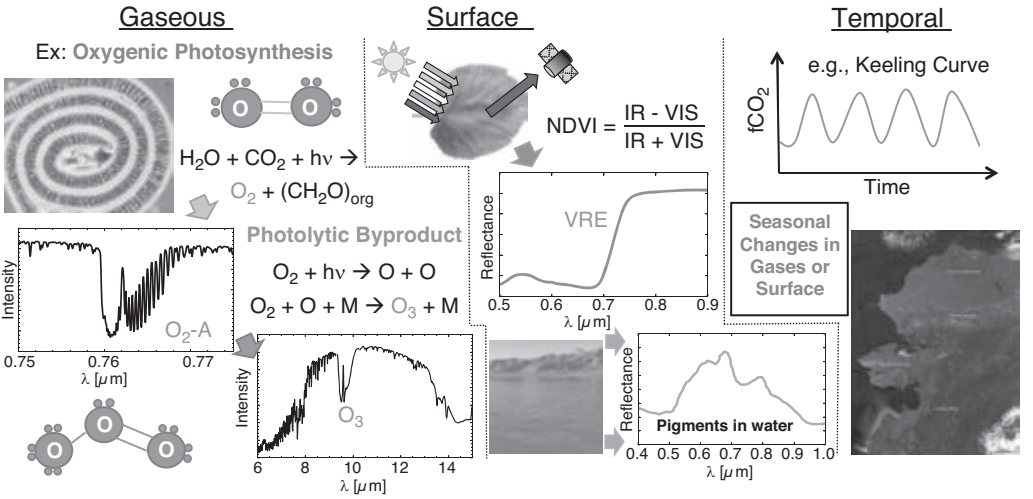


Figure 6.1 Gaseous, surface, and temporal biosignatures. *Left:* Gaseous biosignatures such as O_2 and O_3 arise directly or indirectly by virtue of biological processes (e.g., photosynthesis). *Middle:* Surface biosignatures are spectral features imprinted in reflected light due to interactions with biological organisms and materials; the red edge of vegetation is a quintessential example. *Right:* Seasonal changes in biota translate to temporal variations in observable quantities. For instance, the concentration of CO_2 oscillates because of the growth and decay of vegetation during different seasons. (© The Authors. Published by Mary Ann Liebert, Inc. CC-BY-NC-4.0. *Source:* Edward W. Schwieterman, Nancy Y. Kiang, Mary N. Parenteau, Chester E. Harman, Shiladitya DasSarma, Theresa M. Fisher, Giada N. Arney, Hilairy E. Hartnett, Christopher T. Reinhard, Stephanie L. Olson, Victoria S. Meadows, Charles S. Cockell, Sara I. Walker, John Lee Grenfell, Siddharth Hegde, Sarah Rugheimer, Renyu Hu, and Timothy W. Lyons [2018], Exoplanet biosignatures: A review of remotely detectable signs of life, *Astrobiology* 18[6]: 663–708, fig. 1.)

subsequent analysis, we will closely mirror the approach outlined in Schwieterman et al. (2018) and demarcate biosignatures into three categories: (1) gaseous, (2) surface, and (3) temporal. The basic principles underlying this classification scheme, along with plausible candidates, are depicted in Figure 6.1. An older, but equally comprehensive, overview of biosignatures is expounded in Des Marais et al. (2002); see also Woolf and Angel (1998) for an early observation-centric synopsis.

In this chapter, we will primarily focus on discerning biosignatures by dint of *remote sensing* (i.e., utilizing space- or ground-based telescopes) as opposed to carrying out in situ observations. At least insofar as the near future is concerned, the latter will be feasible for only a handful of targets

within our Solar system, whereas the deployment of telescopes opens up the prospect of studying hundreds of worlds in the next decade. For those who wish to delve further into the myriad realm of on-site techniques, we note that thorough and up-to-date treatments of in situ biosignatures vis-à-vis Solar system exploration are elucidated in Summons et al. (2008), Neveu et al. (2018), and Chan et al. (2019).

Lastly, we do not attempt to cover every single method for detecting exoplanets. The reason is because we are primarily interested in approaches that permit detailed characterizations of planetary atmospheres and surfaces in the near future. An important method precluded in this chapter is astrometry—whereby precise measurements of the "wobble" of stars are used to infer the existence (and basic parameters) of planets—as it does not enable the discovery of temperate Earth-sized planets in the next decade (Perryman et al. 2014); to systematically do so calls for future missions whose current status is far from being confirmed. As of now, astrometry allows for the identification of large planets on long-period orbits.

Another notable approach absent from this chapter is gravitational microlensing, which is particularly useful in constraining the abundances of free-floating planets and finding planets that are situated at distances of $\gtrsim 1$ kpc from the Earth. The basic idea is that light from the "source" star is gravitationally bent under the influence of the "lens" star with its associated planets. The perturbations in the observed light curve due to the planets orbiting the lens star enable their detection. The interested reader is referred to the treatises by Gould and Loeb (1992) and Tsapras (2018) for review. We exclude this method since the planet's existence is inferred only once, and the large distances do not permit follow-up spectroscopy of its atmosphere by present-day telescopes.

6.1 TRANSITING PLANETS

The reader is referred to Winn (2010), Fujii et al. (2018), and Deming et al. (2019) for diligent overviews of the transit method; our approach will closely parallel the former two references. Transits and occultations belong to the category of eclipses, wherein one celestial body obscures another. A transit occurs when the smaller of two objects passes in front of the larger object, whereas an occultation entails the opposite phenomenon. Note that only a small fraction of all planets will be observed to transit, as this technique requires observers to view the planet's orbit almost edgeways. The idea of

detecting planets via the transit method dates back, at the very least, to the mid-nineteenth century, as described in the historical account by Perryman (2018). In this chapter, we will restrict ourselves to planets orbiting single stars, although it must be recognized that a sizable fraction ($\gtrsim 40$ percent) of all stars are multiple systems (e.g., binaries and triples).

6.1.1 Basic physical parameters

The transit probability ($P_{\rm tra}$) can be estimated as follows. The half-angle Θ of the cone, whose vertex is situated at the center of the star, satisfies the condition $\sin\Theta \approx R_\star/a$. Note that $R_\star$ denotes the stellar radius and a is the semimajor axis that is nearly equal to the orbital radius for a circular orbit; here, we have implicitly assumed that the planetary radius R is much smaller than the stellar radius. The transit probability is essentially the solid angle fraction traced out by the cone and is therefore expressible as

$$P_{\rm tra} \approx \frac{2\pi \cdot (2\Theta)}{4\pi} \approx \frac{R_\star}{a} \approx 4.6 \times 10^{-3} \left(\frac{R_\star}{R_\odot}\right) \left(\frac{a}{1\ {\rm AU}}\right)^{-1}, \tag{6.1}$$

where we have made use of the fact that the opening angle is 2Θ along with $\sin\Theta \approx \Theta$ for $\Theta \ll 1$. A more accurate calculation has been shown to yield

$$P_{\rm tra} = \left(\frac{R_\star + R}{a}\right) \left(\frac{1 + e\sin\omega}{1 - e^2}\right), \tag{6.2}$$

where e is the eccentricity of the orbit and ω represents an orbital element known as the argument of periapsis, which will not be discussed here. Henceforth, unless explicitly stated, we will proceed with the assumption that the eccentricity is very small and can therefore be neglected.

One of the chief advantages with the transit method is that it enables the determination of many basic physical parameters of the planet. Two of the foremost among them are the radius and the mass. As the planet transits, a fraction of the starlight will be blocked, along the lines of Figure 6.2. In turn, this causes a reduction in the observed stellar flux, denoted by $\delta_{\rm tra}$. We can estimate $\delta_{\rm tra}$ by noting that it represents the cross-sectional area of the planet in comparison to the star, i.e., we obtain

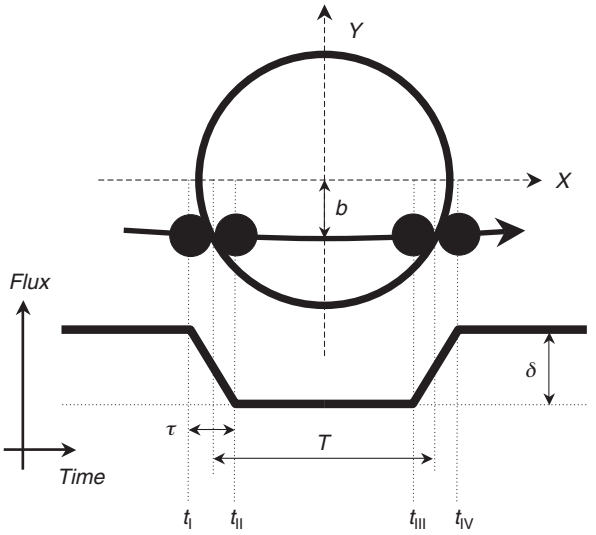


Figure 6.2 A transiting object along with the accompanying light curve. Note that δ refers to the transit depth defined in (6.3). In general, the ingress ($t_{II} - t_I$) and egress ($t_{IV} - t_{III}$) intervals are nearly equal and therefore jointly represented by τ. The total duration of the eclipse T_{total} is defined as $t_{IV} - t_I$, with the full eclipse roughly lasting over the duration $T \equiv T_{total} - \tau$. The transit impact parameter b, in heuristic terms, is the projected distance between the center of the stellar disk and the object's location at midtransit as seen from the image. It could be calculated by utilizing $\left(1 - b^2\right) \tau / T = R/R_\star$ for a transiting exoplanet. (© 2011 The Arizona Board of Regents. Reprinted by permission of the University of Arizona Press. *Source:* Winn, J. N. (2010). Exoplanet transits and occultations, fig. 2. In Seager, S., editor, *Exoplanets,* pages 55–77. The University of Arizona Press, Tucson, AZ.)

$$\delta_{tra} \approx \left(\frac{R}{R_\star}\right)^2 = 8.4 \times 10^{-5} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{R_\star}{R_\odot}\right)^{-2}. \tag{6.3}$$

Hence, by empirically determining δ_{tra}, it becomes possible to solve for R provided that $R_\star$ is independently known, based on the bolometric luminosity and effective temperature of the star. Next, let us turn our attention to the mass M of the planet.

The best-known methodology for determining the mass is known as the radial velocity (RV) method. It appears to have been first contemplated, albeit cursorily, by David Belorizky in 1938, but this method was rejected in favor of transits (Belorizky 1938). The RV method was explicated in 1952 by the distinguished astronomer Otto Struve in what has

now come to be recognized as a seminal publication.[1] Struve's paper posited that Jupiter-sized planets could exist at distances of ~ 0.02 AU from their host stars (Struve 1952). Because such "Hot Jupiters" were unknown in that era and anticipated to be theoretically implausible, this premise led the majority of astronomers to dismiss Struve's method as being impractical, thus consequently delaying the progression of exoplanetary science by decades. The basic physical mechanism underpinning the RV method is that the gravitational tug caused by exoplanets on the host star leads to variations in the latter's radial velocity that can be detected, thereby enabling the determination of M.

We will sketch a heuristic derivation of the RV method; the reader should consult Lovis and Fischer (2010) and J. T. Wright (2018c) for the general case. We begin by supposing that the orbital period $\mathcal{P}_{\rm pl}$ of the planet constitutes a known quantity. For instance, this quantity can be estimated from the detected periodicity in the variations of the host star's radial velocity, since the periods of the star and the planet are equal to one another. For perfectly circular orbits, we have

$$V_\star = \frac{2\pi\, a_\star}{\mathcal{P}_{\rm pl}}, \qquad V_{\rm pl} = \frac{2\pi\, a_{\rm pl}}{\mathcal{P}_{\rm pl}}, \tag{6.4}$$

where $V_\star$ and $V_{\rm pl}$ denote the orbital velocity of the star and planet, whereas $a_\star$ and $a_{\rm pl}$ signify the orbital radii of the star and the planet around the center of mass. Note that the latter duo are expressible as

$$a_\star = \left(\frac{M}{M + M_\star}\right) a, \qquad a_{\rm pl} = \left(\frac{M_\star}{M + M_\star}\right) a. \tag{6.5}$$

Hence, in the limit $M/M_\star \to 0$, we see that $a_\star \to 0$ and $a_{\rm pl} \to a$. Next, we note that the radial velocities measured by the observer constitute the projection of the total velocity along the line of sight. Hence, we introduce the inclination i of the system that represents the angle between the orbital plane and the observer's line of sight. The corresponding radial velocities of the planet and star are given by $V_{r{\rm pl}} = V_{\rm pl} \sin i$ and $V_{r\star} = V_\star \sin i$, respectively.

1. Although Struve was arguably the first to propound a cogent case for the RV method, earlier investigations within this field ought not be consigned to the wilderness.

Another important relation that we make use of is Kepler's Third Law:

$$\frac{a^3}{\mathcal{P}_{\rm pl}^2} = \frac{G\,(M_\star + M)}{4\pi^2}. \tag{6.6}$$

By combining (6.4), (6.5), and (6.6) to solve for the stellar radial velocity $V_{r\star}$, we obtain

$$\frac{M}{(M + M_\star)^{2/3}} = \frac{V_{r\star}}{\sin i}\left(\frac{\mathcal{P}_{\rm pl}}{2\pi\, G}\right)^{1/3}. \tag{6.7}$$

If eccentricity is incorporated, the only difference is that the right-hand side must be multiplied with an extra factor of $\sqrt{1-e^2}$. Hence, the expression for the stellar radial velocity semi-amplitude is given by (Lovis & Fischer 2010):

$$V_{r\star} \approx 9 \times 10^{-2}\,{\rm m/s}\,\frac{1}{\sqrt{1-e^2}}\left(\frac{M\sin i}{M_\oplus}\right)\left(\frac{M + M_\star}{M_\odot}\right)^{-2/3}\left(\frac{\mathcal{P}_{\rm pl}}{1\,{\rm yr}}\right)^{-1/3}. \tag{6.8}$$

Hence, if $V_{r\star}$ and the other variables are known, it is possible to compute $M \sin i$. An important point here is that $M \sin i$ will be obtained via this method (and not M); hence, we can only derive a *lower* bound on the planetary mass. The radial velocity $V_{r\star}$ is conventionally determined through Doppler spectroscopy:

$$\frac{V_{r\star}}{c} \sim \frac{\lambda_B - \lambda_0}{\lambda_0}. \tag{6.9}$$

In this formula, λ_0 and λ_B refer to the photon wavelength in the rest frame of the source and in a suitable reference frame centered on the Solar system. An important point worth mentioning with regard to (6.8) is that the variation in $V_{r\star}$ could also arise from phenomena other than the planetary companion. Some of the prominent sources capable of introducing jitter in the data include magnetic activity of the star and instrumental noise.

The transit timing variation (TTV) method represents an alternative approach that is particularly useful for inferring the existence of additional planets (which need not necessarily transit) and possibly their masses. The basic idea is that the presence of perturbing bodies will induce variations in the interval between two consecutive transits. Such variations are expected to be strongest either when the perturbing planet is in mean-motion resonance (rational ratios of orbital periods) with the transiting planet or when

the transiting planet has a long period (Agol et al. 2005). For the sake of simplicity, we assume that (1) the orbital period of the perturbing planet ($\mathcal{P}_2$) is much higher than the transiting planet ($\mathcal{P}_1$) and (2) the eccentricity of the perturbing planet is much smaller than unity. In this scenario, the mass of the perturbing planet (M_2) was estimated by Holman and Murray (2005) to be

$$M_2 \approx \frac{16}{45\pi} M_\star \frac{\Delta t_{\max}}{\mathcal{P}_1} \left(\frac{\mathcal{P}_2}{\mathcal{P}_1}\right)^3, \tag{6.10}$$

where $\Delta t_{\max}$ signifies the maximum interval between successive transits. The reader is referred to Agol and Fabrycky (2018) for additional details pertaining to this topic as well as to the transit duration variations (TDV) framework that is related to TTV studies and is based on investigating fluctuations in the duration of time taken by the planet to transit.

In the case of transiting exoplanets, we typically have $\sin i \approx 1$ due to the geometric configuration, as a result of which the actual mass M can be computed from (6.7). The orbital inclination is roughly estimated from T and τ, defined in Figure 6.2, as follows:

$$\cos i \approx \frac{\pi}{\delta_{\rm tra}^{1/4}} \left[\frac{T\tau}{\mathcal{P}_{\rm pl}^2}\left(1 - \delta_{\rm tra}^{1/2}\frac{T}{\tau}\right)\right]^{1/2}, \tag{6.11}$$

where $\delta_{\rm tra}$ is the transit depth; we have assumed $\tau \ll T$ and $e \ll 1$ for the sake of simplicity (Winn 2010). Lastly, before moving ahead, we note that the planetary surface gravity, $g \equiv GM/R^2$, is derivable without any dependence on stellar parameters, since it obeys

$$g = \frac{2\pi}{\mathcal{P}_{\rm pl}} \frac{V_{r\star}\sqrt{1-e^2}}{\sin i} \left(\frac{R}{a}\right)^{-2}. \tag{6.12}$$

6.1.2 Transmission spectroscopy

As the planet transits the star, a small fraction of the photons will pass through the upper atmosphere of the planet and will be absorbed. The resultant absorption is wavelength dependent and modulated by the atmospheric composition. At certain wavelengths, the atmosphere is rendered opaque (and transparent at others), thereby increasing the effective silhouette of the planet. Thus, by measuring the out-of-transit and in-transit spectra, the

transmission spectrum of planetary atmospheres can be deduced (Kaltenegger 2017). Theoretical knowledge of the absorption and scattering properties of molecules in conjunction with the observed spectral features could therefore enable us to infer the atmospheric composition and other characteristics of exoplanets.

Before estimating the strength of the signal, it is instructive to define the atmospheric scale height (H_a). For an isothermal atmosphere at temperature T_a comprising chemical species with mean molecular mass $\bar{m}$, the scale height is intuitively understood as the location at which the potential energy of the particle becomes equal to the thermal energy of the atmosphere. In other words, we have

$$H_a = \frac{k_B T_a}{\bar{m} g} = 7.6\,\mathrm{km} \left(\frac{T_a}{250\,\mathrm{K}}\right) \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{M}{M_\oplus}\right)^{-1} \left(\frac{\bar{m}}{28\,m_p}\right)^{-1}, \quad (6.13)$$

where the choice of normalization for $\bar{m}$ is based on an N_2-dominated atmosphere. A more rigorous derivation can be undertaken by solving the differential equation of hydrostatic equilibrium, assuming an ideal gas isothermal equation of state that yields the definition of H_a. The above equation may be further simplified by using the mass-radius relationship $M \propto R^{3.7}$ for rocky planets comparable in size to the Earth. The density of an isothermal atmosphere ρ_a falls off exponentially with the vertical height z via $\rho_a = \rho_{0a} \exp(-z/H_a)$, where ρ_{0a} is the density at ground level. Hence, as per this model, the effective thickness of the atmosphere is a few scale heights.

Let us now suppose that the planet is optically thick (strongly absorbing) when the radius is R. Including the presence of the atmosphere, because of its attendant absorption and scattering features, will cause a change in the transit depth that we denote by S. By defining the effective atmospheric thickness as $N_H H_a$, where N_H is conventionally of order unity, we have

$$S \approx \frac{\pi\,(R + N_H H_a)^2 - \pi R^2}{\pi R_\star^2} \approx \frac{2 N_H H_a R}{R_\star^2}, \quad (6.14)$$

after using $N_H H_a \ll R$. Note that it is more accurate (and appropriate) to envision N_H as quantifying the strength of the spectral features. Ipso facto, higher values of N_H typically arise from increased absorption or scattering,

consequently amplifying the effective atmospheric thickness. After normalizing the above expression by an Earthlike planet orbiting a Sunlike star, we end up with

$$S \approx 8.4 \times 10^{-7} \left(\frac{N_H}{4}\right) \left(\frac{H_a}{8\,\mathrm{km}}\right) \left(\frac{R}{R_\oplus}\right) \left(\frac{R_\star}{R_\odot}\right)^{-2}. \tag{6.15}$$

Although this equation provides us with an estimate for the strength of the signal, the effects of noise must also be considered. In the idealized limit where only photon noise exists, the signal-to-noise ratio (SNR) is approximately given by $S \times \left(N_{ph}/\sqrt{N_{ph}}\right)$, with N_{ph} representing the stellar photon count.[2] The factor of N_{ph} in the numerator is the average number of detected events, whereas the factor of $\sqrt{N_{ph}}$ in the denominator arises from the standard deviation for Poisson statistics. A more accurate analysis requires the calibration of the in-transit spectrum with the out-of-transit spectrum for equal periods of time, thereby dividing the above SNR with a factor of $\sqrt{2}$. Further details concerning the derivation for the SNR presented below can be found in Fujii et al. (2018):

$$\begin{aligned} \mathrm{SNR} \sim 10 &\left(\frac{N_H}{4}\right) \left(\frac{H_a}{8\,\mathrm{km}}\right) \left(\frac{R}{R_\oplus}\right) \left(\frac{R_\star}{0.1\,R_\odot}\right)^{-1} \\ &\times \left(\frac{n_\lambda(\lambda;\, T_\star)}{n_\lambda(3\,\mu\mathrm{m};\, 2500\,\mathrm{K})}\right)^{1/2} \left(\frac{\Delta\lambda}{0.1\,\mu\mathrm{m}}\right)^{1/2} \left(\frac{\Delta t}{30\,\mathrm{hr}}\right)^{1/2} \\ &\times \left(\frac{d_\star}{10\,\mathrm{pc}}\right)^{-1} \left(\frac{D_t}{6.5\,\mathrm{m}}\right) \left(\frac{\xi}{0.4}\right)^{1/2}, \end{aligned} \tag{6.16}$$

where $d_\star$ denotes the distance of the star from the Earth, D_t is the diameter of the telescope, Δt represents the integration time during the transits, and $\Delta\lambda$ is the wavelength resolution of the telescope. The throughput ξ can be thought of as an efficiency factor that quantifies the actual number of photons reaching the telescope's focal surface. Lastly, $n_\lambda(\lambda;\, T)$ is the spectral photon flux (with units of m^{-3}) and is defined as

2. In actuality, other sources of noise include the instrument, molecular absorption in Earth's atmosphere, and background sky brightness.

$$n_\lambda(\lambda;\ T) = \frac{2c}{\lambda^4}\left[\exp\left(\frac{hc}{\lambda k_B T}\right) - 1\right]^{-1}, \tag{6.17}$$

where λ denotes the photon wavelength and T is the effective temperature of either the star ($T_\star$) or the planet ($T_{\rm eq}$), depending on the context. We note that the telescope parameters in (6.16) have been normalized based on the James Webb Space Telescope (JWST). It has a wavelength range of 0.6–28 μm with an associated spectral resolution of $\mathbb{R} \approx 4$–3250.[3] The instruments onboard JWST are capable of characterizing atmospheres via transmission spectroscopy, and its transit depth spectroscopic precision is ~ 10–100 ppm ($1\,\text{ppm} = 10^{-6}$).

If we closely inspect (6.15) and (6.16), a discrepancy in the normalization factors is apparent. In the former, we have normalized $R_\star$ by the solar radius, whereas in the latter, we have normalized it by the typical value for a late (i.e., low-mass) M-dwarf. The basic reason for doing so is to demonstrate that the SNR is likely to be diminished when Sunlike stars are considered. Two counteracting effects are at work in (6.16): (1) the factor $1/R_\star$ favors smaller stars, and (2) the converse is true for the factor $n_\lambda^{1/2}$, although its magnitude also depends on the wavelength. In general, because the scaling with respect to $R_\star$ is stronger, we may expect the SNR to decrease for solar-mass stars. One possible means of bypassing this bottleneck is to increase the integration time, but the dependence on Δt is rather weak. Another route is to increase the telescope aperture, since the SNR scales linearly with D_t; this would necessitate the construction of larger telescopes.

It is worth reiterating that (6.16) was derived under the assumption of photon noise alone. In actuality, there will be other sources contributing to the noise. For instance, both the Hubble Space Telescope (HST) and the Spitzer Space Telescope (SST) exhibit noise levels of tens of ppm, implying that S must be comparable to 100 ppm ($\sim 10^{-4}$) in order for the SNR to attain a reasonable value. From (6.15), it is apparent that this ratio is impossible to achieve for Earth-analogs around Sunlike stars, but it is attainable for Earth-sized planets around late M-dwarfs such as TRAPPIST-1 and Proxima Centauri. Hence, owing to technological limitations, most of

3. The spectral resolution is defined as $\mathbb{R} = \lambda/\delta\lambda$, with $\delta\lambda$ representing the smallest change in wavelength (i.e., increment or decrement) that is resolvable at the wavelength λ.

the upcoming observations in the near future involving transmission spectroscopy will need to be centered on low-mass stars. Finally, we note that our discussion primarily pertained to telescopes with comparatively low spectral resolving power. Future ground-based telescopes with apertures of $\gtrsim 30$ m will possess spectral resolutions of $\mathbb{R} \gtrsim 10^5$, thereby allowing numerous spectral lines to be resolved and compared against template spectra.

As noted earlier, one of the advantages with transmission spectroscopy is that it could enable the identification of atmospheric gases based on their distinctive spectral features arising from absorption (or scattering). Upon inspecting (6.16), we find that the SNR will be maximized (for a given star) at the wavelengths where the stellar flux peaks. The optimal wavelength range depends not only on stellar properties but also on instrument sensitivity and the required spectral resolution. The mid-infrared (mid-IR) range is considered a good choice since many prominent gases in Earth's atmosphere (e.g., O_2 and CO_2) exhibit distinctive molecular features in this regime (Rodler & López-Morales 2014); the same is true for numerous organic species, which may serve as indicators of biological activity (Seager et al. 2016).

The transmission spectra for an Earthlike planet transiting a Sunlike star as well as a late M-dwarf is depicted in Figure 6.3. Some of the potentially discernible molecular features include those located at approximately

- 2.7, 4.3, and 15 μm for CO_2,
- 0.94, 1.13, 1.9, and 6 μm for H_2O,
- 0.69, 0.76, and 1.27 μm for O_2,
- 0.5–0.7, 3.3, 4.7, and 9.6 μm for O_3,
- 2.3, 3.3, and 7.7 μm for CH_4,
- 2.9, 4.5, and 16 μm for N_2O,
- 4.6 μm for CO, and
- 14 μm for CH_3Cl.

Another point worth noting in connection with Figure 6.3 is that the effective atmospheric thickness is enhanced for O_2 in the ultraviolet (UV) regime of 115–200 nm. Due to O_2 absorption, the thickness is ~ 180 km over this range, whereas it drops by about an order of magnitude as we enter the IR regime. Hence, from (6.16), we would be predisposed to think that the SNR will be boosted accordingly, thus implying that the UV is a viable range for future observations. However, note that the stellar photon flux—one

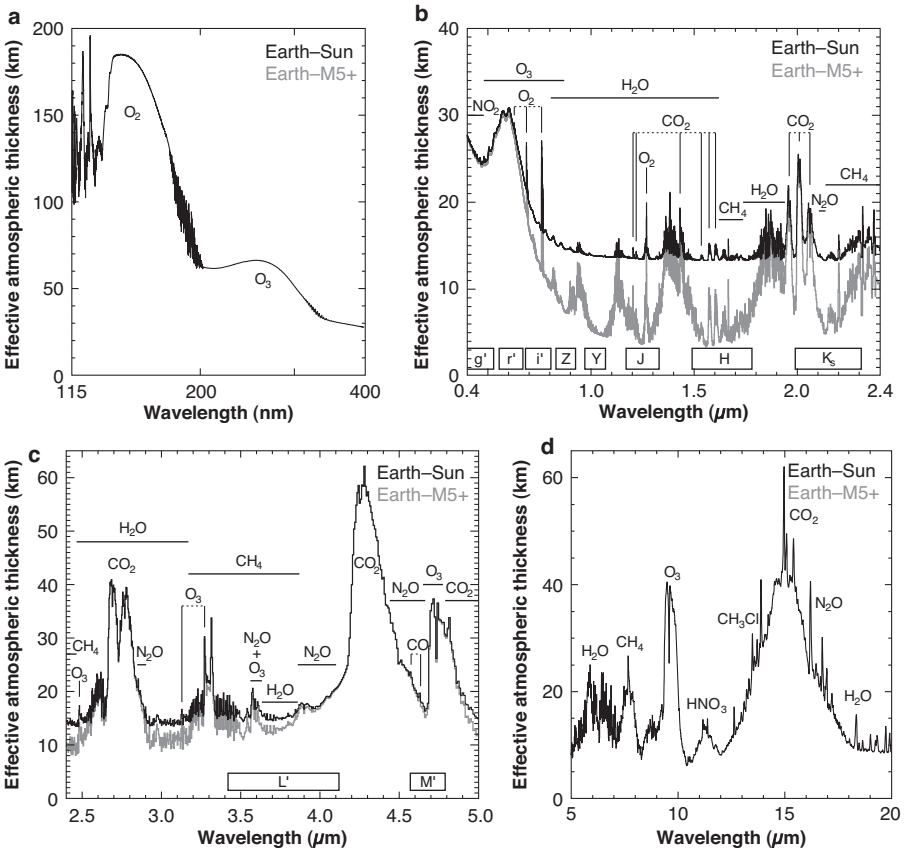


Figure 6.3 The transmission spectra for an Earthlike planet for different wavelength ranges. The black curve corresponds to an Earthlike planet transiting a solar twin, whereas the lighter curve is applicable to an Earthlike planet transiting a late M-dwarf (with $M_\star < 0.15\ M_\odot$). The differences between the two cases originate because refraction governs the depth to which the atmosphere can be probed for Earth-analogs around Sunlike stars, but this restriction does not apply to late M-dwarfs owing to the star-planet size ratios and separations (Kaltenegger 2017). The black boxes at the bottom of the panels represent the spectral coverage conferred by several commonly employed filters. In panel (*a*), depicting the ultraviolet range, note that the *x*-axis is expressed in nm and *not* μm. (© Annual Reviews. *Source:* Lisa Kaltenegger [2017], How to characterize habitable worlds and signs of life, *Annual Review of Astronomy & Astrophysics* 55: 433–485, fig. 10.)

of the factors in (6.16)—becomes smaller in the UV, and this may offset the gain engendered by the increased atmospheric thickness (Bétrémieux & Kaltenegger 2013).

Apart from probing atmospheric composition, transmission spectroscopy could enable the identification of clouds or hazes. These layers may be manifested as either a broad slope in transmission spectra or muted spectral features (Morley et al. 2015). With regard to the latter, if the spectra are rendered featureless, care must be taken to distinguish between the following two scenarios: (1) an atmosphere with thick clouds or hazes and (2) the absence of an atmosphere. It is also possible, in principle, to probe the vertical structure of planetary atmospheres in some instances. This is because of the fact that the atmospheric density gradient promotes refraction of light through the atmosphere (that leaves imprints in the transmission spectrum) and the altitudes to which the transmitted and refracted rays penetrate change with time. Hence, time-resolved transit spectroscopy is a potentially useful tool for determining the vertical structure of the atmosphere. For optically thin atmospheres sans atmospheric refraction, the surface pressure might be determinable, but this is unlikely in most cases since these two conditions are rarely fulfilled.

6.1.3 Occultation spectroscopy

During an occultation, recall that the total flux will decline to a small degree since we transition from illumination provided by the star and the dayside of the planet to that arising from the star alone. The amplitude of the reduction will depend not only on the planetary and stellar sizes but also on the photon fluxes emitted by the star (at temperature $T_\star$) and the planet (at temperature $T_{\rm eq}$). This contrast, denoted by $\mathcal{C}_{\rm IR}$, is therefore given by

$$\mathcal{C}_{\rm IR} \approx \frac{\pi R^2 n_\lambda(\lambda;\ T_{\rm eq})}{\pi R_\star^2 n_\lambda(\lambda;\ T_\star)} \approx \left(\frac{R}{R_\star}\right)^2 \frac{B_\lambda(\lambda;\ T_{\rm eq})}{B_\lambda(\lambda;\ T_\star)}, \tag{6.18}$$

where $B_\lambda\ (\lambda;\ T)$ is the spectral radiance of a blackbody, and the last equality follows from $n_\lambda = B_\lambda/(hc/\lambda)$. In the Rayleigh-Jeans limit where $\lambda \gg hc/(k_B T)$, it is easy to verify that (6.18) simplifies to

$$\mathcal{C}_{\rm IR} \sim \left(\frac{R}{R_\star}\right)^2 \frac{T_{\rm eq}}{T_\star}. \tag{6.19}$$

The normalized version of (6.18) for a late M-dwarf is expressible as

$$\mathcal{C}_{\rm IR} \approx 5.4 \times 10^{-5} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{B_\lambda(\lambda;\ T_{\rm eq})}{B_\lambda(10\,\mu{\rm m};\ 300\,{\rm K})}\right) \times \left(\frac{R_\star}{0.1\,R_\odot}\right)^{-2} \left(\frac{B_\lambda(\lambda;\ T_\star)}{B_\lambda(10\,\mu{\rm m};\ 2500\,{\rm K})}\right)^{-1}. \tag{6.20}$$

The SNR for occultation spectroscopy is given by

$${\rm SNR} \sim \frac{\zeta_a \mathcal{C}_{\rm IR} N_{ph}}{\sqrt{N_{ph}}}, \tag{6.21}$$

with ζ_a representing the relative depth associated with the spectral features arising from atmospheric gases. Note that the factors of N_{ph} and $\sqrt{N_{ph}}$ originate from the assumption of Poisson statistics for the photon noise. After simplification, the final expression for SNR becomes (Fujii et al. 2018):

$${\rm SNR} \sim 1.6\,\zeta_a \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{R_\star}{0.1\,R_\odot}\right)^{-1} \left(\frac{d_\star}{10\,{\rm pc}}\right)^{-1} \left(\frac{D_t}{6.5\,{\rm m}}\right) \times \left(\frac{n_\lambda(\lambda;\ T_\star)}{n_\lambda(10\,\mu{\rm m};\ 2500\,{\rm K})}\right)^{-1/2} \left(\frac{n_\lambda(\lambda;\ T_{\rm eq})}{n_\lambda(10\,\mu{\rm m};\ 300\,{\rm K})}\right) \times \left(\frac{\Delta t}{30\,{\rm hr}}\right)^{1/2} \left(\frac{\Delta\lambda}{0.1\,\mu{\rm m}}\right)^{1/2} \left(\frac{\xi}{0.4}\right)^{1/2}. \tag{6.22}$$

Inspecting the formula reveals that the SNR is quite low even for late M-dwarfs. If we consider solar-type stars, due to the factors $1/R_\star$ and $n_\lambda(\lambda;\ T_\star)^{-1/2}$ in (6.22), we see that the SNR would be lowered even further, probably by an order of magnitude. Hence, at least in the near future, it is unlikely that occultation spectroscopy can be used to probe Earth-sized planets around solar-type stars. For the same reasons, it is not easy to utilize occultation spectroscopy in the visible wavelength range. In this regime, $\mathcal{C}_{\rm IR}$ is predicted to drop below 1 ppm, thereby rendering such observations difficult. We remark, before proceeding further, that alternative equations for (6.16) and (6.22) are elucidated in Kempton et al. (2018).

Occultation spectroscopy is more effective at wavelengths where the thermal emission from the planets is not too low—that is, when $\mathcal{C}_{\mathrm{IR}}$ is sufficiently high. It has been proposed that a wavelength range of ~ 8–$30\ \mu$m is best suited for occultation spectroscopy. The major signatures of molecules in this range include

- 8.9, 9.6, and 14 μm for O_3,
- 8.6 and 17 μm for N_2O,
- 7.7 μm for CH_4, and
- 15 μm for CO_2.

The reader can consult Morley et al. (2017) and Lustig-Yaeger et al. (2019), wherein the predicted thermal emission spectra and/or transmission spectra are presented for the seven planets of the TRAPPIST-1 system (and the temperate M-dwarf exoplanets GJ 1132b and LHS 1140b in the former reference). The resultant thermal emission (from planets) is less likely to experience obstruction by hazes (Burrows 2014)[4] and may enable us to probe the lower atmosphere, provided that sufficient instrumental sensitivity (more so than JWST) to precisely measure thermal emission in the mid- and far-IR is realizable (Batalha et al. 2018). Aside from deducing the atmospheric composition, occultation spectroscopy might also facilitate the determination of the vertical structure, the surface temperature, and rocky materials (e.g., features from Si-O bonds) near the surface, provided that the atmosphere is optically thin at the wavelengths of choice.

6.2 NON-TRANSITING PLANETS

Hitherto, we dealt with methods for identifying the properties of transiting exoplanets. However, as noted earlier, only a small fraction of all planets are observed to transit. The question arises: How do we determine their atmospheric composition and other planetary characteristics? Broadly speaking, there are three different channels that present themselves. We will tackle each of them below, with a particular focus on direct imaging, by adopting the approach and notation presented in Fujii et al. (2018).

4. In contrast, transmission spectra of temperate planets around M-dwarfs—including those in the celebrated TRAPPIST-1 system—could be dominated by clouds, consequently muting their spectral features (Fauchez et al. 2019; Komacek et al. 2020; Suissa et al. 2020).

6.2.1 Direct imaging of exoplanets

Recent overviews of this rapidly growing topic are furnished in Traub and Oppenheimer (2010), Bowler (2016), and Biller and Bonnefoy (2018). As the name indicates, direct imaging relies on detecting photons that are either stellar photons reflected from the planet or those thermally emitted by the planet. In either instance, as one would expect, the stellar photon flux will overwhelmingly dominate the planetary photon flux, thereby making the direct imaging of exoplanets a very challenging endeavor.

There are two methods for suppressing starlight: coronographs and starshades. One of the chief differences between the two is that the former constitutes a part of the telescope payload, whereas the latter is a separate spacecraft; this had led coronographs and starshades to be dubbed internal and external occulters, respectively. A coronograph is a device that blocks out light from the disk of the star, thus permitting the imaging of objects close to the star. The coronograph derives its name from the corona, since it was originally designed by Bernard Lyot in 1930–1931 with the purpose of studying the Sun's corona and prominences. A starshade operates on the same principles, except that it is even easier to visualize given that its location is external to the telescope. In a manner of speaking, the role of starshades is analogous to that of the Moon during a solar eclipse.

Starshades have several advantages, but also some disadvantages, with respect to coronographs. For starters, as we shall see below, the angular resolution achievable by starshades is higher than coronographs; compare (6.24) with (6.25). Second, since coronographs are internal, they present engineering challenges in terms of designing the telescope aperture for achieving contrasts of smaller than $\sim 10^{-9}$; this problem is not so severe for starshades. Lastly, starshades are capable of functioning at larger bandwidths and require fewer optics. Against these advantages, the main challenge with starshades is that they need to be positioned accurately at a large distance with respect to the telescope and the process of reorientation may take ~ 1 to 2 weeks. The latter, in particular, will make it difficult to carry out long-term observations of seasonal changes on exoplanets.

Next, note that the angular separation θ_{ps} between the planet and the star as viewed from Earth is given by

$$\theta_{ps} \approx \frac{a}{d_\star} = 100\,\text{mas}\,\left(\frac{a}{1\,\text{AU}}\right)\left(\frac{d_\star}{10\,\text{pc}}\right)^{-1}, \tag{6.23}$$

where $1\,\mathrm{mas} = 4.85 \times 10^{-9}$ radians. In order for the planet to be resolvable, it is necessary for θ_{ps} to exceed the inner working angle (IWA). The inner working angle is the smallest angular separation that is resolvable by the coronograph or starshade. Not surprisingly, it is closely related to the theoretical diffraction limit of $1.22\lambda/D'$, with D' representing the appropriate diameter of the aperture. In fact, for both coronographs and starshades, the IWA is a few times higher than the diffraction limit, although progress in adaptive optics is driving the IWA toward this bound (Guyon 2018; Ruane et al. 2018; Por & Haffert 2020). For coronographs, the IWA is expressible as

$$\mathrm{IWA} \approx 103\,\mathrm{mas} \left(\frac{\mathcal{R}_c}{2}\right)\left(\frac{\lambda}{0.6\,\mu\mathrm{m}}\right)\left(\frac{D_c}{2.4\,\mathrm{m}}\right)^{-1}, \tag{6.24}$$

with D_c denoting the coronograph diameter, whereas $\mathcal{R}_c$ is dependent on coronograph design and has typical values of $\gtrsim 2$ for the time being. The IWA for starshades is estimated to be

$$\mathrm{IWA} \approx \frac{2F_{st}\lambda}{D_{st}} = 73\,\mathrm{mas} \left(\frac{F_{st}}{10}\right)\left(\frac{\lambda}{0.6\,\mu\mathrm{m}}\right)\left(\frac{D_{st}}{34\,\mathrm{m}}\right)^{-1}, \tag{6.25}$$

wherein D_{st} refers to the diameter of the starshade and F_{st} is known as the Fresnel number. This dimensionless parameter is ~ 10 for starshades that have proven capable of achieving a contrast of 10^{-10} (Harness et al. 2019).

By comparing (6.23) with (6.24) and (6.25), a couple of points stand out. First, the angular separation increases with a, implying that the IWA constraint is more easily satisfied for temperate planets around Solar-type stars relative to those around low-mass M-dwarfs. Second, we see that the IWA increases with the wavelength, implying that it is more advantageous to choose shorter wavelengths. However, one cannot select arbitrarily short wavelengths, since other factors such as the contrast and instrument specifics come into play. In (6.18), we presented the contrast achievable for thermal emission from the planet. In the same spirit, we can calculate the contrast for scattered light ($\mathcal{C}_\mathrm{S}$). The expression takes the form

$$\mathcal{C}_\mathrm{S} = \phi_p A_g \left(\frac{R}{a}\right)^2, \tag{6.26}$$

where A_g signifies the geometric albedo of the planet that is implicitly a function of the wavelength. The factor of $(R/a)^2$ arises because it encapsulates the fraction of the stellar flux intercepted by the planet, and A_g quantifies the fraction of light reflected from it. The extra factor of ϕ_p is called the phase law or phase function and depends on the phase angle α_p of the planet. The phase angle represents the angle between the lines joining the planet and the star and the planet and the observer on Earth. At $\alpha_p = 0$, the star-planet-observer system is aligned, with the planet situated behind the star. The geometric albedo is defined as the ratio of the planet's flux at $\alpha_p = 0$ to that of a Lambertian (i.e., characterized by isotropic scattering) disk at the same distance and same area of cross section as the planet. The phase function for an idealized Lambertian sphere is given by

$$\phi_p\left(\alpha_p\right) = \frac{\sin\alpha_p + \left(\pi - \alpha_p\right)\cos\alpha_p}{\pi}, \tag{6.27}$$

and we refer the reader to Seager (2010) for a more detailed treatment. In the visible and near-IR range, (6.26) can be approximately rewritten as (Fujii et al. 2018)

$$\mathcal{C}_S \sim \frac{2}{3\pi}\left(\frac{R}{a}\right)^2 A_g \sim 10^{-10}\left(\frac{R}{R_\oplus}\right)^2\left(\frac{a}{1\,\mathrm{AU}}\right)^{-2}\left(\frac{A_g}{0.3}\right). \tag{6.28}$$

If we inspect (6.23) and (6.28), a problematic trade-off is manifested. If we decrease a, this increases the contrast $\mathcal{C}_S$ as seen from (6.28). Hence, we may expect that close-in planets are more optimal for direct imaging, since the contrast to be achieved is smaller. On the other hand, decreasing a leads to a decrease in the angular separation θ_{ps} as evident from (6.23). A reduction in θ_{ps} will necessitate an accompanying reduction of the IWA, as otherwise the star-planet system cannot be resolved. However, as we have remarked before, it is difficult to lower the IWA in an arbitrary fashion, given that it calls for technological advances that might be a decade away. With regard to the contrast, one of the highest achieved in recent times is by the Spectro-Polarimetric High-contrast Exoplanet REsearch (SPHERE) instrument on the Very Large Telescope (VLT) in Chile that attained $\sim 3 \times 10^{-7}$ at a separation of 400 mas. Ongoing work on both coronographs and starshades has illustrated that contrast ratios of $\sim 10^{-10}$ are theoretically

feasible and could be implemented in the near future, subject to funding uncertainties.

A number of potentially detectable biosignatures become feasible as a result of directly capturing the exoplanetary photons, as seen in Figure 6.4. Spectroscopic analysis in the visible and near-IR can reveal molecular features of common atmospheric gases. Some of the notable examples in this regard include

- 0.63, 0.69, and 1.27 μm for O_2,
- 0.82, 1.13, and 1.4 μm for H_2O,
- 1.2, 1.6, and 2.0 μm for CO_2,
- 0.78, 0.97, and 1.66 μm for CH_4, and
- 0.5–0.7 μm for O_3.

Note that the spectral resolution required for resolving these features is dependent on their width. For example, to detect H_2O absorption bands around 0.94 μm, $\mathbb{R} \gtrsim 20$ is necessary, whereas detecting the presence of O_2, based on the narrow band at 0.76 μm, at high confidence necessitates $\mathbb{R} \gtrsim 150$ (Brandt & Spiegel 2014). In contrast, lower spectral resolution may suffice to identify certain features such as hazes and clouds.

Not merely atmospheric gases can be detected but also certain distinctive surface features. This becomes feasible when the atmosphere is optically thin, allowing the light scattered off from the surface to be captured by telescopes. As the reflectance of materials on the surface changes with wavelength, measuring variations in the albedo could enable us to discern them via spectroscopy. The reflectance spectra for some common materials are illustrated in Figure 6.4. The most notable among them is the *red edge* of vegetation corresponding to a sharp increase in the reflectance at ~ 0.7 μm (Seager et al. 2005). We will discuss the rationale behind the red edge at a later point in this chapter, but for now it suffices to say that it may constitute a fairly robust signature of photosynthetic organisms. Apart from the red edge, other biological pigments are capable of generating distinctive edges, as seen in Figure 6.4.

Direct imaging opens up the prospects for detecting extraterrestrial oceans through the "glint" effect (Williams & Gaidos 2008; Robinson et al. 2010). To understand this phenomenon, it is important to distinguish between diffuse and specular reflection. In diffuse reflection, the light from an object on the surface is scattered into the solid angle of 2π. In contrast,

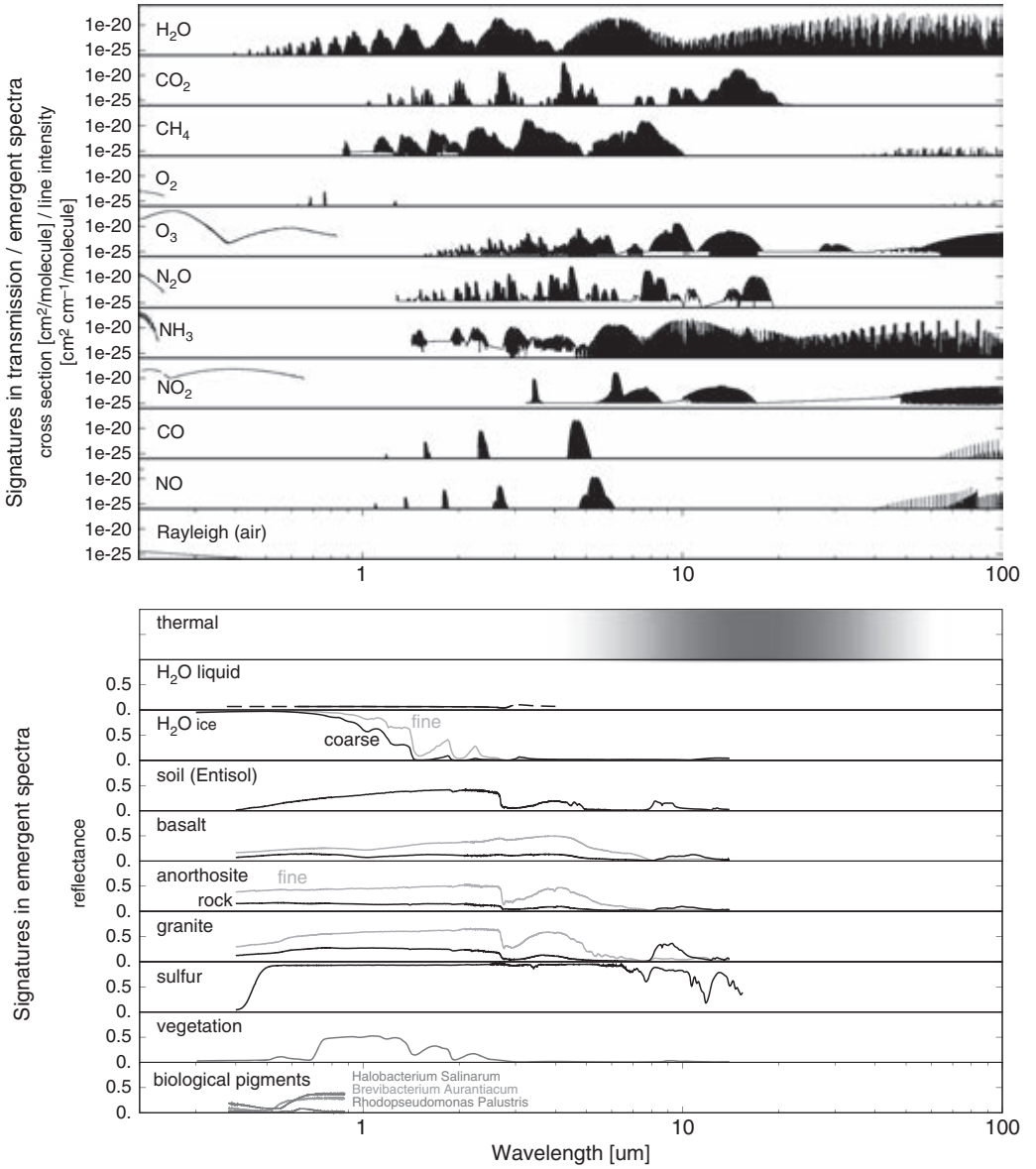


Figure 6.4 *Top:* At shorter wavelengths, the continuous molecular features in terms of absorption cross sections at 300 K. In contrast, at longer wavelengths, the intensities of spectral lines at 296 K assuming a 1 bar atmosphere. *Bottom:* The reflectance spectra of commonly found materials (both biological and abiotic) on the surface of Earth as well as the thermal radiation. (*Source:* Yuka Fujii, Daniel Angerhausen, Russell Deitrick, Shawn Domagai-Goldman, John Lee Grenfell, Yasunori Hori, Stephen R. Kane, Enric Pallé, Heike Rauer, Nicholas Siegler, Karl Stapelfeldt, and Kevin B. Stevenson [2018], Exoplanet biosignatures: Observational prospects, *Astrobiology* 18[6]: 739–778, fig. 2.)

specular reflection is akin to how light is reflected from a flat mirror, in which the angle of incidence and reflection equal one another. In the case of a planet orbiting the star, the angle subtended by light from the star on the planet is $2\Theta \approx 2R_\star/a$, as noted in Section 6.1.1. Thus, as light is reflected at this angle, the corresponding solid angle Ω_{sr} is given by

$$\Omega_{sr} = \int_0^{2\Theta} \sin\theta' \, d\theta' \int_0^2 \pi \, d\phi' = 2\pi \left[1 - \cos\left(\frac{2R_\star}{a}\right)\right]$$
$$\approx 4\pi \left(\frac{R_\star}{a}\right)^2 \approx 2.7 \times 10^{-4}\,\text{sr} \left(\frac{R_\star}{R_\odot}\right)^2 \left(\frac{a}{1\,\text{AU}}\right)^{-2}. \qquad (6.29)$$

Now, suppose that we consider a terrestrial planet comprising 35 percent land, 10 percent ice, and 55 percent water. The typical albedos for land and ice are chosen to be 0.2 and 0.8, respectively. In this specific case, the fractional contribution from the land and ice components to the diffuse scattered radiation (f_{dr}) is

$$f_{dr} = \frac{(0.35)(0.2) + (0.1)(0.8)}{2\pi} \approx 2.4 \times 10^{-2}, \qquad (6.30)$$

where the factor of 2π reflects the solid angle over which the scattering occurs. Next, consider the situation in which the ocean is obliquely illuminated. The resultant specular reflection is from small waves on the ocean surface that have a probability $p_{\text{wav}} \approx 1.3 \times 10^{-5}$, and the albedo of ocean water in this configuration is around 0.6. The fractional contribution from the ocean to specular reflection (f_{sr}) is expressible as

$$f_{sr} = \frac{(0.55)(0.6)p_{\text{wav}}}{\Omega_{sr}} \approx 1.6 \times 10^{-2} \left(\frac{R_\star}{R_\odot}\right)^{-2} \left(\frac{a}{1\,\text{AU}}\right)^2. \qquad (6.31)$$

It is therefore apparent that the contribution from the specular reflection of light by oceans is approximately equal in magnitude to the component from diffuse reflection. Depending on the orbital configuration and the phase angle, the signal from specular reflection could even exceed that originating from diffuse reflection. Subsequent numerical models have revealed that the glint is particularly pronounced during the crescent phase of the planet, i.e., when α_p approaches π. The difficulty with detecting the glint at this configuration is that the star-planet angular separation is small, with the contrast also diminished, as seen from (6.26) and (6.27). When it comes

to false positives for the glint effect, planets with moderate obliquity and substantial ice coverage at higher latitudes may mimic specular reflection (Cowan et al. 2012).

From (6.28), it is evident that measuring the contrast at different wavelengths ought to yield an estimate for the albedo, provided that the planetary radius R is known through some other procedure. Knowledge of the albedo can, in turn, be used to infer the effective temperature of the planet. It is also potentially feasible to place constraints on the surface pressure P_s via the atmospheric column density, because the latter is given by P_s/g. The column density may, in principle, be determined from Rayleigh scattering features as they depend on the molecular cross section (that could be calibrated for known species) and the atmospheric column density. Determining the surface temperature, on the other hand, is much more difficult since its effects on spectra in the wavelength regimes of interest appear to be minimal.

Hitherto, we have discussed only "snapshots" of exoplanets through direct imaging. It is expected that time-resolved direct imaging will provide additional information insofar as the planetary climate and physical parameters are concerned. For example, if the resolvable timescale is smaller than the rotation period, then one will discern periodic variations in the disk-integrated light emanating from a nonuniform planetary surface. By measuring the periodicity, it will therefore be possible to calculate the rotation rate. Knowledge of the rotation rate is particularly important for climate models and habitability. For example, the inner edge of the habitable zone has been shown to depend strongly on the rotation rate, with slowly rotating planets (e.g., Venus) being capable of sustaining Earthlike climates at ~ 2 times the stellar flux compared to planets with rapid rotation rates (Yang et al. 2014).

Other parameters that are deducible from time-resolved direct imaging include the planetary obliquity (axial tilt) and the existence of a heterogeneous surface. With regard to the latter, as seen in Figure 6.4, the overall albedo (at different wavelengths) of the planet will vary depending on the fraction of surface covered by diverse materials (e.g., rock, ice). Hence, by investigating the scattered light spectra at multiple durations and wavelengths, it may be possible to infer whether the planet is geologically active. In the idealized limit where surface features cause isotropic scattering and the only source of noise is photon noise, it might even be possible to compute the fraction of the surface covered in land, oceans, and vegetation (Ford et al. 2001; Fujii et al. 2010). The presence of a time-varying cloud cover is also detectable in theory by measuring the deviations in daily light curves.

In closing, we note that the class of Extremely Large Telescopes (ELTs), with apertures of $\gtrsim 30$ m, expected to start functioning in the mid-to-late 2020s, would be capable of achieving IWA on the order of 10 mas, as seen from (6.24). In turn, this would enable the angular resolution of planets orbiting M-dwarfs, as is evident from (6.23). The contrast ratio of $\sim 10^{-7}$ for such systems is achievable by these ground-based telescopes. Hence, the direct imaging of nearby temperate exoplanets around M-dwarfs will be attainable in the near future. On the other hand, the situation with respect to Earth-sized planets around Sunlike stars is much more challenging, because the contrast ratio of $\sim 10^{-10}$ is very difficult to achieve from the ground, thereby necessitating space observatories equipped with either starshades or coronographs.

6.2.2 Thermal phase curves

As noted earlier, direct imaging of exoplanets is currently a very challenging endeavor. An alternative approach for studying non-transiting exoplanets entails the extraction of planetary spectra from the time-varying component of the star+planetary spectra; the temporal evolution arises as a result of the planet's orbit around the star, thus causing the phase angle α_p to change. The variations in the phase curves originate from scattered light as well as thermal emission from the planet, but it is the latter that provides a better contrast ratio. To see why, note that the contrast ratio for this method is functionally identical to occultation spectroscopy and is therefore given by (6.20). Next, from inspecting (6.20) and (6.28), after normalizing the latter by 0.1 AU instead of 1 AU, we see that the former (thermal emission contrast) is higher than the latter (scattered light contrast) by ~ 2 orders of magnitude. The corresponding SNR is expressible as

$$\begin{aligned}\mathrm{SNR} \sim 1.6\,\zeta_{pc} &\left(\frac{R}{R_\oplus}\right)^2 \left(\frac{R_\star}{0.1\,R_\odot}\right)^{-1} \left(\frac{d_\star}{10\,\mathrm{pc}}\right)^{-1} \left(\frac{D_t}{6.5\,\mathrm{m}}\right) \\ &\times \left(\frac{n_\lambda(\lambda;\,T_\star)}{n_\lambda(10\,\mu\mathrm{m};\,2500\,\mathrm{K})}\right)^{-1/2} \left(\frac{n_\lambda(\lambda;\,T_{\mathrm{eq}})}{n_\lambda(10\,\mu\mathrm{m};\,300\,\mathrm{K})}\right) \\ &\times \left(\frac{\Delta t}{30\,\mathrm{hr}}\right)^{1/2} \left(\frac{\Delta\lambda}{0.1\,\mu\mathrm{m}}\right)^{1/2} \left(\frac{\xi}{0.4}\right)^{1/2}, \end{aligned} \tag{6.32}$$

where ζ_{pc} denotes the relative amplitude of variations in the light curve stemming from the changing orbital phase.

One of the chief advantages associated with this method is that it enables the detection of an atmosphere and the accompanying redistribution of heat. This feature is particularly important for tidally locked exoplanets, since the atmosphere can facilitate the transport of heat to the permanent nightside. We explored this aspect in Section 5.3; in Figure 5.4, it is apparent that a sufficiently thick atmosphere ensures the nightside temperature may allow for habitable conditions, whereas the absence of an atmosphere would result in a very cold (and uninhabitable) nightside. Bearing these facts in mind, let us consider how the variations in thermal emission with orbital phase would permit us to discern the presence of an atmosphere.

As the planet orbits the host star, the fraction of its dayside visible to the observer will change; the same holds true for its nightside. In the limit where no atmosphere exists, the temperature gradient between the dayside and nightside is particularly pronounced since the latter is very cold (see Figure 5.4). Hence, in turn, we expect that the amplitude of variations in the observed planetary thermal emission would be correspondingly higher. Next, consider a planet with a thick atmosphere that permits the efficient redistribution of heat. In this event, the temperature gradient between the dayside and nightside is diminished and translates to smaller fluctuations in the thermal emission. This method has been explored by several authors, with Gaidos and Williams (2004) being one of the earlier publications, while Koll et al. (2019) represents a recent example.

In order to explicate this method further, we shall focus on Proxima b, the temperate planet orbiting Proxima Centauri, as it was studied by Kreidberg and Loeb (2016). Figure 6.5 shows the thermal phase curves that depict the contrast between the planetary and stellar fluxes over the orbital period of Proxima b. The first point worth noting is that the contrast increases as one moves from 5 μm to 10 μm; this behavior is along expected lines since the thermal emission of planets will peak at such wavelengths. Second, as explained in the previous paragraph, the thermal phase curves are muted in the presence of an atmosphere that redistributes 35 percent of the incident heat to the nightside, compared to the scenario with only bare rock. Kreidberg and Loeb contend that the JWST is capable of distinguishing between these two cases at 4σ confidence via phase curve measurements. In addition, ozone exhibits an absorption feature centered at 9.8 μm that may be detectable in the thermal emission spectrum by the JWST, provided

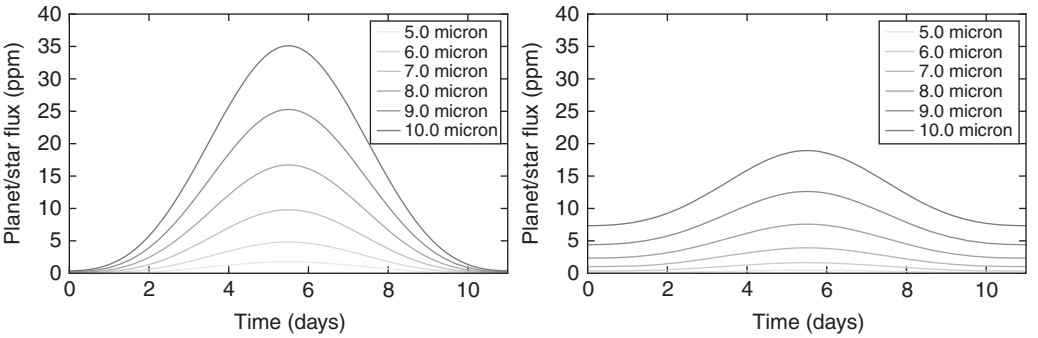


Figure 6.5 The thermal phase curves, expressed as the ratio of the planetary and stellar fluxes in ppm, are depicted for a single orbit of Proxima b (around eleven days). The left panel corresponds to bare rock (sans atmosphere), while the right panel assumes that 35% of the incident stellar energy is redistributed to the nightside via an atmosphere. The thermal phase curves are plotted for different wavelengths. (© AAS. Reproduced with permission. *Source:* Laura Kreidberg and Abraham Loeb [2016], Prospects for Characterizing the Atmosphere of Proxima Centauri b, *Astrophysical Journal Letters* 832[1]: L:12, fig. 2.)

that the SNR is sufficiently high; a potential difficulty, however, is that the amplitude of the signal is small ($\lesssim 1$ ppm).

Clouds are another feature that may be manifested in thermal phase curves. Climate models indicate that tidally locked planets with oceans are likely to engender the formation of thick clouds at the substellar point (zenith), thus leading to characteristic patterns in the orbital phase curves. Hence, identifying such patterns might point to the presence of not only clouds but also underlying oceans. Lastly, spectrally resolved phase curves could enable the identification of atmospheric molecules (as noted above for ozone) due to their absorption bands. It is important to recognize that thick atmospheres reduce the likelihood of probing near-surface atmospheric layers and therefore diminish the prospects of detecting molecular signatures.

In closing, we remark that thermal phase curves have been gainfully employed to study the hot rocky exoplanet LHS 3844b, with a radius of $\sim 1.3 R_\oplus$ and a dayside temperature of ~ 1040 K (Kreidberg et al. 2019). The observed phase variation, which was symmetric and distinguished by its large amplitude of order 100 ppm, led the authors to conclude that this planet either has a tenuous atmosphere or is virtually airless, with the latter being favored by models of atmospheric escape.

6.2.3 High-contrast imaging and high-resolution spectroscopy

One of the chief difficulties associated with the remote sensing of biosignatures sensu lato stems from having to discern the planetary component amidst the combined spectra derived from the host star and planet; this is a difficult endeavor because the contribution from the latter is much smaller than the former. The last major method that we shall cover entails combining high-resolution spectra with high-contrast imaging (Sparks & Ford 2002). Recent summaries of this promising approach have been expounded in Birkby (2018) and Mazin et al. (2019).

To appreciate the significance of high-resolution spectroscopy (with $\mathbb{R} \gtrsim 10^5$), note that molecular bands will be revealed to comprise tens to hundreds of individual lines. The conundrum, then, is how the stellar and planetary lines are distinguishable from one another. Let us turn our attention to (6.4) and (6.5). It is clear from these formulae that the orbital velocity of the planet is much higher than that of the star. Hence, to leading order, the star may be considered stationary, whereas the spectral lines associated with scattered light from the planet will be Doppler-shifted due to the radial component of its velocity. For Hot Jupiters, the change in wavelength could reach several times the spectral resolution per hour. Hence, by searching for the Doppler-shifted individual lines, the planetary component can be duly isolated. The identification of specific molecular species is undertaken by cross-correlating the observed spectra with templates generated from models. When employed in this manner, high-resolution spectroscopy has been shown to potentially achieve planet-to-star contrast on the order of 10^{-5}.

Next, let us consider again high-contrast imaging (Section 6.2.1). The use of a coronograph or a starshade is capable of blocking the light from the star by several orders of magnitude; even with current technology, a contrast of $\sim 10^{-7}$ is attainable for ground-based telescopes. Let us err on the side of caution and consider coronographs that achieve a contrast of $\sim 10^{-5}$. Therefore, based on the preceding discussion, it is apparent that the overall planet-to-star contrast realizable is $\lesssim 10^{-5} \times 10^{-5} = 10^{-10}$. This number is important since it corresponds to the contrast that must be achieved for characterizing planets around Sunlike stars, as evidenced by (6.28). Earlier, we mentioned that this contrast ratio is beyond the scope of current ground-based telescopes reliant solely on high-contrast imaging. In principle, however, this bottleneck could be surmounted by ground-based

telescopes that combine high-resolution spectroscopy with high-contrast imaging.[5] Snellen et al. (2015) predicted that putative super-Earths (with $R = 1.5\,R_{\oplus}$) around Alpha Centauri A at an effective temperature of $T_{\rm eq} = 300$ K might be detectable at an SNR of 5 from one night of observations using a telescope with ~ 40 m aperture in the wavelength range of 4.82–4.89 μm.

Apart from the determination of molecular species in the atmosphere, this approach also enables the calculation of the planet's line-of-sight velocity. By combining this information with the orbital velocity, the inclination of the planet can be determined, thus permitting the estimation of its actual mass via the radial velocity method outlined in Section 6.1.1. In the regime where thermal emission from the planet dominates over the scattered light (mid-IR), the vertical structure of the atmosphere may be obtainable under ideal conditions.

6.3 ALTERNATIVE OBSERVATIONAL CONSTRAINTS ON HABITABILITY

Hitherto, we have primarily concerned ourselves with signatures of molecules in the atmospheres of exoplanets, although we also elucidated how surface features such as oceans and land are potentially detectable. These characteristics are either directly indicative of life (e.g., vegetation) or indirectly increase the prospects for habitability (e.g., oceans). We will explore the latter category further by studying how other factors that influence habitability—including the properties of the host star—could be constrained through observations.

6.3.1 Planetary magnetic field

The presence of a strong (usually dipolar) planetary magnetic field is conventionally perceived as one of the requirements for habitability. It has been argued to mitigate the flux of cosmic rays reaching the surface and the escape of atmospheric particles due to interaction with stellar winds; for more details, consult Section 4.2. Yet, on the other hand, emerging

5. The same principle is applicable to future space-based telescopes, as this technique can relax the light suppression requirements by a few orders of magnitude (J. Wang et al. 2017).

theoretical models suggest that the significance of magnetic fields vis-à-vis planetary habitability might have been overstated (Lingam 2019). Nonetheless, detecting or ruling out exoplanetary magnetic fields is advantageous from at least two different standpoints.

First, despite the above concerns, the presence of magnetic fields could boost the chances for complex biospheres. Second, the sustenance of long-term planetary dynamos is intimately linked to plate tectonics (another potentially crucial ingredient of habitability), as noted in Section 5.2. For Mars and Venus, it is conceivable that the growth of the inner core—an ostensibly necessary condition for a robust dynamo—was truncated owing to the cessation of plate tectonics (Breuer et al. 2010). Hence, discovering a strong magnetic field might imply that the planet experiences tectonic activity. A number of avenues have been proposed for determining the existence and strength of exoplanetary magnetic fields. The proposed methodologies include the likes of chromospheric emission, early transit ingress, H_3^+ emission, and Lyα absorption profiles. A recent review of this subject can be found in Grießmeier (2015); we will restrict ourselves to a couple of closely related examples below.

It has been more than forty years since the idea of searching for magnetized exoplanets by looking for intense, low-frequency radio emission from their auroral regions was proposed, as summarized in Zarka (2007). The planetary magnetic field lines originating in these regions merge with interplanetary magnetic field lines and are therefore regarded as being "open." The reason behind this auroral emission is a plasma instability known as the electron-cyclotron maser instability (ECMI). A mathematical derivation of this instability is beyond the scope of the book (see Treumann 2006), but it suffices to say that the ECMI arises when the following resonance condition is satisfied between an electromagnetic wave and energetic electrons:

$$\omega - k_{\parallel} v_{\parallel} = \ell \frac{\omega_{ce}}{\gamma}, \tag{6.33}$$

where ω and $k_{\parallel}$ denote the angular frequency and parallel wave number of the electromagnetic wave; the parallel direction is determined by the ambient magnetic field. Note that $v_{\parallel}$ is the parallel velocity component of the energetic electrons, while $\omega_{ce} = eB/m_e$ represents the cyclotron angular frequency; here, e and m_e are the charge and mass of an electron, whereas B denotes the magnetic field strength. Finally, ℓ signifies the cyclotron

harmonic number, and γ is the Lorentz factor for electrons moving at velocity v.

The ECMI has been likened to a laser or maser in some respects. Note, for instance, that the ratio of the rates of stimulated emission to spontaneous emission for atoms is determined via the Einstein coefficients. This ratio (Δ_{er}) is consequently expressible as

$$\Delta_{er} = \left[\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1\right]^{-1}, \tag{6.34}$$

where ω is the angular frequency of the photon and T is the ambient temperature. If all things are held equal, we see that this ratio increases as the angular frequency is lowered or, equivalently, at longer wavelengths. In the 1950s, this fact was invoked to explain why the emission is dominated by radio wavelengths. In actuality, the specifics behind auroral radio emission involve the complex interplay of energetic particles in the stellar wind interacting with the constituents of the planet's upper atmosphere; see Treumann (2006) for a comprehensive overview of this topic.

The essential point to take away from this discussion is that radio waves are emitted at a frequency $f_{\max}$ close to the electron gyrofrequency given by

$$f_{\max} \approx \frac{eB_s}{2\pi m_e} = 2.8\,\text{MHz}\left(\frac{B_s}{10^{-4}\,\text{T}}\right), \tag{6.35}$$

with B_s denoting the *average* magnetic field strength at the surface of the planet. A more accurate model requires us to replace B_s with the magnetic field strength at higher latitudes (the so-called polar caps) where most of the radio emission occurs; however, the result changes by less than a factor of 2. At this stage, an important limitation must be highlighted that we shall return to later. At frequencies $\lesssim 10$ MHz, the Earth's ionosphere blocks the radio waves from penetrating to the surface. Hence, this method is more effective when it comes to planets with magnetic field intensities that are several times higher than the Earth's, preferably by an order of magnitude; an alternative route is to rely on space-based observatories, where the interference from the ionosphere is absent.

Therefore, by guessing the value of B_s and searching at the corresponding wavelengths, it may be feasible to detect auroral radio emission and thereby estimate the exoplanetary magnetic field strength. From a

practical standpoint, however, we are confronted by another potential issue: Is the flux of radio emission received at Earth sufficiently high so as to be detectable? In order to answer this question, we must calculate the radio flux density $\Phi_{\rm rad}$, defined as

$$\Phi_{\rm rad} = \frac{P_{\rm radio}}{4\pi\,\Delta f d_\star^2}, \tag{6.36}$$

where $\Delta f \approx f_{\rm max}/2$ and $P_{\rm radio}$ signifies the power associated with the planetary radio emission. Hence, from inspecting the above formula, it is clear that both $P_{\rm radio}$ and $f_{\rm max}$ are unknown. With regard to the latter, several dynamo scaling laws may be utilized to potentially predict the magnetic field strength; the issue, however, is that the dependence on basic physical parameters (e.g., planetary rotation rate) varies from model to model.

Estimating the radio power $P_{\rm radio}$ is also difficult since it depends on the nature of the interaction between the stellar wind and the planet. The wind and the planet can be either magnetized or unmagnetized, thus leading to four permutations in total. Of these, the interaction between an unmagnetized wind and an unmagnetized planet does not yield radio emission, but the other three cases remain relevant. Broadly speaking, most of these models posit that the radio emission from the exoplanet is proportional to the power carried by the stellar wind, as the latter constitutes the source of energetic charged particles. However, it is important to appreciate that the stellar wind has two major components—the kinetic and the electromagnetic (thermal energy is set aside)—for which reason, there will be two components of the power transported by the wind.

The kinetic power ($P_{\rm kin}$) carried by the stellar wind is given by

$$P_{\rm kin} \approx \left(\rho_{sw} v_{sw}^2\right)\left(\pi R_{mp}^2 v_{sw}\right) = \pi \rho_{sw} v_{sw}^3 R_{mp}^2, \tag{6.37}$$

where ρ_{sw} and v_{sw} signify the density and velocity of the stellar wind, respectively, while R_{mp} is the magnetopause standoff distance delineated in Section 4.2.1. $P_{\rm kin}$ can be understood as the product of the dynamic pressure (i.e., energy density) and the volumetric flow rate that is defined as the product of the cross-sectional area πR_{mp}^2 and the velocity v_{sw}. In contrast, the magnetic power ($P_{\rm mag}$) of the stellar wind is expressible as

$$P_{\rm mag} = \int \left(\frac{\mathbf{E}_{sw} \times \mathbf{B}_{sw}}{\mu_0}\right) \cdot d\mathbf{S} \approx \frac{\pi B_\perp^2 v_{sw} R_{mp}^2}{\mu_0}, \tag{6.38}$$

wherein $\mathbf{E}_{sw}$ is the electric field of the stellar wind, and $B_{\perp}$ is the projection of the interplanetary magnetic field in the direction perpendicular to the velocity $\mathbf{v}_{sw}$ of the stellar wind.[6] The above formula is obtained by integrating the Poynting flux $\mathbf{E}_{sw} \times \mathbf{B}_{sw}/\mu_0$ over the cross-sectional area of the magnetopause distance. The electric field $\mathbf{E}_{sw}$ is determined via Ohm's Law for ideal magnetohydrodynamics: $\mathbf{E}_{sw} + \mathbf{v}_{sw} \times \mathbf{B}_{sw} \approx 0$; see (4.30) for more details. Note that (6.38) must be multiplied by a factor of order 0.1 to account for the efficiency of magnetic reconnection, but this is absorbed into the parameter η_{mag} introduced below.

The radio power (P_{radio}) in (6.36) consequently has two components (kinetic and magnetic), respectively given by

$$P_{\text{radio}} = \eta_{\text{kin}} P_{\text{kin}} \tag{6.39}$$

and

$$P_{\text{radio}} = \eta_{\text{mag}} P_{\text{mag}}, \tag{6.40}$$

wherein $\eta_{\text{kin}} = 10^{-5}$ and $\eta_{\text{mag}} = 2 \times 10^{-3}$ (Zarka 2007); these coefficients were computed for our Solar system and are not guaranteed to be valid for other planetary systems. After employing these two formulae in conjunction with (6.36), we arrive at

$$\frac{\Phi_{\text{rad,kin}}}{\Phi_{\text{rad,mag}}} \approx 3.2 \times 10^{-5} \, \frac{\rho_{sw} v_{sw}^2/2}{B_{\perp}^2/(2\mu_0)}, \tag{6.41}$$

with the relevant details provided in Vidotto and Donati (2017). Hence, in most instances, $\Phi_{\text{rad,mag}}$ will become dominant since the kinetic energy density of the stellar wind is not usually four to five orders of magnitude higher than its (perpendicular) magnetic energy density. For Proxima b, theoretical calculations indicate that $\Phi_{\text{rad}} \gtrsim 100$ mJy is potentially feasible (Burkhart & Loeb 2017); note that 1 Jy is 10^{-26} W m^{-2} Hz^{-1}. In contrast, numerical simulations and analytical modeling suggests that the radio emission from the TRAPPIST-1 planets may be on the order of 0.1 mJy (Dong, Jin, et al. 2018). The difference between the two cases stems partly from the fact that $\Phi_{\text{rad}} \propto d_{\star}^{-2}$ and the distance to TRAPPIST-1 is about ten

6. Note that $B_{\perp} \leq B_{sw}$, wherein B_{sw} is the magnetic field strength inherent in the stellar wind at the location of the planet (interplanetary magnetic field).

times higher compared to Proxima Centauri. For both systems, the radio emission might peak in the range ~ 0.01–1 MHz.

The last point that we wish to bring up is that our discussion pertains to the steady-state stellar wind. At this stage, it is worth recalling the basic aspects of coronal mass ejections (CMEs) introduced in Section 4.4.2. CMEs are particularly important for M-dwarfs as these stars tend to flare more often and are therefore more likely to produce CMEs. During a sufficiently large CME, such as the 1859 Carrington event, the kinetic energy density can be amplified by a factor of $\sim 10^3$ (Dong, Jin, et al. 2018), thus implying that $\Phi_{\mathrm{rad,kin}}$ will also increase by the same degree. Likewise, $\Phi_{\mathrm{rad,mag}}$ could undergo amplification by an order of magnitude or more. Hence, for active stars, we may expect to see spikes in the radio flux density corresponding to the impact of space weather events.

Earlier, we mentioned that one of the potential caveats with this method is that the detection of weak planetary magnetic fields is difficult, since the emitted radio waves are blocked by the ionosphere. Hence, an alternative approach entails searching for auroral emission from exoplanets in the visible range. One of the best-known examples in this category is the oxygen line at 557.7 nm (O I) that represents the brightest auroral feature and is responsible for the compelling green glow observed during the Aurora Borealis and Aurora Australis. The emitted power at the O I line (P_{OI}) is estimated to be

$$P_{\mathrm{OI}} \sim 2 \times 10^{-4}\, P_{\mathrm{kin}}, \tag{6.42}$$

with P_{kin} given by (6.37). The advantage with this method, apart from the wavelength, is that most of the emission occurs in a narrow band centered on the O I line. As a result, it can be shown that the planet-star contrast ratio at this band is potentially on the order of 10^{-7} for planets around late M-dwarfs after taking typical stellar wind parameters into account. Moreover, when the planet is impacted by large CMEs, the steady-state contrast ratio will be boosted by two to three orders of magnitude. Hence, coronographs or starshades that achieve the desired contrast with minimal instrumental noise may be capable of detecting O I auroral emission and thereby constraining the magnetic field strength of the corresponding exoplanets. This approach was explicitly applied to Proxima b by Luger et al. (2017), who concluded that (1) $P_{\mathrm{OI}} \sim 10^{11}$ W, and (2) future telescopes could observe O I auroral emission from Proxima b after a few nights of observation time.

6.3.2 Planetary system architecture and composition

Although we have dealt hitherto with approximately Earth-sized rocky planets in the temperate zone, observations will also reveal the presence of gas giants akin to Jupiter and the orbital architecture of planetary systems. Giant planets could be comparatively easier to detect as the transit depth, contrast ratio, and other relevant parameters depend on the size of the planet.

A knowledge of the orbital architecture can help us gain additional information concerning the habitability of temperate rocky planets, if any do exist. For starters, it is expected to provide a window into planet formation theories (Udry & Santos 2007; Winn & Fabrycky 2015; Armitage 2020). Hot Jupiters are believed to have formed at larger orbital radii and then migrated to their current positions. During the course of this migration, volatiles may also have been dragged inward, thus resulting in the formation of small planets with volatile-rich inventories. Similarly, discovering planets in orbital resonance with one another is potentially indicative of the fact that they formed beyond the snow line and migrated inward,[7] thereby ending up with high volatile inventories. For example, theoretical models as well as the inferred densities suggest that the TRAPPIST-1 system was characterized by inward planetary migration (Unterborn, Desch, et al. 2018).

Apart from informing us about the orbital evolution and the composition of terrestrial planets, knowing the architecture of exoplanetary systems informs us about their habitability in other crucial respects. For instance, there has been much debate about whether giant planets akin to Jupiter have positive or negative effects on the development and sustenance of life on temperate rocky planets. Arguments in favor, for instance, tend to posit Jupiter-like planets as a shield against impacts by asteroids and comets that could cause regular mass extinction or, alternatively, deliver volatiles such as water. Planetary companions, especially gas giants, are expected to induce the slow temporal evolution of the target planet's obliquity and eccentricity (Waltham 2019). These long-term variations in the orbital parameters are known to engender quasiperiodic changes in the climate (known as Milankovitch cycles); on Earth, Milankovitch cycles

7. The snow line, sensu lato, embodies the distance from the host star at which simple volatile molecules like water, carbon dioxide, and methane condense to form solid ice grains.

drive the onset of glaciation and deglaciation over timescales of $\sim 10^4$ to 10^5 yr. Therefore, gaining a thorough understanding of planetary architecture may be necessary for assessing the long-term climatic stability of temperate exoplanets.

In closing, it must be said, that our current understanding of planet formation and the ramifications of exoplanetary architecture for habitability is not definitive. Hence, while determining both of these factors is advantageous, any conclusions drawn from them should be interpreted with due caution.

6.3.3 Atmospheric and ocean losses

The presence of liquid water on the surface and an atmosphere are both essential requirements for surface-based life. However, as we saw in Chapter 4, it is possible for M-dwarf exoplanets to lose their atmospheres and / or oceans over timescales shorter than the age of the Universe, consequently becoming uninhabitable. It is thus of paramount importance to identify observational metrics by which the extent of atmospheric and ocean losses are quantifiable.

It is only recently that the significance of isotopic fractionation for observations has gained appreciation. *Isotopic fractionation* refers to the enrichment of one isotope with respect to another via physical, chemical, geological, or biological mechanisms. Among other causal factors, atmospheric escape is known to drive isotopic fractionation. The overall process is rather complex, owing to which we will restrict ourselves to a heuristic description. As seen from (6.13), the scale height is larger for lighter elements. Hence, as one approaches higher altitudes, lighter isotopes will grow increasingly more abundant in comparison to heavier isotopes. In Section 4.2.2, we encountered mechanisms, such as ion pickup and sputtering, that play dominant roles in the escape of ions and neutrals. These mechanisms act to preferentially remove lighter isotopes owing to their relatively greater abundances for reasons explained above. In addition, from inspecting (4.45), we see that hydrodynamic escape also yields selectively elevated loss rates for lighter isotopes because the number of particles escaping per unit time is inversely proportional to the particle mass.

Thus, on planets where significant atmospheric escape has occurred, the abundance of heavier isotopes in the atmosphere should be higher than the "norm." By this logic, we may predict that Mars should have a

higher isotope ratio of deuterium-to-hydrogen relative to Earth if much of the Martian atmosphere was lost to space. This qualitative result is consistent with the latest observations by the Echelon Cross Echelle Spectrograph (EXES) instrument, which concluded that the D/H value for Mars is ~ 4 times higher than Earth's oceans (Encrenaz et al. 2018). When it comes to Venus, multiple lines of evidence indicate that the majority of its water has been depleted. It is therefore not surprising that its atmospheric D/H ratio is $\gtrsim 100$ times higher than the corresponding value for Earth's oceans. In a similar vein, extensive depletion of heavier molecules such as O_2 and CO_2 can give rise to isotope fractionation of their constituent elements.

If the fractionation process is highly pronounced (i.e., with ratios that are ~ 100 times higher relative to Earth), new chemical species such as HDO (the isotopic analog of H_2O), CH_3D, and $^{13}CO_2$ will come into play and duly modify the spectra of atmospheres. Thus, by searching for distinctive spectral features induced by the introduction of these species, it might be possible to identify whether the worlds under question have undergone massive atmospheric and/or ocean (i.e., H_2O) losses. Mollière and Snellen (2019) investigated this issue and found that HDO may beget spectral signatures at 3.7 μm that are detectable by 40 m telescopes. If Proxima b is water rich, analyzing its reflected light could yield evidence of HDO after about one night of integration time.

In similar fashion, Lincowski et al. (2019) analyzed the prospects of detecting molecules such as HDO in the atmospheres of the TRAPPIST-1 planets. They determined that if the D/H ratios were comparable to that of Venus and the $^{18}O/^{16}O$ ratios were ~ 100 times higher than that of Earth's oceans, then discerning the isotopologues (molecules with different isotope composition) of H_2O and CO_2 at an SNR of 5 was realizable in as few as four to eleven transits using transmission spectroscopy.

6.3.4 Stellar parameters

One of the chief auxiliary benefits of studying exoplanets is the parallel characterization of their host stars. In Chapter 4, we saw how stellar properties, most notably the stellar mass $M_\star$, influence various aspects of planetary habitability.

Many of the basic physical parameters of the host star—such as the mass, radius, effective temperature, and age—can be determined to varying degrees of accuracy by observing the spectral energy distribution (SED)

and the distance of the star (using stellar parallax) from Earth.[8] These two observational inputs may be combined with stellar evolutionary models to extract further information regarding the host star. Another important means of estimating basic stellar parameters is astroseismology, the study of stellar oscillations. Astroseismology has been successfully employed on the *Kepler* data set to derive the masses and radii of several stars to an accuracy of a few percentage points.

The SED is particularly useful for assessing the propensity of planets to host life, since photons at different wavelengths impact exoplanetary habitability via several routes summarized in Chapter 4. It is worth recalling a couple of examples. XUV radiation facilitates atmospheric loss, the depletion of oceans, and the buildup of abiotic O_2 in the atmosphere. Visible radiation, and perhaps near-IR radiation as well, is utilizable for photosynthesis. Probing stellar atmospheres at distinctive spectral lines (France et al. 2016; Astudillo-Defru et al. 2017), such as the H and K transitions of Ca II (at 396.8 and 393.3 nm, respectively) and Lyα (121.6 nm), has been shown to yield valuable insights into the magnetic activity and stellar magnetic fields, age, and rotational period.

Lastly, on the basis of variations in the observed brightness of light curves, the existence of flares and superflares (with energies $\gtrsim 10^{26}$ J) can be inferred. Understanding the flaring frequency and amplitude is of particular importance because stellar flares have a wide range of effects, both beneficial and detrimental, on planetary habitability. Furthermore, the burst of radiation emitted by flares is reflected by planets, with the ensuing "echoes" enabling their detection (Sparks et al. 2018). Missions such as *Evryscope, Kepler, MOST* (Microvariability and Oscillations of Stars Telescope), and *TESS* (Transiting Exoplanet Survey Satellite) have greatly advanced our understanding of how stellar flares are modulated by characteristics like the spectral type and rotation period.

6.4 GASEOUS BIOSIGNATURES

Gaseous biosignatures are generated directly through biological processes or indirectly via the formation of secondary compounds by dint of biogenic mechanisms (Seager et al. 2012). Many of the gaseous biosignatures we

8. The SED generally comprises the spectral irradiance (irradiance per unit wavelength) as a function of the wavelength.

encounter subsequently could arise from nonbiological phenomena; it is therefore important to distinguish false positives from actual signatures of life. Gaseous biosignatures are detectable, in principle, through the characterization of exoplanetary atmospheres using methods outlined previously; the reader may also consult Heng (2017), Deming and Seager (2017), and Madhusudhan (2019). The idea of detecting extraterrestrial life via biosignature gases has a long and distinguished history; it was propounded by the likes of James Lovelock and Nobel laureate Joshua Lederberg in seminal papers authored during the 1960s (Lederberg 1965; Lovelock 1965; Hitchcock & Lovelock 1967). Here, we will delineate some of the prevalent candidates in the scientific literature on gaseous biosignatures.

6.4.1 Molecular oxygen

Of all the gaseous biosignatures, none has been investigated to the same degree as molecular oxygen (O_2). The reader is referred to Meadows et al. (2018) for an exhaustive summary of the (de)merits of O_2 as a biosignature gas.

On Earth, O_2 is the second most dominant gas in the atmosphere and arose as a consequence of oxygenic photosynthesis. The mechanisms underlying the advent of oxygenic photosynthesis and the complex nonmonotonic rise in atmospheric O_2 levels on Earth are reviewed in Chapter 3, owing to which we shall not retread this subject here. It suffices for now to recall a few points of import. First, one of the central advantages associated with oxygenic photosynthesis is that the requisite raw materials—such as CO_2, water, and photons in the visible range—are abundant insofar as the Earth is concerned. Second, oxygenic photosynthesis does not add O_2 directly to the atmosphere per se, because the buildup of O_2 occurs via the burial of organic matter; see Section 4.3.5 for more details. Last, the history of Earth's O_2 levels is complicated, but it comprises three broad phases (anoxic, hypoxic, oxic), outlined in Chapter 3.

A number of spectral features are associated with molecular oxygen, some of which we encountered earlier. In the visible and near-IR regions, the most noteworthy features include the O_2-A band at $0.76\ \mu$m, the O_2-B band at $0.69\ \mu$m, and the O_2-g band at $0.63\ \mu$m. Two additional features at $1.06\ \mu$m and $1.27\ \mu$m, which either partly or wholly originate from dimer O_2-O_2 (referred to henceforth as O_4) collisionally induced absorption,

come into play when looking for signatures of ocean loss and planetary desiccation (Leung et al. 2020). In the mid-IR a prominent absorption band exists at 6.4 μm, whereas in the UV range strong absorption (due to photodissociation) by O_2 is manifested at $\lesssim 0.2$ μm. The 6.4 μm feature, also a result of collision-induced absorption, has gained attention recently since it could effectively diagnose high O_2 pressures suggestive of abiotic origin (Fauchez et al. 2020). Of the various features, it is believed that the O_2-A band merits higher preference for direct imaging studies, owing to its low probability of being mistaken for a different gas and the strength of this band.

With reference to atmospheric O_2, not one but two major hindrances arise in connection with its status as a biosignature gas. The first concerns the issue of *false negatives,* corresponding to planets hosting widespread life that is nevertheless hard or impossible to detect via gaseous biospheres; such biospheres are classifiable under the category of cryptic biospheres. On Earth, the O_2-A band is predicted to have an appreciable depth only when the partial pressure of O_2 was $\gtrsim 1$ percent PAL (present atmospheric level). In other words, because this criterion has been consistently satisfied in the past ~ 0.5 Gyr, detecting atmospheric O_2 may be feasible only during the last ~ 10 percent of Earth's current lifetime (Reinhard, Olson, et al. 2017). A potential exception to this trend is that the atmospheric O_2 levels could have exceeded ~ 1 percent PAL during the Lomagundi Event, which unfolded in the Paleoproterozoic from ~ 2.2 to ~ 2.0 Ga. Looking beyond Earth, it is plausible that a subset of worlds with only oceans on the surface are susceptible to false negatives, whereby atmospheric O_2 content is muted for reasons elucidated in Section 5.5.

In order to bypass the issue of false negatives pertaining to atmospheric O_2, detailed observations in conjunction with theoretical modeling are necessary (Lisse et al. 2020). For instance, as noted in Chapter 3, the eventual rise of atmospheric O_2 to modern levels might be attributable to oxygenic sources either dominating, outlasting, or suppressing environmental sinks. In this event, ceteris paribus, it would make sense to target older planets that have had sufficient time to permit the buildup of O_2. It will also be necessary to search for signatures of volcanism or plate tectonics that can exude reducing gases into the atmosphere, thereby acting as O_2 sinks. Therefore, viewed collectively, there are plausible grounds for supposing that younger planets with reducing atmospheres and volcanic activity are not well suited for detecting signatures of oxygenic photosynthesis.

The second major issue that we wish to point out is false positives, which were introduced at the beginning of this chapter. Several pathways have been identified for the production of abiotic O_2, but we shall address only a couple of major ones herein. Planets with sizable CO_2 inventories are susceptible to photolysis via the reaction

$$CO_2 + h\nu \rightarrow CO + O, \tag{6.43}$$

provided that the wavelength of the photon is < 175 nm. The oxygen (O) thus liberated through the above photochemical reaction can undergo further reactions to yield O_2 and O_3 (ozone). Hence, if CO is detected at high concentrations ($\gtrsim 100$ ppmv) in conjunction with O_2, it might imply that the latter is abiogenic in origin. The most distinctive spectral features worth searching for in this context are the CO absorption bands at 2.35 μm and 4.6 μm; as this process requires substantial CO_2 inventories, detecting IR features at 1.6, 2.0, and 4.3 μm stemming from CO_2 could also reveal the existence of abiotic O_2.

While this approach comprises a promising means of identifying false positives, it represents an oversimplification. Schwieterman, Reinhard, Olson, Ozaki et al. (2019) performed numerical simulations that took multiple microbial metabolic pathways and atmospheric chemistry into account to calculate the expected levels of atmospheric CO for various exoplanets. It was shown that high CO levels ($\gtrsim 100$ ppmv) are actually compatible with the presence of O_2-rich atmospheres when it comes to late M-dwarfs analogous to Proxima Centauri. Hence, it appears as though CO ought not always be construed as an antibiosignature (indicative of the absence of life); on the other hand, high levels of CO might prove toxic for complex life. From this example, we see that assessing the likelihood of life on a given exoplanet must be painstakingly undertaken on a case-by-case, contextual basis.

The second route for the buildup of abiotic O_2 was covered in Section 4.3.1. Massive O_2 atmospheres on the order of hundreds of bars may build up on planets around late M-dwarfs, as these stars pass through an extended luminous pre-main-sequence phase during which the desiccation of oceans and the escape of hydrogen (H) to space can occur. At such high concentrations, the collisionally induced absorption O_4 (O_2-O_2 dimer) features become prominent, as the absorption coefficient for O_4 exhibits a quadratic dependence on the density of O_2 molecules. Two distinctive features in

the IR associated with O_4 absorption are the bands at 1.06 and 1.27 μm, as mentioned in Section 6.4.1. Additionally, future telescopes capable of probing the UV and visible ranges (unlike the JWST) would search for O_4 features at 0.36, 0.445, 0.53, and 0.63 μm.

6.4.2 Ozone

The concentration of ozone (O_3) is closely connected to the abundance of O_2 in the atmosphere. The Chapman scheme describes how the photolytic formation and destruction of O_3 functions; it can be expressed as follows:

$$\begin{aligned} &O_2 + h\nu(\lambda < 240\,\text{nm}) \rightarrow O + O \\ &O + O_2 + \mathcal{X} \rightarrow O_3 + \mathcal{X} \end{aligned} \tag{6.44}$$

$$\begin{aligned} &O_3 + h\nu(\lambda < 320\,\text{nm}) \rightarrow O_2 + O \\ &O + O_3 \rightarrow O_2 + O_2 \end{aligned} \tag{6.45}$$

$\mathcal{X}$ denotes any molecule that transports the excess energy released during this reaction, and λ represents the desired photon wavelength. Clearly, the formation of ozone depends on the availability of UV photons, implying that planets with similar concentrations of O_2 are likely to have divergent O_3 abundances due to the differences in stellar properties. For instance, planets around M-dwarfs have a higher propensity to build up abiotic O_3 because of the comparatively higher UV fluxes at $\lambda < 200$ nm originating from chromospheric emission. Yet, at the same time, we saw in Section 4.4.3 that energetic particles contribute to ozone depletion, and they are more prominent for M-dwarf exoplanets due to higher stellar activity. Hence, it is not easy to draw a clear conclusion, prima facie, about which class of planets will exhibit stronger O_3 features.

O_3 has a number of distinctive spectral features, of which the most important may be the Hartley-Huggins bands centered at 0.25 μm and ranging between 0.15 and 0.35 μm. The chief advantage of the Hartley-Huggins feature is that it remains distinctive even when the O_2 concentration is ~ 1 percent PAL. Ipso facto, even if the reflected spectra do not evince any clear O_2 signals, the Hartley-Huggins bands are still detectable (Meadows et al. 2018). Hence, this feature opens up the possibility of surmounting the O_2 false negative conundrum discussed earlier. While we saw

that biological O_2 was probably detectable only during the last ~ 10 percent of Earth's history, biogenic O_3 was potentially discernible since the Great Oxidation Event (GOE) at ~ 2.4 Ga, thus corresponding to ~ 50 percent of our planet's lifetime.

Note, however, that the Hartley-Huggins bands exist in the UV and therefore require telescopes to have the capacity to probe this range. In the visible range, the Chappuis bands are manifested at 0.5 to 0.7 μm and are partly responsible for the overall U-shaped spectrum of the Earth in this regime. The Chappuis bands are sensitive to the abundance of O_3 in the lowest layer of the atmosphere (troposphere) and constitute a viable means of identifying O_3. Moving on to the IR, a number of weak features have been documented at 2.05, 2.5, 3.3, and 4.8 μm. Lastly, in the mid-IR, O_3 produces a strong band at 9.65 μm, but it must be cautioned that there may be some degree of overlap with the CO_2 band at 9.4 μm.

6.4.3 Methane

In Sections 3.2 and 5.4.2, we encountered some of the abiotic and biological pathways for the production of methane. Most of the abiotic routes for CH_4 synthesis rely on Fischer-Tropsch-type reactions, expressible as

$$CO_2 + 4H_2 \leftrightarrow CH_4 + 2H_2O. \tag{6.46}$$

At the high temperatures where this reaction usually operates (~ 473–623 K), CO_2 is thermodynamically more stable, owing to which the yield of methane is quite low. The yield of methane increases at lower temperatures or when the mantle is less oxidized—that is, when its oxygen fugacity is lower relative to Earth. The biological production of methane proceeds via two different channels, with (6.46) constituting one of them. The second involves the decomposition of acetic acid to yield methane and carbon dioxide:

$$CH_3COOH \rightarrow CH_4 + CO_2 \tag{6.47}$$

Distinguishing between abiotic and biotic CH_4 is potentially feasible through a couple of interconnected avenues (Krissansen-Totton, Olson & Catling 2018). First, the fluxes of methane produced via biology ($\sim 10^{15}$ $m^{-2}\,s^{-1}$) are predicted to be at least an order of magnitude higher than those

generated sans biotic activity. Hence, if the mixing ratio of methane in the atmosphere is $\sim 10^{-3}$–10^{-2}, it may possess a biogenic origin. Second, even if the abiotic fluxes of methane are comparable to biological fluxes, the former are expected to be accompanied by correspondingly high fluxes of CO. In contrast, when it comes to the latter, microbes known as acetogens are capable of consuming CO to yield acetic acid, as per the simplified reaction:

$$4CO + 2H_2O \rightarrow 2CO_2 + CH_3COOH \tag{6.48}$$

Therefore, the simultaneous detection of CO and CH_4 has been argued to comprise a signature of abiotic methane production. Yet, this statement is not robust since Schwieterman, Reinhard, Olson, Ozaki et al. (2019) demonstrated via simulations that the simultaneous existence of CO and CH_4 in the presence of life is feasible under certain circumstances.

The detection of methane is very sensitive to the atmospheric composition as well as to the choice of host star. Let us first consider the former. To leading order, the methane abundance will be inversely proportional to the concentration of O_2 and O_3. This is because CH_4 and O_2 should react rapidly along the following lines,

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O, \tag{6.49}$$

and therefore lead to the depletion of CH_4. The methane abundance prior to the GOE was potentially $\gtrsim 10^2$ times higher than the modern value of $\sim$ 1 ppmv. Hence, it is believed that methane would not have been detectable in Earth's atmosphere for nearly 50 percent of its history—namely, over the past $\sim$ 2.4 Gyr (Reinhard, Planavsky, et al. 2017). Now, let us consider the effect of the host star. The primary sink for CH_4 is the hydroxyl (OH) radical via the reactions

$$\begin{aligned} CH_4 + OH &\rightarrow CH_3 + H_2O, \\ CH_3 + OH &\rightarrow HCHO + H_2. \end{aligned} \tag{6.50}$$

The OH radical is derived, either directly or indirectly, from the photolysis of water and depends on the UV flux in the appropriate wavelength range. In the case of late M-dwarfs akin to Proxima Centauri, for the same CH_4 flux as modern Earth, the mixing ratio of methane is $\sim 10^{-3}$ for

atmospheres that resemble either that of modern (oxic) or Hadean-Archean (anoxic) Earth (Schwieterman, Reinhard, Olson, Ozaki et al. 2019).

In the infrared regime, the strongest features of methane are manifested at 1.65, 2.4, 3.3, and 7–8 μm. There are also weaker features in the visible and IR at 0.6, 0.7, 1.0, and 1.4 μm. The difficulty with using the canonical 7–8 μm band is that it overlaps partly with H_2O absorption as well as the N_2O band at the same wavelengths. Hence, the identification of methane requires high spectral resolution to distinguish its spectral lines from those produced by other species.

6.4.4 Nitrous oxide

Nitrous oxide is an important byproduct during the incomplete denitrification of nitrate to molecular nitrogen. We briefly touched on the denitrification pathway in Section 3.2, which can be expressed schematically as follows:

$$NO_3^- \rightarrow NO_2^- \rightarrow NO + N_2O \rightarrow N_2 \tag{6.51}$$

In identifying potential gaseous biosignatures, it is important to ensure that the abiotic fluxes of the species under consideration are generally much smaller than the biological fluxes. This criterion will indicate that the detection of the species may constitute a reliable proxy for the existence of life. The concentration of N_2O in the Earth's atmosphere prior to the explosive growth of technology was ~ 270 ppb (1 ppb translates to an abundance of 10^{-9}).

Although the above mixing ratio comes across as rather low (making N_2O harder to detect), two points must be borne in mind. First, during the denitrification pathway, the transformation of N_2O into N_2 necessitates the action of an enzyme known as nitrous oxide reductase. The key point here is that the catalytic action of nitrous oxide reductase is made possible by a central site (Cu_Z) that is a copper-sulfide cluster. Hence, if the access to copper was constrained, as was perhaps the case in the sulfidic and anoxic (euxinic) oceans of Proterozoic Earth, the concentration of N_2O could have been higher. Second, for the same input flux of N_2O, the atmospheric abundance may be orders of magnitude higher for planets around M-dwarfs. The underlying reason has to do with the lower fluxes of near-UV photons that contribute to the depletion of N_2O.

In terms of generating abiotic N_2O fluxes, a couple of pathways merit a mention. The first is lightning, but it has been estimated to produce only $\sim 2 \times 10^{-5}$ of the total atmospheric N_2O. The second possibility involves a combination of XUV radiation and stellar energetic particles (SEPs), as noted in Section 4.4.3. In particular, N_2O could be synthesized from the radicals NO and NH (Airapetian et al. 2016).

$$NO + NH \rightarrow N_2O + H \tag{6.52}$$

There are a couple of points worth appreciating here. The N_2O flux from this pathway is dependent on the frequency at which stellar proton events occur and the fluence of SEPs impacting the planetary atmosphere. Both of these factors are conceivably higher by a couple of orders of magnitude for planets orbiting active M-dwarfs relative to young G-type stars. Thus, further work is necessary to assess the abiotic contribution for M-dwarf exoplanets. A potential means of distinguishing false positives on such planets is that SEPs also produce high abundances of nitrogen oxides (NO_x). Hence, if the latter are found to coexist with N_2O at high concentrations, it may indicate that the observed N_2O is abiogenic in origin.

The most significant bands associated with N_2O exist in the IR at 3.7, 4.5, 7.8, and 8.6 μm. There are, however, a couple of caveats. First, most of these bands are relatively weak for planets with Earthlike abundances of N_2O. Second, many of these features overlap with those arising from other gases. Collectively, the detection of N_2O is challenging and will require the use of high-resolution spectroscopy to demarcate and identify N_2O.

6.4.5 Sulfur gases

A number of sulfur-bearing gases are produced by lifeforms on Earth, but most of the simplest ones (e.g., hydrogen sulfide and sulfur dioxide) are also generated copiously by abiotic processes. Instead, we must look toward organosulfur gases, such as dimethylsulphide (CH_3SCH_3), dimethyl disulfide ($CH_3S_2CH_3$), and methanethiol (CH_3SH), which are primarily generated by microbes on Earth. The major source of dimethylsulphide, for instance, is traceable to the biogenic decomposition of the compound dimethylsulfoniopropionate that is found in marine eukaryotes. Microbial mats have been documented to produce dimethylsulphide, dimethyl

disulfide, and CH_3SH, possibly by way of harnessing sulfide compounds generated by sulfate-reducing bacteria.

The abundances of these gases on modern Earth are low, but it has been hypothesized that their concentrations were higher in the comparatively reducing environments of early Earth. However, even at fluxes that are about an order of magnitude higher than the currently observed values, spectral detection of these molecules is likely to be very challenging. The only viable scenario for identifying them is on planets orbiting M-dwarfs, because these worlds receive relatively lower UV fluxes, which are responsible for the degradation of organosulfur gases (Domagal-Goldman et al. 2011). Since the direct detection of these molecules is difficult, indirect strategies become necessary.

The most prominent among them entails searching for signatures of ethane (C_2H_6). UV light is capable of splitting CH_3 radicals from dimethylsulphide and dimethyl disulfide, resulting in the synthesis and buildup of ethane above expected levels. It has been estimated that the abiotic production of ethane from methane is much smaller than the generation of ethane from organosulfur gases, thus implying that detecting C_2H_6 would serve as a proxy for the likes of dimethylsulphide and dimethyl disulfide. However, to properly assess whether the detected ethane is indirectly biogenic in origin, it would require the measurement of CH_4 abundance in conjunction with extensive theoretical modeling that properly accounts for other atmospheric components, thermal and UV profiles.

The prominent spectral features for the organosulfur gases are as follows:

- 6–7, $\sim$10, and $\sim$15 μm for dimethylsulphide,
- 7, 8–9, and 17 μm for dimethyl disulfide, and
- 6–7, 8–12, and 14–15 μm for methanethiol.

Ethane exhibits strong features at 6–7 and 11–13 μm. In many cases, the spectral bands overlap with one another and thus require sufficiently high spectral resolution. Numerical simulations suggest that the mixing ratios of dimethylsulphide, and ostensibly its kin, evince distinct spatial anisotropy on tidally locked M-dwarf exoplanets, with the dayside and nightside magnitudes diverging by a factor of $\sim$1.5 (H. Chen, Wolf, et al. 2018).

6.4.6 Methyl chloride

Methyl chloride (CH_3Cl), also known as chloromethane, is primarily generated by biogenic sources on Earth. These include marine phytoplankton, aquatic and land plants, fungi, and the decay of organic matter. Volcanic activity constitutes a notable abiotic source of methyl chloride. Note that this compound has also been observed by (1) the Atacama Large Millimeter / submillimeter Array (ALMA) around the binary star system IRAS 16293-2422 and (2) the *Rosetta* spacecraft in the tenuous atmosphere of the comet 67P/Churyumov-Gerasimenko (67P/C-G). The fluxes of CH_3Cl produced by biological and abiotic sources are not well constrained on Earth; it is therefore difficult to assess whether methyl chloride can be regarded as a robust gaseous biosignature.

Some of the notable spectral features associated with CH_3Cl include 3.3, 7, 9.7, and 13.7 μm. Many of these bands overlap with those of other gases, with one example being the 9.65 μm feature of ozone. The accumulation of CH_3Cl to significant levels is unlikely for solar-type stars but might be feasible for planets around quiescent M-dwarfs that receive low UV fluxes. CH_3Cl may also constitute a viable biosignature on planets with H_2-dominated atmospheres; the same could apply to other gaseous biosignatures encountered earlier, such as dimethylsulphide and N_2O (Seager et al. 2013). Ammonia (NH_3) has been posited as a prospective biosignature in N_2-H_2 atmospheres since the following reaction is exothermic in nature:

$$N_2 + 3H_2 \rightarrow 2NH_3 \tag{6.53}$$

6.4.7 Hazes and other gaseous biosignatures

We discussed atmospheric hazes and the conditions for their production in Section 5.4.3. In this context, it suffices to note that the required flux of methane to generate thick hazes on Hadean-Archean Earth was on the order of 10^{15} m^{-2} s^{-1} (Arney et al. 2016), which happens to be comparable to the present-day methane flux. Hence, for planets with microbial biospheres akin to that of Earth, the buildup of thick hazes is conceivable provided that the atmospheres are anoxic. Hazes are particularly useful as gaseous biosignatures since they drastically modify the spectra of reflected light from planets, more so at UV and blue wavelengths, as seen clearly in Figure 6.6. Hence, even at low spectral resolution ($\mathbb{R} \sim 10$), the identification of organic hazes may be feasible.

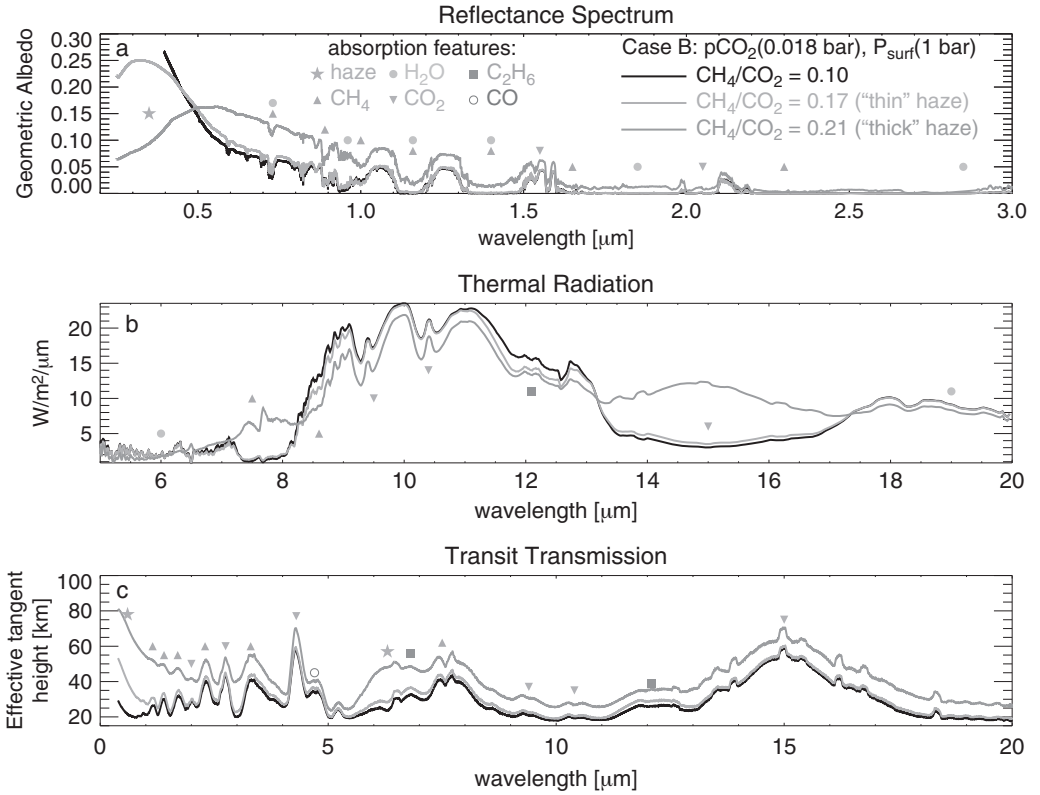


Figure 6.6 The spectral features manifested in the reflected light, thermal, and transmission spectra for different ratios of CH_4/CO_2 that functions as a proxy for the thickness of the haze. *Top:* Striking differences between the haze-free ($CH_4/CO_2 = 0.1$) and thick haze cases are manifested in the reflectance spectrum at UV and blue wavelengths. Note that the effective tangent height quantifies the change in the planet's effective radius arising from an increase in absorption by hazes. (© The Authors. Published by Mary Ann Liebert. CC-BY-NC-4.0. *Source:* Giada Arney, Shawn D. Domagal-Goldman, Victoria S. Meadows, Eric T. Wolf, Edward Schwieterman, Benjamin Charney, Mark Claire, Eric Hébrard, and Melissa G. Trainer [2016], The pale orange dot: The spectrum and habitability of hazy Archean Earth, *Astrobiology* 16[11]: 873–899, fig. 8 a, b, c.)

Although the detection of organic hazes does not suffice by itself to confirm the presence of life,[9] it nonetheless represents a promising sign

9. Titan, for instance, possesses a thick haze, but whether life exists on this moon remains unknown.

that merits follow-up observations. Organic hazes, by virtue of their origin from methane, enable us to constrain the methane flux if the abundance ratio CH_4/CO_2 is known. Likewise, the detection of sulfur aerosols (H_2SO_4 and S_8) could permit the determination of the H_2S/SO_2 ratio and the flux of H_2S. Note that the H_2S flux, viewed in isolation, does not inform us about the existence of life since it also originates from volcanic activity. Sulfur hazes do, however, allow us to distinguish between oxidizing and reducing atmospheres: identifying S_8 aerosols is emblematic of a neutral or reducing atmosphere, whereas finding H_2SO_4 implies oxidizing conditions.

The list of gaseous biosignatures that we have delineated so far is by no means comprehensive. A vast number of other gases are generated as secondary metabolic products on Earth, although most of them exist only in trace amounts. For a comprehensive list of volatile compounds that are possible biosignature gases, refer to Seager et al. (2016). One such gas that has recently garnered much attention is phosphine (PH_3), given that one of its chief sources on Earth is anaerobic bacteria. Numerical simulations indicate that the concentration of PH_3 might reach $\mathcal{O}(100)$ ppm on worlds with anoxic atmospheres under optimal circumstances. Phosphine has three major spectral features at 2.7–3.6 μm, 4.0–4.8 μm, and 7.8–11 μm, and a potential detection of this gas by JWST could necessitate an integration time on the order of ten hours (Sousa-Silva et al. 2020). The ostensible identification of phosphine in the cloud decks of Venus (described in Section 5.7) has stimulated further interest in this gas, especially with respect to fathoming the various channels (biological or otherwise) by which it is produced.

6.5 SURFACE BIOSIGNATURES

Surface-based life has the capacity to alter the observed spectrum via the absorption and reflection of light. These twin processes may be regulated by life in a number of ways such as biological pigments, scattering from physical assemblages of organisms, organic matter degradation, and fluorescence. In many instances, life has been shown to produce distinctive "edges" at certain wavelengths in the reflectance spectra, as we shall see shortly.

6.5.1 Photosynthesis: oxygenic and anoxygenic

Oxygenic photosynthesis, as we saw earlier, is already responsible for one of the most prominent biosignature candidates: molecular oxygen. In addition,

photosynthetic pigments used in the capture of light energy also produce discernible biosignatures that form the subject of our discussion. Before tackling this topic, we refer the reader to Sections 3.2, 3.3, and 4.3.5 for an overview of the basic principles underlying photosynthesis (both anoxygenic and oxygenic).

The primary role of pigments is to harvest light energy. Antenna pigments are responsible for absorbing photons across a certain wavelength range and transmitting this energy to the primary photopigment, which is also known as the reaction center (RC) pigment. The energy provided to the primary photopigment leads to the excitation of an electron across the band gap, followed by its ejection and capture by an electron acceptor. A physical constraint on the photopigment is that it must possess a higher oxidation power than the reducing agent (e.g., water in oxygenic photosynthesis) employed in photosynthesis. Clearly, pigments play a multitude of roles in photosynthesis; hence it is instructive to further unpack their properties. The wavelengths absorbed by pigments depend on their oxidation state, functional groups, and surrounding proteins.

In Figure 6.7, which shows the absorbance spectra for various photopigments, it is evident that the bacteriochlorophylls found in anoxygenic photoautotrophs absorb in the near-IR ranging from $\sim$710 to 1040 nm. Such pigments are well suited for functioning on M-dwarf exoplanets as the host stars tend to emit more radiation in this range. In contrast, the chlorophylls extracted from oxygenic photoautotrophs absorb primarily in the visible range, with the exception of the recently discovered Chl d and Chl f, which exhibit absorbance peaks close to the near-IR. Note that Figure 6.7 cannot be directly applied to organisms for a couple of reasons. First, the absorbance spectra is subject to variability by $\mathcal{O}(10)$ nm, based on the ambient proteins. Second, depending on the available photon fluxes, organisms are capable of tuning their composition of pigments to optimize photon capture and usage. For instance, cyanobacteria that occur in environments with enriched far-red light synthesize Chl f and modified phycobiliproteins (antenna pigments) to harness this radiation for carrying out oxygenic photosynthesis (Ho et al. 2016).

With these caveats out of the way, we turn our attention to the famous red edge of vegetation; we will refer to it subsequently as the vegetation red edge (VRE). In Figure 6.7, we see that the absorbance of Chl a and Chl b peaks in the range $\sim$660–700 nm, which happens to coincide with the peak in solar photon spectral flux. In contrast, light is scattered in the

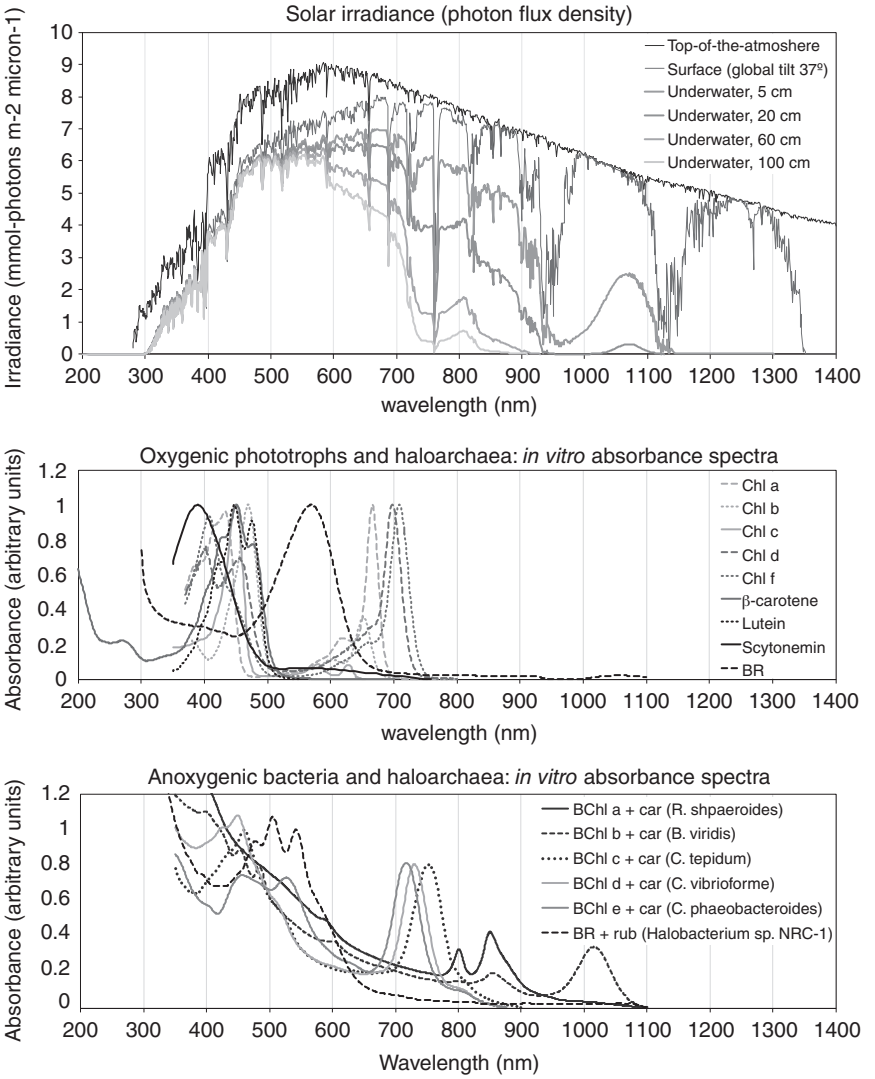


Figure 6.7 *Top:* Solar photon spectral flux at different regions ranging from top of the atmosphere to 1 m underneath the surface of water. *Middle:* Absorption spectra in vitro for pigments extracted from oxygenic photoautotrophs such as chlorophylls (Chls); most pigments have peaks at ~450 nm and ~700 nm. *Bottom:* Absorption spectra in vivo for pigments derived from anoxygenic photoautotrophs such as bacteriochlorophylls (BChls). *Notes:* (1) the absorbance is measured in normalized units, (2) the spectra in the middle graph must be blueshifted by ~5–40 nm to obtain in vivo spectra, and (3) BR, car, and rub denote bacteriorhodopsin, carotenoids, and bacterioruberin, respectively. (© The Authors. Published by Mary Ann Liebert, Inc. CC-BY-NC-4.0. *Source:* Edward W. Schwieterman, Nancy Y. Kiang, Mary N. Parenteau, Chester E. Harman, Shiladitya DasSarma, Theresa M. Fisher, Giada N. Arney, Hilairy E. Hartnett, Christopher T. Reinhard, Stephanie L. Olson, Victoria S. Meadows, Charles S. Cockell, Sara I. Walker, John Lee Grenfell, Siddharth Hegde, Sarah Rugheimer, Renyu Hu, and Timothy W. Lyons [2018], Exoplanet biosignatures: A review of remotely detectable signs of life, *Astrobiology* 18[6]: 663–708, fig. 9.)

near-IR at $\sim$760–1100 nm, due to the absence of absorbing pigments at these wavelengths in conjunction with changes in the refractive index of leaf tissues. Together, these characteristics combine to yield a sharp increase in reflectance at approximately 680 nm (red wavelengths) that flattens at $\sim$760 nm. The precise properties of this red edge—such as the amplitude of the increase in reflectance and the midpoint of the edge—fluctuate depending on both the health and taxa of organisms.

Despite this variation, all oxygenic photoautotrophs ranging from terrestrial and aquatic plants to mosses, lichens, and algae exhibit the red edge spectral feature. The bottom graph in Figure 6.4 shows an example of the VRE, which is widely regarded as being unique to biology, and therefore a robust biosignature, but potential false positives have been proposed. Certain minerals possess reflectance edges that are similar in strength and location to the VRE; cinnabar (HgS) and crystalline sulfur, for instance, have sharp reflectance edges at $\sim$600 nm and $\sim$450 nm, respectively. The extent of the mineral edge will depend on the fraction of exposed rocks on the surface and the abundances of these minerals within the rocks. Another possibility, addressed in Section 9.5.4.1, is that artificial photosynthesis might be employed by technological species to efficiently harvest light energy; in this event, distinguishing between natural and artificial photosynthetic edges will require the consideration of additional environmental factors.

The most common metric used to quantify the strength of the VRE is the Normalized Difference Vegetation Index (NDVI), defined as

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}}}, \tag{6.54}$$

with ρ_{NIR} and ρ_{red} representing the reflectances at the near-infrared and red wavelengths, respectively. Another rubric, known as the VRE index, replaces the denominator in (6.54) with ρ_{red}. Calculations of the VRE index for light reflected by the Earth from the Sun (Earthshine) have illustrated that the values are typically on the order of a few percentage points. The detection of the VRE is challenging and requires at least 10 percent of the surface not obscured by clouds to be covered in vegetation. T. D. Brandt and Spiegel (2014) numerically estimated that an SNR of $\sim$40 to 120 is potentially necessary for identifying the VRE, although the optimal spectral resolution of $\mathbb{R} \sim 20$ is fairly low. In light of our earlier analysis, achieving

such an SNR is challenging and requires either very large telescope apertures or long integration times.

Hitherto, our discussion of oxygenic photosynthesis has implicitly dealt with an Earthlike planet orbiting a Sunlike star. When it comes to M-dwarf exoplanets, however, the situation can become very different. The red edge may shift toward longer wavelengths, thus becoming an infrared edge instead, as described in Section 4.3.5. Resolving the issue of the photon wavelengths at which spectral photosynthetic edges are manifested on planets around other stars is hugely challenging. It requires taking into account not only the stellar spectra but also the temporal fluctuations in photon intensity, accessory and RC pigment characteristics, and physiological responses, to name a few. Therefore, laboratory experiments probing the feasibility of oxygenic photosynthesis in the near-IR along with theoretical models to predict the edge spectra that could be generated by accessory and primary pigments are necessary.

Here, it must be noted that oxygenic photosynthesis is not the only pathway characterized by a spectral edge. Anoxygenic photoautotrophs also possess pigments with spectral edges in the near-IR that are a consequence of the absorbance peaks depicted in the bottom graph of Figure 6.7. Anoxygenic photoautotrophs on Earth are often found in microbial mat communities, owing to which multiple near-IR edges have been distinguished in measurements of reflectance spectra. In view of the fact that similar features may exist on other worlds, the need for rigorous theoretical modeling and empirical studies of microbial communities (Nadell et al. 2016) to further our understanding of ecosystem biosignatures is apparent.

Lastly, the VRE is useful for identifying oxygenic photosynthetic pigments, but it is not extremely difficult to distinguish between unicellular and multicellular photoautotrophs. Let us suppose that multicellular photoautotrophs on exoplanets exhibit treelike structures. This is actually a fairly reasonable premise since the resulting branching networks optimize the capture of light energy and transport of nutrients and water in *generic* multicellular photosynthetic organisms (G. B. West et al. 1997). Due to this fractal structure, the shadows cast by such species vary depending on the phase angle (defined in Section 6.2.1); this feature is not expected to be prominent for microbial photoautotrophs. As a consequence, perhaps multicellular photoautotrophs are identifiable by their distinctive bidirectional reflectance distribution function (BRDF), which is responsible for causing the reflectance to change with the phase angle.

Doughty and Wolf (2010) simulated the variation in albedo as the phase angle was altered from 0.5° to 44°. In the complete absence of clouds, it was found that the change in albedo was (1) 0.027 in the visible and 0.037 in the near-IR for zero vegetation and (2) 0.043 in the visible and 0.096 in the near-IR for Earthlike vegetation. As the magnitudes of these variations are clearly very small, Doughty and Wolf proposed that the second derivative of the planetary albedo could be employed instead to distinguish one case from the other. Although much more research in the vein of Doughty et al. (2020) is required, the BRDF effect might constitute a valuable metric for demarcating multicellular photosynthetic organisms from their microbial counterparts. Hence, it represents a step in the direction of formulating strategies for discerning complex multicellularity on exoplanets (Schulze-Makuch & Bains 2018).

6.5.2 Retinal pigments

While the significance of chlorophylls is well-known, other pigments are also capable of harvesting and utilizing light energy. The best known among them are the so-called retinal pigments that capture and utilize light energy for vision, ATP synthesis, and other fundamental biological processes. The most notable among the retinal pigments is bacteriorhodopsin, which we explored in Section 3.2. Bacteriorhodopsin acts as a light-driven proton pump and facilitates the synthesis of ATP. Given the importance of bacteriorhodopsin in metabolism, its widespread occurrence in microbes, and potentially early emergence, there are compelling grounds for viewing it as a prospective biosignature.

Bacteriorhodopsin absorbs strongly at wavelengths corresponding to the green regions of the visible spectrum, which is evident from inspecting Figure 6.7. This aspect is of much interest because none of the chlorophylls absorb strongly in this range, implying that bacteriorhodopsin constitutes a complementary pigment to the chlorophylls. One of the best-known examples of microbes with bacteriorhodopsin are the Haloarchaea, characterized by their bright-purple color. Retinal pigments akin to bacteriorhodopsin are theoretically predicted to yield green-yellow spectral edges that resemble the VRE and may be detectable by future telescopes. The prospects for their detection are naturally dependent on the spatial extent of these putative microbial biospheres. Recently, DasSarma and Schwieterman (2018) theorized that the Archean Earth might have been dominated

by light-harvesting organisms endowed with retinal-based pigments, thus giving rise to a "Purple Earth."

6.5.3 Miscellaneous pigments

Organisms are known to possess pigments for reasons other than the synthesis of organic compounds and metabolic functioning. The advantages conferred by pigments include UV screening (e.g., carotenoids), insulation against extreme thermal fluctuations, nutrient scavenging, antimicrobial and antioxidant protection, gene regulation mechanisms via *quorum sensing,* and signaling. In some environments such as the subsurface hydrothermal vents, organisms produce pigments even in the absence of light, thus revealing that the primary purpose of pigments is not necessarily to bestow color. In view of the myriad roles played by pigments, it is sensible to contend that the search for pigments should extend beyond those involved in capturing electromagnetic energy.

Several surveys have been conducted to determine the reflectance signatures of microbes. For instance, Hegde et al. (2015) investigated the spectral characteristics of 137 microbes, including a number of extremophiles. Most of the organisms exhibited strong absorption features in the visible range, owing to their pigments. The reflectance signatures for select microbes are illustrated in the bottom graphs in Figure 6.4. For example, the coloration of the archaeon *Halobacterium salinarum* is attributable to the carotenoid pigment bacterioruberin along with bacteriorhodopsin. Apart from the direct spectral signatures, the presence of intracellular water is manifested as hydration bands at 0.95, 1.15, 1.45, and 1.92 μm. However, distinguishing between such absorption bands and atmospheric H_2O is challenging. The extent to which nonphotosynthetic pigments will be discernible in remote-sensing observations depends on their surface coverage. If the microbes under question are exclusively confined to specialized environments such as hypersaline ponds and lakes (as on Earth), it is very unlikely that these signatures will be detectable in the near future.

6.5.4 Fluorescence and bioluminescence

As opposed to photons being absorbed or scattered by biological entities, there is another pathway not considered thus far—to wit, surface-based organisms may emit photons in their own right. Fluorescence represents

one of the classic examples in this category. Autofluorescence, in particular, entails the emission of light from electronically excited states created through the absorption of photons; the energies of emitted photons are usually lower than those absorbed by the material. Autofluorescence has been documented for abiotic substances (e.g., calcite) as well, which is why the fluorescence spectrum must be carefully evaluated to assess whether it is biogenic in nature.

Chlorophyll fluorescence is believed to have emerged with the aim of mitigating physiological stress. The fluorescent emission of Chl a ranges from 680 to 800 nm, with peak values attained at 685 and 740 nm. The fluorescent component arising from vegetation on Earth has been estimated to increase the visible spectral flux by only a few percentage points. Substances other than chlorophyll also exhibit autofluorescence. On planets that receive high fluxes of UV radiation, fluorescence may evolve as a UV-protection strategy by absorbing photons in the UV and emitting them in the visible range. Pigments and proteins derived from corals are well suited for this purpose since they absorb photons at $\sim$350–650 nm and evince emission peaks at $\sim$486–685 nm with quantum yields (number of emitted photons to absorbed photons) of $\sim$1–10 percent. O'Malley-James and Kaltenegger (2018) modeled the fluorescent emission from corals and found that the visible flux on planets around F-type stars could increase by $\mathcal{O}(10\%)$ under optimal conditions. This study assumed that the fraction of coral-like organisms covering the planetary surface was $\sim$30 to 100 percent, as opposed to the Earth, where corals cover merely 0.2 percent of the ocean floor.

Another mechanism for photon emission by lifeforms is bioluminescence. Bioluminescence operates via the oxidation of generic light-emitting molecules known as luciferins, and this mechanism is catalyzed by enzymes called luciferases. Bioluminescence has been documented in multiple taxa on Earth, ranging from bacteria and plankton to fish and insects. It has arisen more than forty or fifty times on Earth (Haddock et al. 2010) and could therefore constitute an example of evolutionary convergence. The photons emitted by luciferins vary widely both among and within species, but the emission peaks are often situated at yellow-green wavelengths ($\sim$500–600 nm). Bioluminescent bacteria might span areas as high as $\sim 10^{10}$ m^2 and emit a faint glow, but the spatial coverage on other worlds is poorly constrained because it depends on both ecological and evolutionary constraints; detecting such features is, however, likely to prove highly challenging in most cases.

6.5.5 Chiral signatures

One of the most distinctive features of life on Earth is homochirality, a topic that we encountered in Section 2.8.4. Due to the fact that many biopolymers (e.g., proteins) comprise only one enantiomeric form, substantial enantiomeric excesses of these compounds are expected in the presence of life. In contrast, abiotic enantiomeric excesses from carbonaceous meteorites are usually < 10 percent. Hence, detecting enantiomeric excess upward of 20 percent has a high likelihood of confirming biogenic origin and thereby serving as a reliable biosignature.

The chiral sites located in amino acids and sugars are responsible for the preferential absorption of left- or right-handed circularly polarized UV light. Likewise, photosynthetic pigments manifest similar behavior in the visible and IR ranges as they also possess chiral centers. This differential absorption (dubbed circular dichroism) has been shown to yield distinctive sinusoidal features in circular polarization spectra, potentially at wavelengths close to the absorption peaks of the appropriate chiral substances. Note, however, that the spectral shape and magnitude of the signals vary by two orders of magnitude for photosynthetic organisms on Earth. The degree of circular polarization ($P_{\rm cp}$) is defined as

$$P_{\rm cp} = \frac{I_{\rm RHC} - I_{\rm LHC}}{I_{\rm T}}, \tag{6.55}$$

where $I_{\rm RHC}$ and $I_{\rm LHC}$ are the intensities of left- and right-handed circularly polarized light, while $I_{\rm T}$ represents the total intensity. Laboratory experiments have illustrated that $P_{\rm cp} \sim 10^{-2} - 10^{-4}$ for many photosynthetic pigments, whereas abiotic materials yield values of $P_{\rm cp}$ that appear to be smaller by a factor of $\gtrsim 10$–100 (Patty et al. 2018). The weak strength of the signal presents a serious issue that is further compounded by the following facts: $P_{\rm cp}$ varies across species and is subject to temporal fluctuations (for the same species) due to physiological effects.

In addition to circular polarization, it is theoretically possible to identify chirality via linear polarization. The polarization degree is defined as the ratio of the polarized light intensity to that of the total light intensity. Berdyugina et al. (2016) carried out laboratory experiments to determine the polarization degree and simulated the spectra resulting from hypothetical planets with varying coverage of photosynthetic pigments. In the absence of clouds, it was shown that the polarization degree, which is roughly

commensurate with pigment absorption, was potentially higher by ~20 to 50 percent for planets with biopigments covering the entire surface compared to those worlds with only deserts; moreover, the shapes of polarization spectra were distinct from each other. Yet, at the same time, we caution that minerals, scattering, and absorption by gases can also give rise to linear polarization signals. Hence, the challenge with utilizing linear polarization for detecting life is to differentiate between abiotic and biogenic sources.

6.6 TEMPORAL BIOSIGNATURES

As the name suggests, temporal biosignatures reveal the presence of biological activity through temporal modulations. Temporal biosignatures are not independent of gaseous and surface biosignatures. Indeed, the time-dependent behavior is manifested as changes in either the concentration of atmospheric gases or the surface albedo. The canonical example offered in this category is the seasonal variation in the abundance of CO_2 that, in turn, reflects how the productivity of land plants evolves with temperature and stellar flux. Modeling temporal biosignatures is currently challenging because it necessitates accounting not only for orbital parameters (e.g., eccentricity) but also for the phase angle and surface heterogeneity.

6.6.1 Fluctuations in gaseous biosignatures

As noted above, CO_2 variability has been widely investigated as a seasonal biosignature. In qualitative terms, the fixation of carbon driving growth in vegetation is higher during the spring, thus leading to a decrease in the content of atmospheric CO_2. On the other hand, during fall and winter, atmospheric CO_2 increases due to a decrease in its consumption allied to the decomposition of plant matter. The amplitude of this oscillation is only ~3 ppmv near the Earth's equator but it can reach ~10 to 20 ppmv at higher latitudes. There are, however, some issues with invoking CO_2 seasonality as a temporal biosignature. One of the more notable ones has to do with the high solubility of CO_2 via the formation of carbonic acid; consequently, these oscillations may be diminished due to ocean-mediated chemical buffering.

Let us turn our attention to methane. The major issue with employing methane seasonality is that the oscillations are not purely biogenic in nature.

Instead, there is an abiotic contribution arising from the photolysis of tropospheric water vapor to yield the hydroxyl (OH) radical. The OH radical in turn contributes to the depletion of CH_4 via the first pathway delineated in (6.50). In other words, the fluctuations in methane abundance are linked to the availability of water vapor; the latter is generally higher in summer (and lower in winter) by virtue of enhanced evaporation at higher temperatures. Hence, although atmospheric methane is primarily derived from biological sources, its modulations are abiotic to a certain degree. Although this factor does not rule out the prospects for using methane seasonality as a temporal biosignature per se, it does render the situation more complex.

Next, we turn our gaze to oscillations in O_2 concentration. The concentrations of O_2 and CO_2 are coupled through the dual processes of aerobic respiration and photosynthesis. When ~ 1 mole of CO_2 is consumed, we expect ~ 1 mole of O_2 to be generated and vice versa. Thus, naively speaking, we may anticipate O_2 oscillations to be perfectly anticorrelated with CO_2 oscillations. However, an important distinction is that O_2 is much less soluble compared to CO_2, owing to which the amplitude of the fluctuations is effectively higher. At warm temperatures, not only is O_2 production (and CO_2 drawdown) higher but also the solubility of O_2 is reduced, permitting higher accumulation of O_2 in the atmosphere. At the middle latitudes of Earth, the amplitude of variations in O_2 abundance due to seasonality is predicted to be ~ 50 ppm.

Finally, we turn our attention to the seasonality evinced by ozone. As we saw in Section 6.4.2, the synthesis of ozone is directly connected to the availability of molecular oxygen as well as the UV flux. On account of the highly oxygenated atmosphere of modern Earth, the formation of O_3 is not limited by O_2 levels. Instead, the seasonal changes in O_3 are primarily driven by atmospheric transport in the stratosphere, christened the Brewer-Dobson circulation. The situation is potentially very different on worlds that are weakly oxygenated, with O_2 abundances of ~ 10 ppmv. S. L. Olson et al. (2018) modeled such an atmosphere where the oxygen levels fluctuated between 5 and 25 ppmv and found that the strength of the reflected signal at the Hartley-Huggins absorption band changes by a factor of ~ 2.5 between the winter minimum and the summer maximum. Hence, it is conceivable that the best diagnostic for O_2 seasonality on weakly oxygenated worlds is via the study of O_3 features in the UV.

6.6.2 Fluctuations in surface biosignatures

A number of the surface biosignatures we have encountered are characterized by temporal fluctuations. For instance, the VRE evolves with time and tracks the growth and death of vegetation during the seasons. The NDVI defined in (6.54) is subject to variability since dead, decaying, or desiccated vegetation is likely to have a lower value when compared to healthy vegetation. Apart from the VRE, the spatial extent of microbial mats is predicted to change with time, thus modifying the reflectance spectra from the corresponding biopigments accordingly. The circular and linear polarization signals are also time dependent because their amplitude depends on physiological responses to varying temperature and stellar flux.

Setting aside reflected or scattered light from pigments, fluorescence and bioluminescence are subject to inherent temporal variability. If, for example, fluorescence constitutes a UV-protection strategy, we would expect to see a spike in the emitted fluorescent light in the immediate aftermath of large stellar flares (O'Malley-James & Kaltenegger 2019a). Likewise, bioluminescence in the ocean appears to persist on timescales of days before subsiding. Strong tidal forces permit amplified nutrient upwelling that may trigger the explosive growth of unicellular organisms (Lingam & Loeb 2018b); the ensuing result would be analogous to harmful algal blooms (HABs) on Earth. These phenomena are transient in nature and have therefore been proposed as candidates for temporal biosignatures, although their detectability depends on the surface fraction encompassed by such blooms.

6.7 FALSE POSITIVES VERSUS REAL BIOSIGNATURES

When confronted with signals that may be indicative of biological activity, it is important to gauge whether they are robust in nature. Hence, we will briefly outline a few methods that have been proposed for identifying the plausibility of potential biosignatures.

6.7.1 Thermodynamic disequilibrium

The importance of thermodynamic disequilibrium in the atmosphere as a means of confirming the presence of life has been extensively discussed since the 1960s, as pointed out toward the beginning of Section 6.4. The

best example of thermodynamic disequilibrium is the simultaneous coexistence of CH_4 and O_2 in the atmosphere; in the absence of life, these two gases should yield CO_2 and H_2O, as seen from (6.49). The fact that these two gases coexist is only because they are being continuously replenished by biological sources. However, the simultaneous detection of CH_4 and O_2 is very challenging for planets akin to modern Earth, owing to the comparatively low abundance of CH_4 and the manifestation of strong spectral features at widely separated wavelengths.

A number of metrics have been proposed to quantify the extent of thermodynamic disequilibrium. For example, Simoncini et al. (2013) proposed that the central quantity of interest is not the "distance" from equilibrium but the amount of power required to sustain thermodynamic disequilibrium. Let us consider the following toy reaction:

$$\mathrm{A} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{B}, \tag{6.56}$$

where k_f and k_r are the rate constants for the forward and backward reactions, respectively. The total flux of species A produced by the driving process is denoted by $\mathcal{J}_\mathrm{A}$ (in units of particles/s), and the equilibrium constant is given by $K_\mathrm{eq} = k_f/k_r$. Simoncini et al. showed that the power (P_drive) required by the driving process to maintain the system in disequilibrium is given by

$$P_\mathrm{drive} = -\mathcal{J}_\mathrm{A} k_B T \ln\left(\frac{p - p_A}{K_\mathrm{eq} p_A}\right), \tag{6.57}$$

where T is the ambient temperature and p_A and p are the partial pressure of species A and the total pressure, respectively.

A different approach was adopted in Krissansen-Totton et al. (2016), who suggested that the amount of "available" Gibbs free energy constitutes a good measure of thermodynamic disequilibrium. Hence, the parameter (Φ_{DE}) used for quantifying the extent of disequilibrium was defined as

$$\Phi_{DE} = G_{(\mathrm{T,P})}\left(n_j\right) - G_{(\mathrm{T,P})}\left(\bar{n}_j\right), \tag{6.58}$$

where $G_{(\mathrm{T,P})}\left(n_j\right)$ is the Gibbs free energy at the observed state and $G_{(\mathrm{T,P})}\left(\bar{n}_j\right)$ is the Gibbs free energy at the equilibrium state; Φ_{DE} was calculated in units of J/mol. Note that n_j and $\bar{n}_j$ denote the number of moles of species j observed and in equilibrium, respectively. Krissansen-Totton

et al. determined that $\Phi_{DE} \sim 2.3 \times 10^3$ J/mol for Earth, $\Phi_{DE} \sim 1.4 \times 10^2$ J/mol for Mars, $\Phi_{DE} \sim 1.2$ J/mol for Titan, and $\Phi_{DE} \sim 6 \times 10^{-2}$ J/mol for Venus. As one may have anticipated, the value of Φ_{DE} is highest for Earth. The primary source of thermodynamic disequilibrium for Earth's coupled atmosphere-ocean system in the modern era as well as in the Proterozoic is the simultaneous coexistence of N_2, O_2, and H_2O. This is because these three species should be converted into the stable nitrate ion (NO_3^-) via the formation of nitric acid (HNO_3) as follows:

$$5O_2 + 2N_2 + 2H_2O \rightarrow 4H^+ + 4NO_3^- \tag{6.59}$$

Thus, the simultaneous detection of N_2, O_2, and liquid water might reveal the existence of life (Lammer et al. 2019). When one considers the Archean, prior to the rise in O_2 levels, the chief contributor to Φ_{DE} (~ 70 percent of the total) is probably attributable to the simultaneous coexistence of CO_2, N_2, CH_4, and liquid water (Krissansen-Totton, Olson & Catling 2018). The rationale is that most of the methane should be depleted via the formation of ammonium and bicarbonate ions:

$$5CO_2 + 4N_2 + 3CH_4 + 14H_2O \rightarrow 8NH_4^+ + 8HCO_3^- \tag{6.60}$$

Hence, Krissansen-Totton, Olson and Catling (2018) proposed that the detection of these four compounds, especially CH_4 and CO_2, could be suggestive of biospheres akin to those prevalent on Archean Earth.

Thermodynamic disequilibria also emerge as a consequence of abiotic planetary processes. Hence, further research is necessary in order to determine and sharpen potential metrics for distinguishing between disequilibria generated by biological and abiotic phenomena in the manner of Wogan and Catling (2020), with *context* playing a pivotal role. Another point worth bearing in mind is that quantitative estimates of thermodynamic disequilibria usually presuppose the existence of oceans or metabolic pathways with properties similar to early or modern Earth, but it is not clear whether these characteristics are sufficiently generic.

6.7.2 Biomass surface density calculations

In Section 3.3, we remarked that one of the chief advantages associated with oxygenic photosynthesis is that the electron donor employed in this pathway (H_2O) was abundant on Earth. In contrast, when one considers

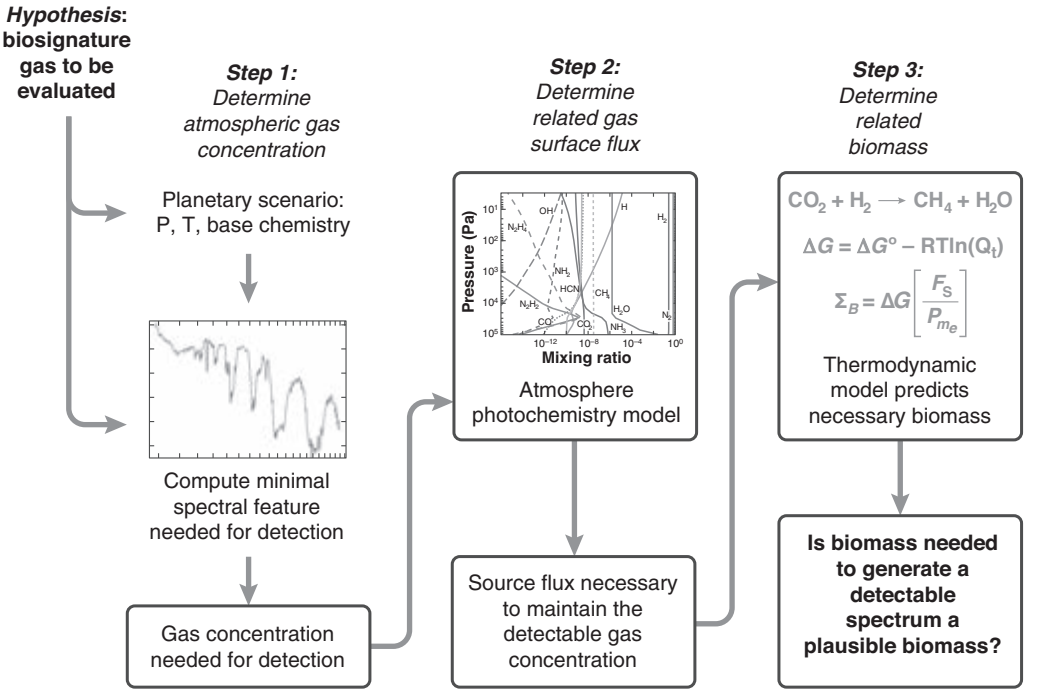


Figure 6.8 The procedure advocated for determining whether the putative concentrations of certain biosignature gases are compatible with biogenic origins. (© The American Astronomical Society. *Source:* S. Seager, W. Bains, and R. Hu [2013], A biomass-based model to estimate the plausibility of exoplanet biosignature gases, *Astrophysical Journal* 775[2]: 104, fig. 1.)

anoxygenic photosynthesis, the corresponding electron donors were far less abundant. In turn, this geochemical constraint may have restricted the rates of carbon fixation by orders of magnitude relative to present-day Earth; as a corollary, the biomass density of anoxygenic photoautotrophs would also have been lower. In principle, therefore, knowing the flux of electron donors (or photons) enables us to infer an upper bound on the biomass density and thereby assess the detectability of surface biosignatures.

Moving on to potential gaseous biosignatures, an ingenious line of reasoning was advanced by Seager et al. (2013) to infer the biomass density on the basis of the measured concentration of certain gases. The basic steps involved in this methodology are depicted in Figure 6.8. There are three components in total. First, the atmospheric concentration of a particular biosignature gas required for detectability is computed from the spectral

features. Second, the required influx of this gas from the planetary surface to achieve the desired atmospheric concentration is calculated from atmospheric photochemistry models. The last stage involves determining the amount of biomass required to generate this flux and subsequently gauging whether this biomass estimate is plausible or unrealistic.

Although this strategy has an innate appeal, it cannot be applied to all gases. The method is well suited for studying biogenic gases generated as the products of metabolic reactions reliant on harnessing chemical energy gradients in the environment; these gases were dubbed Type I biosignature gases by Seager et al. (2013). The classic example of a Type I gas is methane that is liberated during methanogenesis, which involves the reduction of dissolved CO_2 by H_2 to yield CH_4, as seen from (6.46). In this pathway, the flux of methane produced is directly correlated with the amount of biomass present. In contrast, the proposed approach does not work well for pathways where the biogenic gases are a by-product of biomass building. For such gases, to leading order, once the required biomass is built, no further activity is expected. Hence, estimating the flux of such gases requires knowledge of the biomass turnover time, which is completely unknown. Hence, unless one enforces a biomass turnover time akin to the Earth, quantitative estimates are not feasible.

Let us now specialize to Type I biosignature gases and adopt the prescription delineated in Seager et al. (2013). The minimum rate of energy required for the maintenance of an organism is denoted by P_{me} and has units of power per unit mass (W kg^{-1}). It is expressible as the product of the Gibbs free energy change (ΔG_r) associated with the energy-yielding reaction and the production rate per unit mass of the metabolic by-product ($\mathcal{R}_m$); observe that $\mathcal{R}_m$ and ΔG_r have units of mol kg^{-1} s^{-1} and J mol^{-1}, respectively. Thus, we end up with

$$P_{me} \sim \Delta G_r \mathcal{R}_m, \tag{6.61}$$

provided that ΔG_r is entirely devoted to organismal maintenance. The production rate $\mathcal{R}_m$ is further given by

$$\mathcal{R}_m \sim \frac{F_{\mathrm{source}}}{\Sigma_B}, \tag{6.62}$$

where F_{source} (units of mol m^{-2} s^{-1}) represents the source flux of the gaseous biosignature, and Σ_B is the biomass density (units of $\mathrm{kg\,m}^{-2}$). Lastly,

we note that P_{me} is given by the Arrhenius relationship,

$$P_{me} = A_{me} \exp\left(-\frac{E_a}{k_B T_r}\right), \tag{6.63}$$

where A_{me} is analogous to the Arrhenius rate constant, and E_a is the activation energy at the reaction temperature T_r. It should be recognized that ΔG_r is dependent on the temperature at which the energy-yielding reaction takes place as well as the measured concentrations of reactants and products. It is known that $A_{me} \approx 3.9 \times 10^{13}$ W kg^{-1} for aerobic growth, $A_{me} \approx 2.2 \times 10^{13}$ W kg^{-1} for anaerobic growth, and $E_a \approx 1.15 \times 10^{-19}$ J. Hence, by simplifying the above three equations, we obtain

$$\Sigma_B \sim \Delta G_r \left(\frac{F_{\text{source}}}{A_{me}}\right) \exp\left(\frac{E_a}{k_B T_r}\right). \tag{6.64}$$

Sensu Seager et al., we will tackle a couple of illustrative examples to highlight how this formalism works. We caution that both of these cases are heuristic, and the ensuing results are by no means definitive.

The concentration of Martian methane exhibits spatial and temporal variations (Yung et al. 2018), with a measured spike of ~ 15 ppb near the Gale crater (Giuranna et al. 2019). Several studies have argued in favor of a subsurface Martian biosphere situated a few kilometers beneath the surface that might serve as the biogenic source for the detected methane spikes. Let us consider $T_r \sim 265$ K, which corresponds roughly to the maximum daytime temperature observed near the Gale crater. Substituting this value into (6.63) and utilizing A_{me} for anaerobic growth, we obtain $P_{me} \approx 0.5$ W kg^{-1}. The Gibbs free energy based on the putative abundances of CO_2, CH_4, H_2O, and H_2 is $\Delta G_r \sim 8 \times 10^4$ J mol^{-1}. Lastly, the surface flux of CH_4 consistent with the measured atmospheric concentration is chosen to be $F_{\text{source}} \sim 1.7 \times 10^{-15}$ mol m^{-2} s^{-1}. By plugging all of these values into (6.64), we obtain a biomass density of $\Sigma_B \sim 2.7 \times 10^{-10}$ kg m^{-2}.

The value of Σ_B calculated via this simple model is much lower than the biomass densities documented for typical microbial biofilms present on the surface of Earth ($\gtrsim 10^{-4}$ kg m^{-2}). Moreover, it may be commensurate with the biomass density of microbial communities located in Earth's subterranean biospheres. Hence, as per this framework, the maximal observed abundance of atmospheric methane is consistent with the existence of a

(sub)surface microbial biosphere on Mars. Yet, it is equally vital to appreciate that compatibility does not amount to proof, since the latter is only possible through meticulous in situ observations of the Martian surface and atmosphere allied to precise laboratory experiments.

The next scenario we contemplate is Titan, the only moon in our Solar system with a substantial atmosphere (Hörst 2017). Some models theorized that acetylene (C_2H_2) is absent from Titan's atmosphere because it is consumed by microbes that rely on the following reactions:

$$C_2H_2 + 2H_2 \rightarrow C_2H_6 \tag{6.65}$$

$$C_2H_2 + 3H_2 \rightarrow 2CH_4 \tag{6.66}$$

Let us suppose that these reactions unfold in the lakes of Titan on the surface comprising liquid methane and ethane at $T_r \sim 100$ K. Now, we must make a leap of faith and utilize the values of E_a and A_{me} introduced earlier; in general, this assumption is not guaranteed to be valid for alternative biochemistries. By proceeding along these lines, we arrive at $P_{me} \sim 1.4 \times 10^{-23}$ W kg^{-1}. Next, for the measured chemical abundances and at $T_r \sim 100$ K, the Gibbs free energy for (6.65) is $\Delta G_r \sim 3 \times 10^5$ J mol^{-1}. Lastly, the source flux is determined via numerical modeling and is given by $F_{\text{source}} \sim 3.3 \times 10^{-10}$ mol m^{-2} s^{-1}. Hence, after making use of (6.64), we obtain a biomass density of $\Sigma_B \sim 7.1 \times 10^{22}$ kg m^{-2}. If the same calculation is repeated for (6.66) using $\Delta G_r \sim 3.85 \times 10^5$ J mol^{-1}, it is found that $\Sigma_B \sim 9.1 \times 10^{22}$ kg m^{-2}.

In comparison, the maximum biomass densities observed on Earth are ~ 100 kg m^{-2}, which are clearly many orders of magnitude removed from the above estimates for Σ_B. In other words, modulo the many assumptions involved, the model predicts that the putative drawdown of acetylene on Titan is unlikely to have a biogenic origin. It is crucial to appreciate that this model does not rule out the prospects for life on Titan altogether. In fact, it might not even rule out the existence of this particular metabolic pathway, for two reasons. First, as noted previously, we were reliant on Earth-centric values of E_a and A_{me}. Second, if the same model is invoked for an ammonia-water brine at $T_r \sim 176$ K, the biomass density drops by many orders of magnitude ($\Sigma_B \sim 2 \times 10^4$ kg m^{-2}), although it is still much higher than what is observed on Earth.

What the examples of Mars and Titan do illustrate, however, is that this theoretical model yields a wide spectrum of values for the biomass density,

of which only some will be compatible with the Earth. Hence, calculating the predicted biomass densities for a given atmospheric composition and checking whether they are commensurate with those observed on Earth is a potentially valuable diagnostic for distinguishing between real biosignatures and false positives.

6.7.3 Chemical reaction networks

Networks are studied mathematically by mapping them to graphs. In rudimentary terms, the nodes of a graph correspond to the entities under question, whereas the edges represent the interactions; in a chemical reaction network (CRN), the entities are chemical species and the edges signify chemical reactions among species. The resulting network can be subjected to topological analysis to determine key properties such as the mean degree (average number of edges per node). One of the most notable among them is the degree distribution $P(k)$ that quantifies the fraction of nodes that have k edges. Network analysis has been successfully applied to a number of complex systems ranging from social networks and the World Wide Web to paleontology and biochemistry (Boccaletti et al. 2006).

Metabolic networks have been studied via this formalism, and they display several interesting features. For starters, they belong to the category of scale-free networks—namely, graphs characterized by $P(k) \propto k^{-\upsilon}$, whereas simple random graphs tend to obey the binomial distribution (Albert & Barabási 2002). The study by H. Kim et al. (2019) revealed that biochemical networks manifest universal scaling laws across three different levels: individuals, ecosystems, and the biosphere; the scaling parameter was determined to be the number of biochemical compounds. It is conceivable that this universality, which is potentially attributable to the hierarchical organization of evolutionary units, comprises a valuable biosignature for in situ life-detection experiments.

However, in the context of remote-sensing studies, of more relevance is the possibility that the aforementioned topological properties of biochemical networks are imprinted in the reaction networks of atmospheric chemistry. Solé and Munteanu (2004) carried out a CRN analysis of the atmospheres of Venus, Mars, Titan, Earth, and the giant planets. It was found that the Earth's atmospheric CRN exhibited a hierarchical organization reminiscent of biochemical networks. Solé and Munteanu also found that

the degree distribution was best fitted by a power law with $\upsilon \approx 2.16$ for the Earth, whereas it displayed exponential falloff for the other worlds. Thus, it is tempting to theorize that network analyses of planetary atmospheres offer a potential means of distinguishing between worlds with and without life.

However, a few caveats are worth highlighting here. First, our knowledge of chemical constituents in atmospheres other than the Earth is comparatively limited, which may introduce observational biases. Second, we do not have a clear understanding yet of how physicochemical factors such as the temperature and atmospheric composition influence the network's topology. In this regard, it is imperative to understand how Earth's atmospheric CRN and its topological characteristics have evolved from the Hadean to the present day. Lastly, given the limited data that will be available from observations, it is not clear whether this data would suffice to distinguish between inhabited worlds and false positives. Despite these potential barriers, the study of CRNs using graph theory represents a promising line of inquiry that merits further research.

6.8 ASSESSING THE PLAUSIBILITY OF LIFE DETECTION

In view of the numerous false positives and other caveats that we have encountered, the need for a systematic framework for determining the plausibility of a given biosignature is self-evident. We will adopt the methodology, notation, and recommendations espoused in Catling et al. (2018) henceforth. Other methodologies, such as signal detection theory and utility theory, also possess their share of merits, especially the former (Pohorille & Sokolowska 2020).

6.8.1 A Bayesian approach

Before embarking on our analysis, introducing some notation is in order. P denotes the probability; H and E stand for "hypothesis is correct" and "evidence," respectively; by $\bar{H}$, we refer to the statement that "hypothesis is false." As per Bayes' theorem, one obtains

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E)}. \tag{6.67}$$

$P(H|E)$ quantifies the likelihood of a particular hypothesis being true given certain evidence, whereas $P(E|H)$ is the likelihood of the evidence being manifested provided that the hypothesis is correct. $P(H)$ is known as the *prior* and signifies the likelihood of the hypothesis being true. The marginal likelihood $P(E)$ can be simplified further for a hypothesis that admits binary values—that is, true or false. The above equation can be rewritten as

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E|H)\,P(H) + P\left(E|\bar{H}\right)P(\bar{H})}, \tag{6.68}$$

wherein $P\left(E|\bar{H}\right)$ is the likelihood of the evidence occurring when the hypothesis is false.

We are now in a position to utilize (6.68) appropriately. The probability that we seek to determine is $P(L|D, C)$. This measures the likelihood for the specified hypothesis of life (L) existing on a world given observational data (D) that may comprise biosignatures and contextual information (C) pertaining to this world; to calculate this probability, we introduce the notation $\bar{L}$ for the hypothesis of "no life." By employing (6.68), we have

$$P(L|D, C) = \frac{P(D, C|L)\,P(L)}{P(D, C|L)\,P(L) + P\left(D, C|\bar{L}\right)P(\bar{L})}, \tag{6.69}$$

and this can be further simplified by making use of (6.67) along with the axioms of conditional probability to yield

$$P(L|D, C) = \frac{P(D|C, L)\,P(C|L)\,P(L)}{P(D|C, L)\,P(C|L)\,P(L) + P\left(D|C, \bar{L}\right)P\left(C|\bar{L}\right)P(\bar{L})}. \tag{6.70}$$

Next, we simplify $P(C|L)$ and $P\left(C|\bar{L}\right)$ by using (6.67). Hence, we end up with the final expression:

$$P(L|D, C) = \frac{P(D|C, L)\,P(L|C)}{P(D|C, L)\,P(L|C) + P\left(D|C, \bar{L}\right)P\left(\bar{L}|C\right)} \tag{6.71}$$

To recap, $P(D|C, L)$ signifies the likelihood of the data being manifested in a given context provided that life exists. More importantly, $P\left(D|C, \bar{L}\right)$ quantifies the likelihood of the data occurring in the absence of life for a given context, and it therefore quantifies the false positive contribution (in

the limit of minimal instrumental noise). Lastly, $P(L|C)$ and $P\left(\bar{L}|C\right)$ are the prior estimates for the presence and absence of life given the contextual information C. As L and $\bar{L}$ are mutually exclusive, we have

$$P\left(\bar{L}|C\right) = 1 - P(L|C)\,. \tag{6.72}$$

It is very hard at this juncture to quantify $P(L|C)$ because our understanding of abiogenesis and subsequent major evolutionary innovations is limited. Clearly, addressing this issue will necessitate synthesizing models and data from fields as diverse as biochemistry, computer science, theoretical physics, and geology. The reader is referred to Chapters 2 and 3 along with the overview by S. I. Walker et al. (2018), wherein potential approaches for constraining $P(L|C)$ are outlined.

At this juncture, it is instructive to introduce the following ratio,

$$\mathcal{D}_\ell = \frac{P(D|C, L)}{P\left(D|C, \bar{L}\right)}, \tag{6.73}$$

which shall be referred to henceforth as the detectability ($\mathcal{D}_\ell$). This ratio is important primarily because it inherently epitomizes the relative likelihood that a given signal (i.e., data) is due to biogenic causes. In this chapter, we have indirectly addressed several approaches for identifying false positives and determining the robustness of the putative biosignatures; see Section 6.7 in particular. Given the data available, one or more of these methods may be employed to estimate $\mathcal{D}_\ell$. By employing the above definition for $\mathcal{D}_\ell$, we find that (6.71) simplifies to

$$P(L|D, C) = \frac{P(L|C)\,\mathcal{D}_\ell}{1 + P(L|C)\,(\mathcal{D}_\ell - 1)}, \tag{6.74}$$

after making use of (6.72).

It is helpful to evaluate a few limiting cases to understand how $P(L|D, C)$ varies with $\mathcal{D}_\ell$ and $P(L|C)$. In the regimes $\mathcal{D}_\ell \to 0$ or $P(L|C) \to 0$ (provided $\mathcal{D}_\ell$ is not much greater than unity), we see that $P(L|D, C) \to 0$. This behavior is along expected lines because extremely low values of the detectability or the prior for life's emergence will greatly diminish the chances of rigorously establishing the presence of life on a given

world. Similarly, in the idealized limits of $P(L|C) \to 1$ and $\mathcal{D}_\ell \to \infty$ (provided $P(L|C)$ is not much smaller than unity), it is easy to verify that $P(L|D, C) \to 1$. Intuitively, a high detectability or prior for life's emergence should translate to an enhanced likelihood of detecting life.

Next, let us consider the case where $\mathcal{D}_\ell \sim 1$. This corresponds to the scenario in which the detectability is neither particularly high nor particularly low. We find that (6.74) reduces to

$$P(L|D, C) \sim P(L|C) . \tag{6.75}$$

As a result, in this regime, the posterior probability will be dependent only on the prior for emergence of life. For the sake of completeness, we should consider the case $P(L|C) \approx 0.5$. If $P(L|C)$ was a uniform distribution, then $P(L|C) \approx 0.5$ would correspond to its mean value. In this event, we end up with

$$P(L|D, C) \approx \frac{\mathcal{D}_\ell}{1 + \mathcal{D}_\ell}, \tag{6.76}$$

implying that the posterior probability will be solely dictated by the detectability in this instance.

We shall now turn our attention to the contextual information C. In principle, any number of factors ranging from basic physical parameters such as planetary mass, radius, and semimajor axis to the properties of the host star could be included. The formalism introduced herein can also be used to assess $\langle P(L) \rangle$, i.e., the weighted likelihood of inhabited exoplanets across the Galaxy:

$$\langle P(L) \rangle = \sum_C P(L|C)\, P(C) \tag{6.77}$$

To see how this function works, let us consider the term f_l from the famous Drake equation (introduced in Section 8.1). In most studies, f_l is viewed as the probability for the emergence of life on planets with habitable conditions. For the sake of simplicity, we will consider only F-, G-, K-, and M-type stars, although there is a strong case to be made for the inclusion of A-type stars as well. Using the above equation, we obtain

$$\begin{aligned} f_l = {} & P\left(L|M_{hp}\right) P\left(M_{hp}\right) + P\left(L|K_{hp}\right) P\left(K_{hp}\right) \\ & + P\left(L|G_{hp}\right) P\left(G_{hp}\right) + P\left(F|M_{hp}\right) P\left(F_{hp}\right), \end{aligned} \tag{6.78}$$

where $P\left(L|M_{hp}\right)$ represents the probability of life's occurrence given the contextual information of a habitable world orbiting an M-dwarf, and $P\left(M_{hp}\right)$ denotes the likelihood of habitable planets around M-type stars; similar definitions are applicable to F-, G-, and K-type stars.

Before moving on, note that the Bayesian approach is summarized in Figure 6.9, which makes apparent that simply relying on observational data or theoretical models alone will not suffice to accurately determine the posterior probability $P\left(L|D,C\right)$ that constitutes the central function of interest. For the interested reader, an explicit example of how the Bayesian framework outlined thus far facilitates the identification of biogenic O_2 is presented in Walker et al. (2018).

6.8.2 Input components: Desiderata

A number of factors can be gleaned from observations, and they will either serve as contextual information (C) or provide us with the requisite data (D) for assessing whether a given world is inhabited.

The first set of observations that must be taken into account pertain to the host star and include its mass, age, rotation rate, and spectral energy distribution. We have already seen in Chapter 4 how these parameters influence habitability. Furthermore, they give rise to several false positives, as outlined in this chapter. One such example is the potential abiotic production of O_2, either through the UV-mediated photolysis of H_2O or CO_2. In Section 6.3.4 we briefly indicated how and why stellar parameters are derivable through observations. Stellar spectroscopy might also enable us to gauge the composition of planets and the abundances of certain bioessential elements (e.g., phosphorus) in their crusts. Lastly, the presence of stellar binaries influences habitability in myriad ways ranging from the boundaries of the habitable zone to the dynamic stability of exoplanets orbiting them (Z. Wang & Cuntz 2019).

A number of external physical parameters are worth knowing about the planet, which are discernible through observations. The radius and mass, collectively determinable through transit and radial velocity methods, respectively, enable us to calculate the bulk density and thereby constrain the composition of the planet. The rotation rate and obliquity can be inferred via direct imaging (Section 6.2.1) and inform us in several respects about the potential habitability of exoplanets. As we saw in Section 6.3.2,

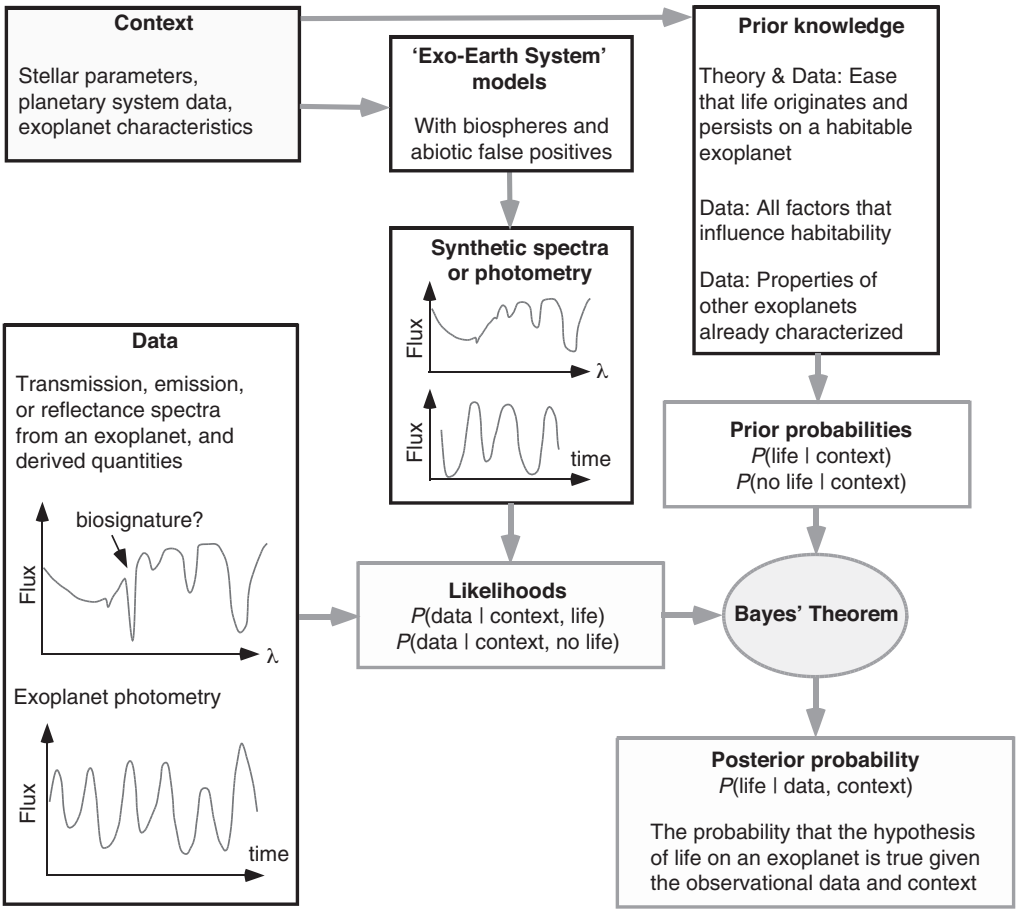


Figure 6.9 Data and modeling necessary to compute the desired probability $P(\text{life}|D, C)$. Note that the context (C) and data (D) are determined through observations, but the likelihoods and prior probabilities require inputs from theoretical models. (*Source:* David C. Catling, Joshua Krissansen-Totten, Nancy Y. Kiang, David Crisp, Tyler D. Robinson, Shiladitya DasSarma, Andrew J. Rushby, Antholy Del Genio, William Bains, and Shawn Domagal-Goldman [2018], Exoplanet biosignatures: A framework for their assessment, *Astrobiology* 18[6]: 709–738, fig. 1.)

exoplanetary architecture is important for estimating long-term habitability and requires observations of planetary companions such as gas giants. Apart from these external parameters, a multitude of internal parameters are possibly measurable.

In this context, observational techniques such as occultation spectroscopy have been proposed as a means of quantifying the surface temperature (Section 6.1.3). Knowing the surface temperature is important from the twin standpoints of ensuring the presence of liquid water and staying within the thermal limits of life. The surface pressure might also be measurable under some circumstances; it would provide further insight into the possibility of liquid water and the dayside-nightside thermal gradient on tidally locked exoplanets, as shown in Chapter 5. Oceans of liquid water are discernible through the glint effect delineated in Section 6.2.1, and their confirmation would increase the prospects for life. Hazes are ostensibly easier to identify than biosignature gases because of their broad spectral features and the low spectral resolution thus required; see Section 6.4.7.

The final, and arguably the most important, input required for the Bayesian approach is the prospective biosignatures themselves. We have already examined this topic in detail, owing to which only a couple of general comments are in order. It is necessary to combine the search for gaseous biosignatures—the most commonly studied category—with surface and temporal biosignatures to increase (1) the chances of detecting a potential biogenic signal and (2) the likelihood that any putative signal thus detected is not a false positive. A good example to illustrate how this strategy would function is the detection of molecular O_2 (gaseous biosignature) and the VRE (surface biosignature), because both of them are consequences of oxygenic photosynthesis.

6.8.3 Classification of life detection

By drawing on the classification scheme employed by the Intergovernmental Panel on Climate Change (IPCC) to describe global warming, Catling et al. (2018) proposed a five-tier scheme to qualitatively assess the confidence levels associated with life detection on a given world. The schema is based on the posterior probability $P(L|D, C)$ that ranges between 0 and 1. An overview of this classification is presented in Table 6.1. It is, however, important to recognize that the body of evidence and the illustrative examples

Table 6.1 Possible categories for probability of life detection on individual exoplanets

Confidence level for detection of life	Posterior probability $P(L\|D, C)$	Evidence	Suggestive but purely illustrative examples
Level 1: very likely inhabited	90–100%	Multiple lines of evidence for life. Given current understanding of planetary processes, no known abiotic process can plausibly explain all observed features.	An O_2-rich atmosphere with other biosignature gases, including CH_4 and N_2O, and a liquid ocean identified on an Earth-sized exoplanet in the HZ.
Level 2: likely inhabited	66–100%	The body of evidence is consistent with the presence of life.	Atmospheric O_2 detected together with CO_2 and water vapor on an exoplanet in the HZ.
Level 3: about as likely as not inhabited (inconclusive)	33–66%	Some evidence for life, but insufficient contextual information to draw a definitive conclusion because plausible alternative abiotic explanations cannot be ruled out.	O_2 detection in isolation; or an organic haze with abundant CH_4; or pigment-like biosignatures; or N_2-CO_2 atmosphere. Circumstantial evidence for liquid water on a planet in the conventional HZ.
Level 4: likely uninhabited	0–33%	Observational evidence that the planet is habitable, but no biosignatures detected despite an exhaustive search.	Planet is in the HZ and has an atmosphere with abundant water vapor features. But no biosignatures are detected despite extensive data.
Level 5: very likely uninhabited	0–10%	Criteria for habitability are not met or atmospheric antibiosignatures are detected.	CO_2-rich, desiccated planets; or CO_2-H_2 antibiosignature atmosphere; or abundant CO antibiosignature.

Notes: This classification scheme qualitatively captures the likelihood of life detection on a given world given the calculated posterior probability; the examples listed are based on extrapolations from life-as-we-know-it on Earth. *Data source:* Reformatted from David C. Catling, Joshua Krissansen-Totten, Nancy Y. Kiang, David Crisp, Tyler D. Robinson, Shiladitya DasSarma, Andrew J. Rushby, Antholy Del Genio, William Bains, and Shawn Domagal-Goldman [2018], Exoplanet biosignatures: A framework for their assessment, *Astrobiology* 18[6]: 709–738, Table 6.)

provided reflect our current knowledge and should therefore not be viewed as definitive.

When the maximum value for $P(L|D, C)$ drops below 33 percent, it is rather unlikely for the planet to be inhabited and maintain compatibility with the data and the context. Hence, from the perspective of science policy, caution is especially necessary when making statements about level 4 and level 5 worlds that exhibit $P(L|D, C) < 0.33$. Although level 5 worlds, in particular, come across as being dispensable from the standpoint of finding extraterrestrial life, in actuality they play a very important role by allowing us to potentially infer the preponderance or scarcity of life in the Universe.

6.9 CONCLUSION

Imagination! who can sing thy force?
Or who describe the swiftness of thy course?
. . . . . .
We on thy pinions can surpass the wind,
And leave the rolling universe behind:
From star to star the mental optics rove,
Measure the skies, and range the realms above.
There in one view we grasp the mighty whole,
Or with new worlds amaze th' unbounded soul.

—Phillis Wheatley, *On Imagination*

Niche construction, broadly interpreted, has played a major role in shaping Earth's environment across space and time. Given its ubiquity and centrality, it is reasonable to expect that other inhabited worlds would also exhibit the same phenomenon. As a consequence of this active modification of the planetary environment, a number of biosignatures are expected to arise. We have covered most of the primary candidates by classifying them as gaseous, surface and temporal biosignatures. Nearly all of the potential biosignatures have associated false positives and negatives; the former are particularly problematic since they may lead to erroneous conclusions about the existence of life on a given world, even though it is actually devoid of biological activity.

A central focus of this chapter was delineating the various observational methods that have sprung up to identify these biosignatures. Understanding this aspect is evidently of paramount importance, since no amount of theorizing in isolation can definitively establish or exclude the presence of extraterrestrial life on other worlds. Yet, at the same time, we cannot do

away with theory altogether as it plays a key role in estimating the likelihood of false positives, the selection of suitable target planets for follow-up observations, and much more. This marriage between observations and theory is exemplified in the Bayesian framework described in Section 6.8; see Figure 6.9 for more details.

It has become something of a cliché for scientific papers and outreach activities to claim that this is the first time in human history when we can look up at the stars and possess the requisite tools to answer the age-old question, "Are we alone?" This optimism stems from the fact that upcoming telescopes such as the James Webb Space Telescope (JWST) and ground-based Extremely Large Telescopes (ELTs) may be capable of carrying out transit spectroscopy of a few select planets (e.g., TRAPPIST-1 system) and seeking gaseous biosignatures such as O_2. Hence, the claim that we find ourselves on the cusp of entering a new era of exoplanetary observations is no mere hyperbole (Kiang et al. 2018).

Yet, we caution that this optimism should be balanced with a healthy dose of realism. The number of target planets that are amenable to detailed searches for biosignature will be few. Furthermore, many of them would orbit M-dwarfs, which may not be very conducive to habitability for reasons described in Chapter 4. Finally, as emphasized throughout this chapter, virtually none of the current biosignatures can be attributed to biological activity with ~ 100 percent certainty, owing to the complicating effect of false positives. It is for this reason that our posterior probability of detecting life given the data and context will lie anywhere between 0 and 100 percent, as explained in Section 6.8.

In summary, setting aside the fundamental question of whether we will actually find extraterrestrial life through remote-sensing observations, there remains little doubt that the next few decades are poised to spark intense research powered by the access to unprecedented amounts of observational data from multifarious telescopes. Hence, irrespective of the eventual outcome, it is not much of an exaggeration to state that we are truly on the cusp of a pivotal epoch in human history. To put it more simply, borrowing from another cliché, the bends, forks, and cul-de-sacs in our journey toward discovering markers of life are nearly as important as the final destination—namely, the detection of life, itself.

Chapter 7

LIFE IN SUBSURFACE OCEANS

> Free man, you will always cherish the sea! The sea is your mirror. In the infinite roll of a wave you contemplate your soul. . . . O sea, no one knows the riches of your keep, so zealously you guard your secrets!
>
> —Charles Baudelaire, *L'homme et la mer*

The Earth is the only world currently known to host life beyond question, although this statement may change in the near future. Thus, from our anthropocentric perspective, our vision of inhabited worlds typically conjures up planets and moons akin to that of the Earth—that is, orbiting a star and endowed with oceans and continents on the surface. In quantitative terms, this notion is captured by the concept of the circumstellar habitable zone (HZ)—the domain around a host star that is capable of sustaining liquid water on the surface of a world (either planet or moon) for a given atmospheric composition. The extent of the HZ varies depending on the characteristics of the host star (e.g., mass) and several planetary parameters such as the atmospheric composition, rotation rate, and so forth. Nevertheless, there is one point of commonality in most analyses of the HZ—namely, they are concerned with the presence of liquid water on the *surface*.

Let us, instead, take a step back and recall a physical phenomenon that we have all observed. Water, whether it be situated in lakes and ponds or in the refrigerator, tends to freeze from the top. It is well-known that this phenomenon is a consequence of the unusual dependence of its density on temperature. Although the density of water increases as it cools along the expected trend, this behavior is reversed at 277 K (4 °C), and its density

actually decreases with a further decrease in temperature. Eventually, this leads to the formation of ice, which is less dense compared to water. As a result, even when the upper layers of lakes are frozen over, their floor may comprise liquid water (at 277 K). We can, therefore, ask ourselves whether a similar phenomenon could occur on a much larger scale, i.e., at the level of planets and moons.

In other words, are there worlds with surface ice envelopes and subsurface oceans of liquid water? The answer is yes, and our own Solar system comprises at least three worlds confirmed to possess subsurface oceans. The first is Europa, one of Jupiter's Galilean moons (discovered by Galileo Galilei), whereas the next two are the moons of Saturn: Titan and Enceladus. In addition to worlds akin to this trio from our Solar system that may have subsurface oceans, it is important to recognize that these worlds need not be bound to a host star. They could have been ejected from their sites of formation and would thus be traversing the vast void of interstellar space; such worlds are often referred to as *free-floating planets*.

This leads us to a more fundamental question: Are such worlds habitable? To put it differently, do they possess the conditions necessary for life to originate and evolve? Clearly, at this stage, we cannot hope to present a definitive answer, given that we do not know the necessary and sufficient conditions for life (even for life-as-we-know-it) to arise. Moreover, our biological knowledge of the deepest reaches of Earth's oceans, only a tiny fraction of which have been explored, is woefully incomplete. To paraphrase the preceding quotation from Charles Baudelaire, the deep sea is replete with "hidden riches" that remain mysteries to us. Yet, even with our limited understanding, there is much that we can deduce about these worlds. In this chapter, we will examine the internal structure of these worlds, the energy and nutrient sources available for powering putative ecosystems, and the number of such worlds that are likely to exist within our Galaxy.[1] Aside from this broad category of worlds, another class deserves a mention in the same spirit: objects with deep terrestrial biospheres comprising organisms dwelling in rocky environments. The reader is invited to peruse Gold (1992), Magnabosco et al. (2018) and Lingam and Loeb (2020h) for in-depth analyses of this subject.

1. Many, albeit not all, of the topics covered in this chapter are elucidated in a couple of our publications (Lingam & Loeb 2018c, 2019g).

Table 7.1 The physical parameters for moons and dwarf planets that are known or suspected to have subsurface oceans

Name	R (in $R_\oplus$)	M (in $M_\oplus$)	H (in km)	$\mathcal{H}$ (in km)	Status
Europa	0.245	8×10^{-3}	$\sim$ 1–10	~ 100	Moon of Jupiter
Enceladus	0.04	1.8×10^{-5}	$\sim$ 5–40	$\sim$ 20–30	Moon of Saturn
Titan	0.404	2.25×10^{-2}	$\sim$ 55–80	Unknown	Moon of Saturn
Ganymede	0.413	2.48×10^{-2}	$\sim$ 150–250	$\lesssim 200$	Moon of Jupiter
Callisto	0.378	1.8×10^{-2}	$\gtrsim 100$	$\lesssim 200$	Moon of Jupiter
Mimas	0.03	6.28×10^{-6}	Unknown	Unknown	Moon of Saturn
Tethys	0.083	1.03×10^{-4}	Unknown	Unknown	Moon of Saturn
Dione	0.088	1.83×10^{-4}	Unknown	Unknown	Moon of Saturn
Triton	0.212	3.59×10^{-3}	Unknown	Unknown	Moon of Neptune
Charon	0.095	2.66×10^{-4}	Unknown	Unknown	Moon of Pluto
Pluto	0.187	2.18×10^{-3}	Unknown	Unknown	Dwarf planet
Ceres	0.074	1.57×10^{-4}	Unknown	Unknown	Dwarf planet

Notes: R and M denote the mean radius and mass of the subsurface ocean world, while H and $\mathcal{H}$ represent the globally averaged thickness of the ice shell and depth of the underlying ocean, respectively. The values of H and $\mathcal{H}$ are either not sufficiently constrained or altogether unknown and should therefore be regarded as rough estimates. $R_\oplus = 6.37 \times 10^6$ km and $M_\oplus = 5.97 \times 10^{24}$ kg are the radius and mass of Earth, respectively. Among the worlds in this list, only Europa, Enceladus, and Titan can be confidently averred to possess multiple lines of evidence for a subsurface ocean.

7.1 WORLDS WITH SUBSURFACE OCEANS WITHIN OUR SOLAR SYSTEM

Most of the worlds within our Solar system that are either known or presumed to have subsurface oceans are moons as seen in Table 7.1. The study of such worlds represents a highly active area of research and is currently witnessing significant progress by virtue of data garnered from recent space missions. Hence, we briefly summarize some of the essential characteristics of these subsurface ocean worlds. For more details on this subject, the reader may consult Nimmo and Pappalardo (2016), Lunine (2017), and Hendrix et al. (2019). We begin with a description of the three ocean worlds confirmed to have subsurface oceans through multiple lines of evidence—Europa, Enceladus, and Titan—and subsequently mention other candidates.

7.1.1 Europa

Europa, one of Jupiter's four Galilean moons, is about the same size as Earth's moon—that is, about a quarter of the Earth's size. Photographs of Europa from the *Voyager* and *Galileo* missions revealed a terrain that had relatively

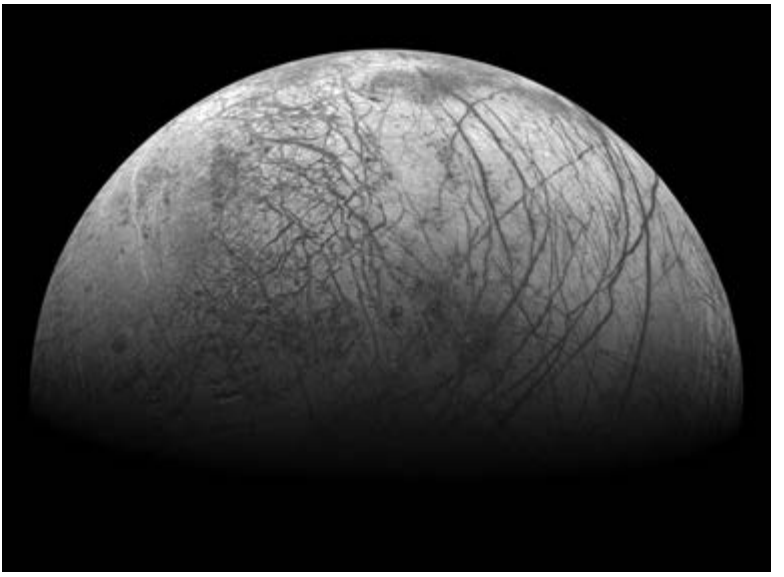


Figure 7.1 A photograph of Europa taken by the Galileo spacecraft in the late 1990s. (*Source:* NASA / JPL-Caltech / SETI Institute.)

few surface craters and was comprised of crisscrossing fractures, as can be seen in Figure 7.1. While not conclusive, the photos show this terrain as consistent with the melting and freezing of ice overlying a liquid ocean. The most unambiguous evidence is, however, not through geology but via magnetic induction.

As per classical electrodynamics, an electrical conductor situated within a time-varying magnetic field will experience an electromotive force, and the induced electrical currents will generate a secondary magnetic field. For the moons of Jupiter, the induced field will vary periodically owing to Jupiter's rotation. By measuring the induction response, it is possible to determine the nature and location of the conductor. The data for Europa revealed that the conductor was located close to the surface—unlike, for example, an iron core—and that the measured response was consistent with the presence of a salty ocean. In addition, there has been recent evidence for the detection of a water plume both remotely via the Hubble Space Telescope (HST) and in situ by the *Galileo* magnetometer. The former is based on finding a transient and compact absorption feature against a uniform UV background due to Jovian scattering, while the latter stems from a brief and intense increase in the plasma density.

The exact magnitudes of the ice envelope and subsurface ocean depths for Europa are not precisely calibrated. With regard to the former, the global thickness of Europa's icy crust is predicted to be ~ 10 km, although there may exist local regions with higher heat flow where the layer declines to $\lesssim 1$ km. Europa's large ocean is situated underneath the icy shell and above its rocky mantle. It seems to possess an average ocean depth of ~ 100 km, implying that the volume of Europa's ocean could be ~ 2 times higher compared to the Earth's oceans.

NASA's *Europa Clipper* mission, which is scheduled for launch in the mid-2020s (probably 2024),[2] will study Europa in more detail during a series of flybys while orbiting Jupiter. The mission is equipped with a pair of mass spectrometers—Mass Spectrometer for Planetary Exploration (MASPEX) and Surface Dust Mass Analyzer (SUDA)—that can sample the plumes (or molecules ejected into space through sputtering) and detect organic molecules (Sephton et al. 2018). In addition, the mission is expected to enable (1) comprehensive imaging and spectroscopy of the surface, (2) detection of active sites ejecting plumes, (3) mapping of the ice envelope's structure, and (4) characterization of the thickness and salinity of the subsurface ocean. On the whole, the *Europa Clipper* mission is anticipated to provide a fairly detailed assessment of Europa's oceanic habitability and pave the way for the launch of a lander in the future.

Another mission scheduled to launch at around the same time (in 2022) is the ESA's Jupiter Icy Moons Explorer (JUICE),[3] which will be the first spacecraft to orbit a moon other than our own. Although the main target of JUICE is Ganymede, and Callisto to a lesser degree, the craft is also expected to conduct two flybys of Europa and may be able to sample the plumes and analyze their composition, as well as study the thickness and structure of the icy crust. JUICE is expected to undertake detailed measurements of the thickness and composition of the crust and underlying ocean for Ganymede and Callisto. Hence, one of its primary objectives is to pave the way toward a deeper understanding of the comparative habitability of Europa, Callisto, and Ganymede.

---

2. NASA Jet Propulsion Laboratory, California Institute of Technology (n.d.), Mission to Europa: Europa Clipper, https://www.jpl.nasa.gov/missions/europa-clipper/

3. European Space Agency (n.d.), JUICE, http://sci.esa.int/juice/

7.1.2 Enceladus

Enceladus is a small moon of Saturn with a radius that is merely 4 percent of the Earth's. Despite its small size, multiple lines of evidence have revealed the presence of a subsurface ocean.

Enceladus is known to possess geyser-like jets that are produced due to cryovolcanism. The *Cassini* spacecraft was able to sample one of these plumes directly and confirmed the presence of water along with sodium salts, silica grains, and other species. The detection of these salts is important since it strongly indicates that the source of water is a subsurface ocean that interacted with silicates, thereby supplying these salts. A second major line of evidence for the subsurface ocean is through measurements of Enceladus's rotation rate. Due to the eccentricity of Enceladus's orbit, its speed also varies and the tidal bulge oscillates backward and forward; this rocking motion goes by the name of *libration in longitude*. For worlds with subsurface oceans, the ice shell is decoupled from the interior, and this results in a higher libration amplitude compared to that of a solid body. This discrepancy has been unambiguously documented for Enceladus, thus providing further evidence for the presence of a subsurface ocean.

The thickness of the ice envelope on Enceladus is not accurately known, but it appears to be within the range of 5–40 km; the most likely value may be $\sim$20 km. Likewise, the depth of the underlying subsurface ocean is not well constrained, but it seems to be around 20 to 30 km, based on the libration measurements and a simple ocean model. An important point worth emphasizing here is that both Enceladus and Europa are predicted to have subsurface oceans in contact with silicate mantles. Many authors perceive the existence of water-rock interactions and hydrothermal vents to be of paramount importance for driving the origin of life (e.g., W. Martin et al. 2008), for reasons explicated in Section 2.7.1.

In the case of Enceladus, the recent detection of molecular hydrogen (H_2) in the plume by means of the *Cassini* mass spectrometer (Waite et al. 2017) was argued to constitute strong evidence in favor of the existence of hydrothermal vents, which are regarded as one of the prime candidates for the sites of abiogenesis. The case for hydrothermal vents was strengthened by the discovery of complex organics with masses greater than 200 atomic mass units (Postberg et al. 2018), since these molecules could have been synthesized at hydrothermal vents in principle. Figure 7.2 depicts the potential

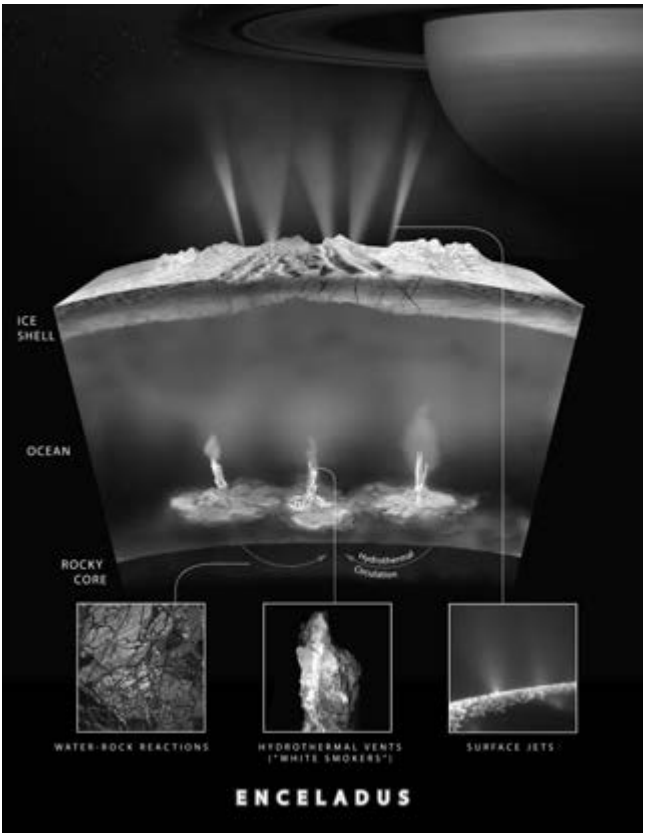


Figure 7.2 An artist's rendering of Enceladus comprising (1) the inferred hydrothermal vent systems on the seafloor, (2) the subsurface ocean and overlaying ice envelope, and (3) the plumes observed by the *Cassini* spacecraft. (*Source:* NASA / JPL-Caltech / Southwest Research Institute.)

internal structure of Enceladus in tandem with its famous plumes surveyed by the *Cassini* instruments.

As the *Cassini-Huygens* mission was completed only recently, there are currently no future missions formally scheduled for Enceladus. However, a couple of recent proposals that were rejected include Enceladus Life Finder (ELF) and Life Investigation For Enceladus (LIFE). Of the two, LIFE was potentially the more promising of the duo since it entailed the return of plume materials for further study on Earth in conjunction with limited in situ analysis. On the other hand, the mission concept for ELF was similar to *Cassini* in scope, as it would have carried out in situ investigations

of the plumes to determine ocean properties (e.g., salinity) and search for molecular biosignatures. While both of these proposals have their own (de)merits, it remains unclear at this stage whether either of them will be selected in the near future. However, in principle, it is conceivable that private organizations, such as the Breakthrough Initiatives,[4] could step in to fill this gap.

7.1.3 Titan

Titan is the second-largest moon in our Solar system, with a radius of 40 percent that of the Earth. Titan is noteworthy for being the only moon in the Solar system with a dense atmosphere, whose surface pressure is higher than the corresponding value on Earth. It is possibly unique among the ocean worlds in another respect as it possesses liquid bodies on both the surface and the subsurface. While the subsurface ocean consists of water, the *Huygens* probe revealed that the surface of Titan has a complex landscape including lakes and seas. The bulk composition of many of these liquid bodies is methane, although some quantities of ethane and nitrogen are also present (Hayes 2016).

The first line of evidence is based on the Huygens probe's discovery of Extremely Low Frequency (ELF) radio waves in Titan's atmosphere. It was observed that the atmospheric electric field did not vanish at the surface and that the ambient electric field and the existence of the ELF waves were best explained by the presence of a conductor beneath the surface; the required conductivity was found to be consistent with that of a salty ocean, perhaps similar to Earth's Dead Sea. Another major line of evidence for Titan's subsurface ocean stems from its eccentric orbit, which generates a time-varying tidal potential. If the ice shell and the interior are separated by a liquid ocean, the tidal response—changes in the gravitational field; lateral and radial surface displacements—is predicted to be higher compared to the solid-body case. Measurements of changes in Titan's gravitational field revealed perturbations in its shape that could not be explained by a solid interior, which therefore necessitated decoupling the ice crust from the core.

Very little is reliably known about both the location and volume of Titan's subsurface ocean. Estimates based on the wave measurements suggest that the ocean may lie 55 to 80 km beneath the surface. On account

4. Breakthrough Initiatives (n.d.), https://breakthroughinitiatives.org/

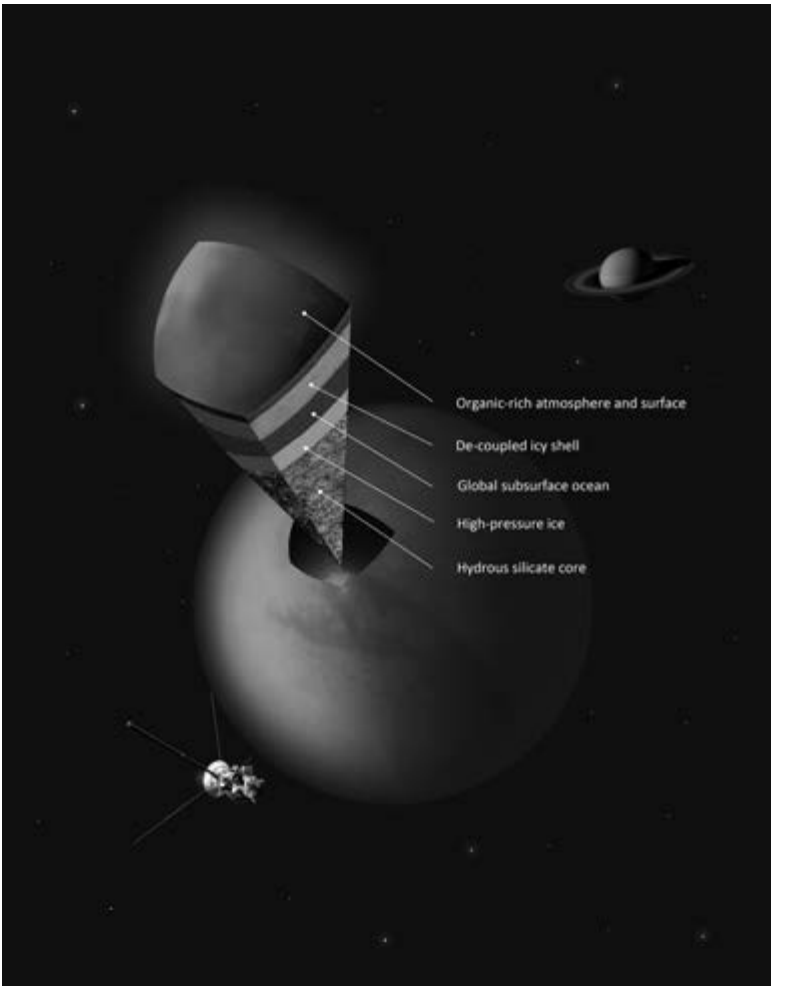


Figure 7.3 An artist's rendering of the complex internal surface of Titan, including the presence of a subsurface ocean that is ostensibly bounded from above and below by ice layers. (*Source:* A. Tavani / NASA.)

of the greater oceanic depth and higher gravity of Titan, it was originally expected that the bottom of this ocean would maintain contact with high-pressure ices (instead of rock), although this conventional viewpoint has been challenged by recent numerical models (Vance et al. 2018). A schematic illustration of Titan's interior, consisting of multiple layers, is presented in Figure 7.3.

Despite the fact that the *Cassini-Huygens* mission was completed only recently, the plans for the next major mission to Titan are already underway. The *Dragonfly* mission by NASA is anticipated to reach this world at the end of 2034 and study Titan's diverse surface environments.[5] Titan's real value to astrobiology is attributable to the presumption that it may represent the optimal place to look for life that is *dissimilar* to that found on Earth; for instance, based on nonpolar (e.g., methane) solvents. *Dragonfly* will deploy a rotorcraft lander, equipped with instruments such as a mass spectrometer and meteorology sensors, to investigate the habitability, rich prebiotic chemistry, prospective biosignatures, and the subsurface ocean of Titan, among other objectives.

7.1.4 Other worlds with subsurface oceans

We have seen that multiple factors indicate the existence of subsurface oceans on Europa, Enceladus, and Titan. However, there are plenty of other worlds in our Solar system that might possess subsurface oceans, but there is only one line of evidence (or none) for the majority of them. We will briefly summarize some of the more intriguing contenders below.

Callisto: One of Jupiter's Galilean moons, its radius is about 38 percent that of the Earth. The primary evidence for a subsurface ocean on Callisto is via induced perturbations arising from Jupiter's magnetic field, akin to Europa's, although the fluctuations measured for Callisto have a lower amplitude. Moreover, the presence of ancient craters on Callisto has been offered as evidence of a thick and cold crust, implying that the ocean may be situated more than 100 km below the surface. JUICE is expected to undertake several flybys of Callisto and will therefore provide more information about the nature of its subsurface ocean.

Ganymede: Ganymede is the largest moon in the Solar system, with a radius that is around 41 percent that of the Earth. Furthermore, Ganymede is the only moon known to possess an intrinsic magnetic field. On account of this reason, it has proven to be more complicated to establish the presence of a

5. Dragonfly (n.d.), We're going to Titan: Dragonfly has been selected as NASA's next New Frontiers mission, https://dragonfly.jhuapl.edu/

subsurface ocean by measuring the induced magnetic field. Instead, HST observations of Ganymede's aurorae offer some of the best evidence to date since the spatial extent of aurorae (auroral ovals) is governed by its internal structure; more precisely, the oscillation of the auroral ovals is reduced in the event of a subsurface ocean by a factor of 2 or 3 owing to the electromagnetic induction within the ocean. The present evidence based on aurorae appears to indicate that Ganymede's ocean may exist at a depth of 150 to 250 km (Saur et al. 2015). The primary objective of the JUICE mission is to move into orbit near Ganymede and characterize its internal structure and ocean in detail.

An important point worth noting with regard to both Callisto and Ganymede is that their ice-rock composition, greater size, and subsurface oceans located deeper beneath the surface are all expected to contribute to the presence of high-pressure ices at the ocean floor (Vance et al. 2018). Thus, in this scenario, there will be no direct water-rock interactions, which could pose impediments to the emergence of life. On a related note, the subsurface oceans of Callisto and Ganymede are predicted to have total depths of $\mathcal{O}(100)$ km (Spohn & Schubert 2003).

Saturnian moons: Mimas is another of Saturn's smaller moons, with a radius that is only 3 percent that of the Earth. The libration amplitude is higher than the expected value by a factor of 2, and the two primary hypotheses are (1) a nonhydrostatic interior and (2) a hydrostatic interior with a subsurface ocean. Owing to Mimas's absence of geological activity, despite the high tidal stresses that it experiences, it seems more likely that (2) is invalid. Early measurements by the *Cassini* spacecraft tentatively suggested the presence of plumes on Dione and Tethys, two of Saturn's moons with radii $0.09\,R_\oplus$ and $0.08\,R_\oplus$, respectively, but no further evidence has been found in subsequent observations.

Kuiper Belt Objects (KBOs): Virtually nothing is known of the potential subsurface oceans in many denizens of the Kuiper Belt, nor can it be said that we have a complete inventory of objects $\gtrsim 100$ km in size. In the case of the dwarf planet Pluto (whose radius is $0.19\,R_\oplus$), its moon Charon (with radius $0.10\,R_\oplus$), and Neptune's moon Triton (whose radius is $0.21\,R_\oplus$), there is some evidence for geological activity based on the terrain or outgassing. Coupled with thermal models of their interiors, there are grounds to believe

that subsurface oceans may exist on some of the KBOs at hundreds of kilometers below the surface. However, there is no direct observational evidence for their existence, although theoretical models favor this outcome (Bierson et al. 2020).

Ceres: Ceres, the only dwarf planet in the Solar system that lies within the orbit of Neptune, is about 7 percent the size of the Earth. The NASA spacecraft *Dawn* arrived at Ceres in 2015 and enhanced our understanding of this world through high-resolution imaging. Currently, there is fairly compelling evidence of geological activity on Ceres, and localized sources producing water vapor have also been identified (Castillo-Rogez et al. 2020). A wide range of compounds, some of which resemble those found in Enceladus's plume, have been apparently detected on the surface of Ceres, including carbonates, hydrated silicates or H_2O ice, and ammonium chloride. The existence of certain compounds on the surface has been interpreted as being consistent with the presence of subsurface H_2O either now or in Ceres's past. One of the other points in favor of a subsurface ocean is that Ceres has a higher surface temperature than the likes of Europa and Enceladus, and its importance will become apparent shortly.

7.2 TEMPERATURE PROFILES OF THE ICE ENVELOPES

In our analysis, we shall consider a simplified icy world with an outer ice envelope, a subsurface ocean, and a mantle and core made up of silicates and metals;[6] to preserve brevity, we opt to dub this region the "core" (see Figure 7.4). Before proceeding further, a few cautionary statements are in order. First, we do not tackle cases wherein the outer envelope consists of a mixture of ice and rock, although such worlds do exist in our Solar system (e.g., Titan). Second, the average H_2O inventory of these worlds will doubtless vary significantly. The water inventory is characterized by the average water depth D, and we require $\mathcal{H} = D - H > 0$, where $\mathcal{H}$ and H are the thickness of the ocean and icy shell, respectively. In qualitative terms, this criterion ensures that all of the water on these worlds is not frozen as ice.

6. NASA Science, Solar System Exploration (n.d.), Europa: Moon of Jupiter, https://solarsystem.nasa.gov/moons/jupiter-moons/europa/overview/

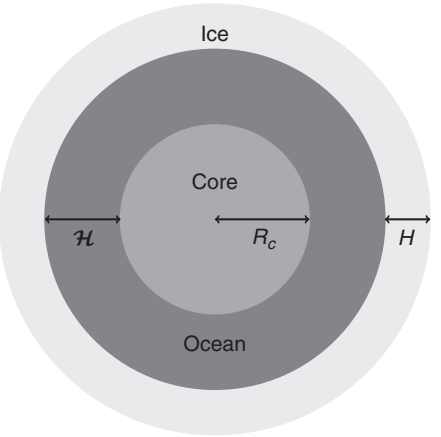


Figure 7.4 The internal structure of a generic subsurface ocean world. It has an ice envelope and a liquid water ocean of thickness H and $\mathcal{H}$, respectively, and an inner mantle and core of radius R_c composed of silicates and / or metals. (© Manasvi Lingam and Avi Loeb.)

We will focus only on those worlds where the geothermal heat flux is primarily due to radiogenic (radioactive decay) and primordial (from gravitational collapse) heating. In the case of modern Earth, the two factors are known to contribute roughly equally to the overall internal heat flux. As a result, even if we focus solely on radiogenic heating, our results are still likely to be correct to within an order of magnitude. There is one major factor that we do not explicitly incorporate within our analysis: tidal heating. Tidal heating occurs when the tidal bulge varies over an eccentric orbit, and this oscillation generates internal friction and heating. The tidal heating $\dot{E}_T$ is given by

$$\dot{E}_T = -\frac{21}{2}\,\mathrm{Im}\,(k_2)\,\frac{R^5 n^5 e^2}{G}, \tag{7.1}$$

where the second factor on the right-hand side is the imaginary part of the complex second-order Love number k_2. In the formula, R, n, and e are the radius, mean motion frequency, and orbital eccentricity of the subsurface ocean world, respectively. Tidal heating is known to be important for the likes of Europa and Enceladus, but it does not play a major role for many of the other subsurface ocean worlds in our Solar system, and it is not a factor when it comes to free-floating planets with subsurface oceans. In addition,

we also neglect the heating due to serpentinization reactions, although this source can become significant on small worlds like Mimas and Enceladus.

The thermal profile of the ice layer is determined on the basis of whether the heat is transported via conduction or convection. Several subtleties are linked with the latter, and we will return to a brief discussion of convection at the end of this section. Thus, in the case of conduction, the temperature can be computed from Fourier's law as follows:

$$\mathcal{Q} + \kappa(T)\frac{dT}{dr} = 0, \tag{7.2}$$

where $\mathcal{Q}(r)$ represents the geothermal heat flux at radius r (in units of $\mathrm{W/m^2}$), $T \equiv T(r)$ denotes the temperature at radius r (with $r = R$ being the outer surface), and $\kappa(T)$ signifies the thermal conductivity of ice. κ is dependent on T as per the relation $\kappa = C_I/T$ with $C_I \approx 651$ W/m (Petrenko & Whitworth 1999). The heat flux $\mathcal{Q}(r)$ can be further expressed as

$$\mathcal{Q} = \left(\frac{Q}{4\pi r^2}\right)\left(\frac{r^3}{R^3}\right) = \frac{Q\,r}{4\pi R^3}, \tag{7.3}$$

with Q representing the total internal heat flow from the planetary interior to space (in units of W). We have obtained the above relation by using $\mathcal{Q} = Q_{\mathrm{enc}}/(4\pi r^2)$ (i.e., the heat flux is the ratio of the heat flow within the enclosed volume and the area of this region). Q_{enc} is further determined by assuming that the radioactive elements are uniformly distributed within the volume of our icy world, which leads us to

$$Q_{\mathrm{enc}} \approx Q\left(\frac{4\pi}{3}r^3\right)\left(\frac{4\pi}{3}R^3\right)^{-1}. \tag{7.4}$$

Now, we require a parametrization for Q in terms of the mass M of the icy world. A reasonable approximation for Q is

$$Q = \Gamma\, Q_{\oplus}\left(\frac{M}{M_{\oplus}}\right)^{\alpha}, \tag{7.5}$$

with $Q_{\oplus} \approx 22$ TW being the radiogenic heat flow of the Earth, while Γ and α are free parameters. It is essential to recognize that both Q and $Q_{\oplus}$ are actually time dependent, since the amount of heat produced through the

decay of radioactive elements falls off exponentially with time. If we restrict ourselves to icy worlds formed a few Gyr ago (or even more recently), Q will change merely by a factor of order unity, provided that only long-lived isotopes are taken into consideration. We have excluded short-lived isotopes from our treatment, but these elements could play a potentially important role in the thermal evolution of icy worlds, thus governing their final water inventories (Lichtenberg et al. 2019).

The standard case often invoked in the literature is the assumption that the heating per unit mass (Q/M) is constant, which amounts to setting $\alpha = 1$. In reality, however, the amount of radiogenic heating ought to depend on the mass of the mantle. Hence, unless the mass of the mantle is linearly proportional to the total mass of the icy world, there is no guarantee that $\alpha = 1$ is valid. Next, when we turn our gaze toward Γ, it can be duly interpreted in two different ways. If $\Gamma = 1$, we may assume that most of the heat is derived from radioactivity. However, as we have pointed out earlier, there are sources such as tidal heating that greatly enhance the overall heating. We can, therefore, construe Γ as a free parameter that quantifies the ratio of the total heat flow to that of radiogenic heating.

There are, however, no a priori reasons for supposing that all icy worlds will have similar chemical composition. In fact, icy worlds with concentrations of short- and long-lived radionuclides markedly distinct from that of Earth are anticipated to exist in galaxies (Lingam & Loeb 2020d). In this event, such variability is encapsulated by the factor Γ that would take values much larger or smaller than unity. A chief advantage of envisioning Γ *in hoc sensu* is that we could constrain it through meticulous analyses of stellar spectra; for instance, uranium is measurable through the U II spectral line at 3860 Å. Stellar spectroscopic surveys collectively suggest that the abundances of long-lived radioactive elements vary by orders of magnitude across stars (Frebel 2010). In this context, observations by Unterborn et al. (2015) concluded that the Sun is depleted in thorium with respect to other solar-type stars, indicating that subsurface icy worlds orbiting them can have higher energy budgets.

Lastly, we must express the mass M in terms of the radius R. Most mass-radius relationships are parameterized as power laws, but this feature is not always universal. Nevertheless, for the sake of simplicity, we use the ansatz

$$M = M_{\oplus} \left(\frac{R}{R_{\oplus}} \right)^{\beta}, \tag{7.6}$$

with β being a free parameter that depends on the exact composition of the icy world. We shall select $\beta \approx 3.3$ for $M \lesssim M_{\oplus}$ (Sotin et al. 2007) and $\beta \approx 3.8$ for $M \gtrsim M_{\oplus}$ (Fu et al. 2010), based on theoretical models of exoplanetary interiors.

We proceed to solve (7.2) by imposing the boundary condition $T(r = R) = T_s$, where the temperature at the surface is T_s. Thus, the temperature profile as a function of the radius is

$$\ln\left(\frac{T}{T_s}\right) = \frac{Q\left(R^2 - r^2\right)}{8\pi C_I R^3}. \tag{7.7}$$

For calculating the thickness of the ice layer, note that the melting point of ice ranges approximately between 250 and 270 K, provided that the total pressure P at the bottom of the ice layer is lower than 620 MPa. From (7.7), we see that the temperature dependence is relatively weak since it is logarithmic, and this allows us to choose a melting point of $T_m = 260$ K for pure ice. However, it should be recognized that the presence of contaminants will lower the melting point of ice. For example, if ammonia is added as a contaminant, the resulting ice-ammonia mixture can remain in liquid form until 176 K. The value of $r = R_m$ at which $T = T_m$ occurs is found to be

$$R_m = R\left[1 - \frac{8\pi C_I R}{Q} \ln\left(\frac{T_m}{T_s}\right)\right]^{1/2}, \tag{7.8}$$

and the thickness H of the ice layer is determined via $H = R - R_m$. Thus, we end up with

$$\frac{H}{R_{\oplus}} = \left(\frac{R}{R_{\oplus}}\right)\left(1 - \left[1 - 2.4 \times 10^{-3} \frac{\ln \Lambda}{\Gamma}\left(\frac{R}{R_{\oplus}}\right)^{1-\gamma}\right]^{1/2}\right), \tag{7.9}$$

and the following auxiliary variables have been introduced: $\gamma = \alpha\beta$ and $\Lambda = T_m / T_s$. An inspection of this formula reveals that H will decrease whenever Λ is decreased or Γ is raised for fixed values of $R / R_{\oplus}$ and γ. In Figure 7.5, the thickness of the ice shell H has been plotted as a function of the radius R for different values of the two free parameters Γ and Λ. Recall that icy worlds host subsurface oceans only when $D > H$, where D denotes the

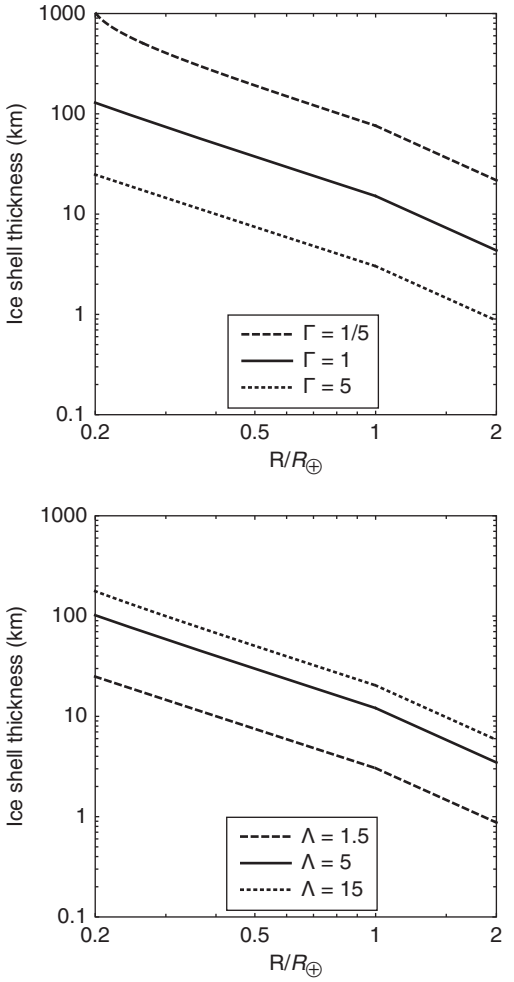


Figure 7.5 *Top:* Ice shell thickness versus the radius for varying concentrations of radioactive elements (encoded within Γ); the surface temperature is taken to be that of the Earth if it was ejected into space (set by $\Lambda = 7.5$). *Bottom:* Thickness of the ice envelopes for varying surface temperatures (different values of Λ) and a fixed concentration of long-lived radionuclides ($\Gamma = 1$). (© Manasvi Lingam and Avi Loeb.)

average water depth and it cannot be quantified a priori since the water inventory can vary widely.

It is instructive to consider what would happen if present-day Earth were to be ejected, and assume that the heat would be transported across the ice shell through conduction. The surface temperature is calculated by solving $\sigma T_s^4 = \mathcal{F}_\oplus$, where $\mathcal{F}_\oplus = 0.087$ W / m^2, and we obtain $T_s \approx 35$ K. Hence, if we make use of $T_m \approx 260$ K for pure ice, we have $\ln \Lambda \approx 2$. Upon making use of this value in (7.9) in conjunction with $\Gamma = 1$ and $R = R_\oplus$, we find $H \approx 15$ km. In contrast, the Earth's average oceanic depth $D_\oplus$ is known to be approximately 3.7 km. Hence, as per this model, the Earth's oceans—including the deepest regions like the Mariana Trench in the western Pacific Ocean—are expected to become wholly frozen. Yet, even if this outcome were valid, it would not necessarily spell the end of life on Earth. This is because the prospects for the survival of chemolithoautotrophic life in the deep biosphere (e.g., hydrated areas in the subduction zones) ought not be discounted.

On the other hand, suppose that we contemplate the ejection of the Earth shortly after its formation in the Solar system nursery. The geothermal heat flux would have been higher by a few times due to the elevated concentrations of short- and long-lived radionuclides alongside greater primordial heating. If the criterion $\Gamma \gtrsim 4$ was valid, there is a strong possibility that the primordial Earth may have retained a global subsurface ocean for a while after becoming a free-floating denizen of interstellar space. The reason is because the condition $D_\oplus > H$ is satisfied since $D_\oplus$ was approximately twice its value for modern Earth (Korenaga 2008), and we obtain $H \approx 3.75$ km after using (7.9) for the choice of $\Gamma \approx 4$.

Let us specialize to the case wherein the second term within the square brackets of (7.9) is much smaller than unity. By using the binomial theorem, we derive the following expression:

$$H \approx 7.6\,\mathrm{km}\,\frac{\ln \Lambda}{\Gamma}\left(\frac{R}{R_\oplus}\right)^{2-\gamma} \tag{7.10}$$

It can be seen that H exhibits a power-law dependence on R and is a monotonically decreasing function of the latter provided that $\alpha = 1$ and the values of β discussed after (7.6) are valid. Before proceeding further, an important aspect of (7.9) is worth highlighting—namely, it becomes mathematically invalid for certain values of R. This cutoff arises because the expression

inside the square brackets must be positive to ensure that the values of H are real. Hence, the lower bound for icy worlds, denoted by R_{crit}, imposed by the constraints of our model is

$$\frac{R_{crit}}{R_{\oplus}} = \left(2.4 \times 10^{-3} \, \frac{\ln \Lambda}{\Gamma}\right)^{1/(\gamma-1)} . \tag{7.11}$$

If we choose the parameters $\ln \Lambda \approx 2$ and $\Gamma = 1$, which are characteristic of a free-floating planet, we obtain $R_{crit} \approx 0.1 \, R_{\oplus}$. Evidently, the cutoff radius R_{crit} becomes higher whenever Λ is increased, but it decreases for larger values of Γ. Finally, we reiterate that our model will not yield accurate results for Europa and Enceladus because the effects of tidal dissipation are not explicitly included, although the phenomenological factor Γ can encapsulate the degree of tidal heating; the reader is directed to Tjoa et al. (2020) for a more elaborate analysis. For instance, when it comes to Europa, the value of H calculated from (7.9) is about four times higher than the results from sophisticated numerical models that incorporate tidal dissipation.

Hitherto, we have restricted ourselves only to heat transport through conduction. We turn our attention to convection: the other major mode of transport across the ice shell. The major issue with studying convection is that it depends on a wide range of factors, some of which cannot be easily quantified. Some of the relevant parameters include the mass of the icy world, availability of water, size of ice grains, presence of contaminants, and rheology (the flow response of matter to applied force). Theoretical modeling is further complicated by the fact that convection is inherently dynamical, implying that the existence of bistability, multiple steady states, and temporal evolution must be taken into account. In spite of these complexities, a sufficiently thick ice shell is expected to ensure the onset of convection. In our Solar system, the ice envelopes of Ganymede and Callisto have been predicted to be convective, but the likes of Europa and Enceladus may possess conducting ice shells instead (Nimmo & Pappalardo 2016).

The total thickness of the ice lid in the convective regime can be found by means of the following procedure. We only describe the essential steps; further details are in Hussmannn et al. (2006) and Abbot and Switzer (2011). In order for convection to occur, the Rayleigh number (Ra) must exceed a critical threshold; it is defined as

$$\mathrm{Ra} = \frac{\rho g \alpha_I \left(T_m - T_c\right) H_c^3}{\kappa_I \eta}, \tag{7.12}$$

where $\rho = 917$ kg / m^3 is the density of ice, $g \approx GM/R^2$ is the acceleration due to gravity, H_c is the thickness of the convecting region, and T_c is the temperature at which convection is initiated. The quantities $\alpha_I = 1.56 \times 10^{-4}$ K^{-1}, $\kappa_I = 1.47 \times 10^{-6}$ m^2 / s, and η are the thermal expansion coefficient, thermal diffusivity, and viscosity of ice, respectively. The viscosity is a complex function of the temperature and is usually approximated by

$$\eta(T) = \eta_0 \exp\left[l\left(\frac{T_m}{T} - 1\right)\right], \tag{7.13}$$

where $\eta_0 \approx 10^{13}$ Pa s represents a fiducial viscosity, and $l \approx 25$. The viscosity within the convecting region varies only by an order of magnitude; hence the condition $\eta\left(T_c\right) \approx 10\eta\left(T_m\right)$ enables us to solve for T_c. The height H_c is computed by means of a scaling between the Nusselt (Nu) and Rayleigh numbers of the form $\mathrm{Nu} = 0.12\,\mathrm{Ra}^{0.3}$, and this leads us to

$$H_c \approx \left[\frac{0.12 k_I \left(T_m - T_c\right)}{\Gamma \mathcal{F}_\oplus}\left(\frac{R}{R_\oplus}\right)^{2-\gamma}\right]^{10}\left[\frac{\rho g_\oplus \alpha_I \left(T_m - T_c\right)}{\kappa_I \eta\left(\bar{T}\right)}\left(\frac{R}{R_\oplus}\right)^{\beta-2}\right]^3, \tag{7.14}$$

where $k_I = 3.3$ W m^{-1} K^{-1} is the thermal conductivity that is assumed to be roughly constant within this layer, given that the temperature variation will not be significant; $\bar{T} = \left(T_m + T_c\right)/2$, and $g_\oplus$ is Earth's surface gravity. The total thickness of the ice sheet H is the sum of the thickness of the convection region H_c and the sum of the stagnant lid H_s, where the latter is given by (7.10) except that Λ is replaced by $\Lambda_c = T_c/T_s$. We leave it as an exercise for the reader to plot the dependence of H on $R/R_\oplus$ for different values of T_m and T_s.

7.3 THE HABITATS FOR SUBSURFACE OCEAN WORLDS

As we have stated, worlds with subsurface oceans may be broadly classified into two different categories: those orbiting a host star and those that float freely in space. We will discuss both of these scenarios in greater detail below and delineate the potential values of Λ and T_s.

7.3.1 Type B worlds

We refer to worlds that orbit their host star as *Type B* worlds,[7] with the letter *B* indicating that they are gravitationally bound. As we have remarked, planets and moons situated beyond the outer edge of the HZ are capable of having subsurface oceans. Apart from the likes of Europa and Enceladus within our Solar system, water-rich extrasolar planets outside the HZ, such as OGLE 2005-BLG-390Lb (Ehrenreich et al. 2006), fall under this category. When we refer to the HZ, we are not only concerned with main-sequence stars but also with those stars that lie within the pre- and post-main-sequences. Thus, subsurface ocean worlds may exist not only around main-sequence stars but also around the likes of red giants and white dwarfs.

It is worth emphasizing that the value of Λ can vary almost by an order of magnitude for Type B worlds. As an example, let us dwell on the dwarf planet Ceres. If we adopt a uniform temperature of 168 K, we end up with $\Lambda \approx 1.5$ upon choosing $T_m \approx 260$ K. Clearly, this value of Λ will become even smaller either in the presence of contaminants that reduce T_m or when Ceres is at perihelion and Λ is evaluated at the subsolar point (zenith). In both these cases, the value of Λ could be approximately 1.1. Next, suppose that we consider Pluto at aphelion with a surface temperature of around 33 K. In this event, if we assume $T_m \approx 260$ K, it is found that $\Lambda \approx 7.9$.

For most Type B worlds, it is fairly straightforward to calculate their surface temperature T_s by means of the Stefan-Boltzmann law; therefore, we obtain

$$T_s \approx 213\,\mathrm{K} \left(\frac{1-A}{0.36}\right)^{1/4} \left(\frac{L_\star}{L_\odot}\right)^{1/4} \left(\frac{a}{1\,\mathrm{AU}}\right)^{-1/2}, \qquad (7.15)$$

where A is the albedo of the icy world, $L_\star$ represents the luminosity of the host star, and a signifies the orbital radius of the planet. In the above equation, we have opted to normalize $1 - A$ by 0.36 because this happens to be the estimate for Europa. We may further express the equation in terms of the stellar mass $M_\star$ via the mass-luminosity relationship $L_\star/L_\odot \propto (M_\star/M_\odot)^\xi$ for

7. Alternatively, we could have called such worlds *Hoths* in honor of the ice-covered world Hoth from *Star Wars,* although the question of whether Hoth has a subsurface ocean remains unresolved. In the same spirit, one may christen ocean worlds (with oceans on the surface and no continents) *Kaminos,* based on the aquatic planet Kamino from the same movie franchise.

main-sequence stars, but the value of ξ varies depending on the mass range. To first order, $\xi \approx 3.5$ can be employed for most G-type stars.

7.3.2 Type U worlds

Dynamical models of planetary formation have revealed that many worlds can be ejected from their host system by means of gravitational interactions; a review of this subject is laid out in Morbidelli et al. (2012). As these objects are freely floating, we refer to them as *Type U* worlds with the letter *U* denoting that they are unbound—that is, not gravitationally bound to a particular star. Current estimates indicate that there are a large number of free-floating objects in interstellar space; this topic will be covered in more detail when we reach Section 7.8. Despite the large number of such free-floating worlds, not much analysis has been devoted to their potential habitability. One of the first papers in this area was by D. J. Stevenson (1999), who presented the requirements for hosting liquid water on the *surface*. Subsequently, Abbot and Switzer (2011) coined the term *Steppenwolf* to describe a free-floating planet with subsurface oceans and derived the thickness of the ice shell. The rationale behind adopting this particular word merits highlighting: "*any life in this strange habitat would exist like a lone wolf wandering the galactic steppe*" (Abbot & Switzer 2011, emphasis ours).

Given that Type U worlds do not orbit a host star, they are typically expected to be very cold, and it would therefore be tempting to conclude that they would be characterized by $\Lambda \sim 10$. When their surface temperature T_s is set by the radiogenic heat flux, we end up with

$$T_s \approx 29\,\mathrm{K}\ \Gamma^{1/4} \left(\frac{R}{R_\oplus}\right)^{(\gamma-2)/4}. \tag{7.16}$$

Thus, for a Europa-sized world with $R/R_\oplus \sim 0.25$, we obtain $T_s \approx 19$ K for $\Gamma = 1$. This corresponds to $\Lambda \approx 13.8$, which appears to confirm our initial impression that $\Lambda \sim 10$. However, there are certain environments (or habitats) where lower values of Λ are feasible; we shall describe a couple of these possibilities below.

First, consider stars that are close to the Galactic center. During a quasar phase, the luminosity of supermassive black holes is very high, which can raise the ambient temperature of Type U (or even Type B) planets in the vicinity. In quantitative terms, the luminosity during the quasar phase

$L_{\rm BH}$ could be the same order as the Eddington luminosity and is therefore given by

$$L_{\rm BH} \approx L_{\rm Edd} = 1.3 \times 10^{37}\ {\rm W} \left(\frac{M_{\rm BH}}{10^6\, M_\odot} \right), \tag{7.17}$$

with $M_{\rm BH}$ denoting the mass of the supermassive black hole. Now, let us suppose that we are interested in obtaining $T_s \approx 100$ K, which corresponds to $\Lambda \approx 2$ if the ice has some contaminants such as ammonia. The distance from the supermassive black hole $\mathcal{R}$ at which this value of T_s occurs is found from

$$\sigma\, T_s^4 \approx \frac{L_{\rm BH}}{4\pi \mathcal{R}^2}, \tag{7.18}$$

and this leads us to the distance:

$$\mathcal{R} \approx 12\,{\rm pc} \left(\frac{L}{10^{37}\,{\rm W}} \right)^{1/2} \left(\frac{T_s}{100\,{\rm K}} \right)^{-2}. \tag{7.19}$$

The stellar density in this region of the Galactic center is $\sim 10^5$ pc^{-3}, implying that there are $\sim 10^9$ stars in a spherical region of 12 pc. The theoretical estimates for free-floating objects based on quasar microlensing suggest that there may exist $\sim 2 \times 10^3$ worlds per star that are bigger than the size of the Moon (Dai & Guerras 2018). Thus, it becomes apparent that the total number of Type U objects in this region could be very large, with an upper bound of $\sim 10^{12}$ worlds.

The distance derived in (7.19) is about two to three orders of magnitude smaller than the inner edge of the Galactic Habitable Zone (GHZ), which is usually taken to be a few kiloparsecs from the center. The GHZ represents the region within the Galaxy where life might have a higher likelihood of evolving, and it takes into account a number of factors ranging from the metallicity (required for planet formation) to the rate of astrophysical catastrophes like supernovae and Gamma Ray Bursts (GRBs); we touched on the GHZ in Section 4.1.[8] Yet, the delineation of the GHZ is contingent

8. When we use the word *metallicity* specifically, it should be understood that we are employing the astrophysical definition unless stated otherwise—namely, we are referring to the mass fraction of elements heavier than hydrogen or helium. The parameter [Fe / H] is

on certain implicit assumptions, including the existence of surficial liquid water and life.

In contrast, sufficiently thick ice layers on Type U worlds potentially shield them from ionizing radiation emitted by supernovae and GRBs. Another advantage is that deleterious surface-based effects such as atmospheric erosion due to hydrodynamic escape are rendered irrelevant. The depletion of planetary atmospheres is especially a bona fide concern during the active phase of the quasar. Theoretical models indicate that ~ 10 percent of all planets in the Universe may lose a total atmospheric mass comparable to that of Earth driven by quasar activity (Balbi & Tombesi 2017; Forbes & Loeb 2018). One possible caveat worth mentioning here is that the higher flux of energetic particles expected to be prevalent in this environment will impact the surface and contribute to the erosion of the ice layer via the mechanism of sputtering. If we consider the fairly intense radiation environment at Europa, by virtue of Jupiter's strong magnetic field and large magnetosphere, sputtering is predicted to erode ice at the rate of $\sim$ 20 m per Gyr (Cooper et al. 2001). Hence, unless the sputtering rate is two to three orders of magnitude higher near the Galactic center, its effects should not be significant over Gyr timescales. One other distinct possibility is that the high density of stars leads to frequent gravitational interactions (McTier et al. 2020), which could prove disruptive to subsurface habitability.

Another important factor should be taken into consideration. The characteristic lifetime of the quasar phase is quite short: typically around 10^7 to 10^8 years. After this period, the surface temperature will drop by a factor of a few, and the thickness of the ice envelope will increase in accordance with (7.9). Interestingly, although this timescale is very short compared to the age of the Earth, some proponents suggest that a timescale of $\sim 10^7$ yrs may have sufficed for life to originate on Earth, based on the observed increase in genome complexity and geological considerations (Lazcano & Miller 1994). On the other hand, we caution that this timescale remains unconfirmed even on our own planet, to say nothing of other worlds where the uncertainties are higher.

---

widely used as a proxy for the metallicity and is defined as the logarithm of the ratio of Fe and H *relative* to the solar Fe / H ratio.

Observations of giant molecular clouds situated near the Galactic center (e.g., Sagittarius B2) have detected a wide range of (in)organic molecules at relatively high concentrations (Herbst & van Dishoeck 2009), of which some are believed to play an important role in prebiotic chemistry. Examples of the molecules found include hydrogen cyanide, aldehydes, and nitriles. There was equivocal and contested evidence for the presence of glycine, the simplest of the amino acids, in the Central Molecular Zone, but subsequent observational studies have not yielded any clear-cut positive results. Thus, taken collectively, it seems plausible that Type U (or Type B) worlds existing in these regions may be characterized by $\Lambda \sim 1$ and have access to a steady supply of these prebiotic molecules, although they need to be transported from the icy surface to the subsurface ocean.

A second avenue for achieving $\Lambda \sim 1$ is through the background heating supplied by the cosmic microwave background (CMB). At the current epoch, the temperature of the CMB is very low, but the CMB energy density is strongly dependent on the age of the Universe through the redshift z. Thus, upon equating this energy density with σT_s^4, the surface temperature is

$$T_s \approx 82\,\mathrm{K}\left(\frac{1+z}{30}\right). \tag{7.20}$$

From this equation, it becomes apparent that selecting $z \approx 30$ leads us to $\Lambda \approx 2$ for ice-ammonia mixtures. Hence, any planets that formed during this epoch may have possessed thinner ice shells owing to the higher surface temperatures. Naturally, there are several challenges that life in a high-redshift universe must confront; some of them are delineated below.

- It is important for the first stars to have formed and seeded the Galaxy with heavy elements after exploding as supernovae. As per our current theoretical understanding, the first stars would have formed at $z \lesssim 70$ (Loeb & Furlanetto 2013), and there is observational support for metal-poor stars at high redshifts of $z \approx$ 10–20 (Bouwens 2018). Hence, there are grounds for believing that stars were present at the redshifts under consideration.
- Despite the fact that exoplanets have been observed around stars with a diverse range of metallicities, the formation of planetesimals is expected to depend to some degree on the metallicity. In this event, the low-metallicity environment of the early Universe may have posed difficulties for planetary formation, although

theoretical models suggest that the first terrestrial planets had formed by 13 Ga.

- The essential elements for life-as-we-know-it,[9] further described in Section 7.6, must be present at sufficient concentrations for life to originate and evolve. The potential existence of carbon-enhanced metal-poor (CEMP) planets at high redshifts can be viewed as evidence in favor of the availability of carbon, but the abundances of phosphorus and sulfur remain poorly constrained for exoplanets, even at low redshifts.
- Clearly, the availability of liquid water is necessary for life-as-we-know-it. As a result, H_2O formation in molecular clouds and its delivery to protoplanetary disks are of high importance. Numerical models indicate that water vapor is common in these clouds even at low metallicities that are $\sim 10^{-3}$ of the solar value (Bialy et al. 2015), implying that the prevalence of liquid water at high redshifts ought not be ruled out.

Observational evidence has established that the present-day energy densities of interstellar radiation, cosmic rays, and the CMB are similar to each other in the Milky Way. Collectively, they add up to a surface temperature of a few Kelvin, and this yields an absolute lower bound on the magnitude of T_s.

7.4 THE ROUTES TO ABIOGENESIS ON SUBSURFACE OCEAN WORLDS

Hitherto, we have been content with evaluating the thickness of ice layers and examining the prospects for subsurface oceans of liquid water. Needless to say, this constitutes an important component of habitability due to water's necessity as a solvent for life. However, a vital point that often goes unappreciated is that water is an active medium whose temperature and pressure, salinity, and ionic strength inter alia are subject to considerable variations. As a result, the vast majority of Earth's aquasphere (around 88 percent) does not appear to host life (E. G. Jones & Lineweaver 2010). In

9. As explained in Chapter 1, we use the word "life" as shorthand for the longer phrase "life-as-we-know-it." From a pragmatic standpoint, we trust that the readers will recognize and account for the subtle distinctions between these two concepts as and when the occasion calls for it.

view of these caveats, while the mere existence of liquid water represents a necessary condition for life, several other criteria must be satisfied at the minimum.

One of the most notable among them is the availability of free energy sources for facilitating prebiotic chemistry and the origin of life, for reasons elucidated in Section 2.3. Moreover, energy sources are essential for the maintenance of ecosystems and biospheres, as we shall witness in Section 7.5. Here, we will adopt a "follow the energy" strategy to quantify the various energy sources within reach for prebiotic chemistry, and we will briefly sketch the possible routes to abiogenesis.

7.4.1 Energy sources for prebiotic synthesis

The first source worth considering in our analysis is UV radiation. The irradiation of interstellar ice analogs at low temperatures has been shown to produce a wide range of vital prebiotic compounds (Öberg 2016; Sandford et al. 2020). Some of the organic molecules that have been synthesized via this route include amino acids (e.g., alanine, glycine, serine) and the canonical RNA / DNA nucleobases. Most of these experiments have been conducted by using a flowing-hydrogen discharge lamp, whose output is divided almost equally between the Lyα line and a 20 nm band centered at around 160 nm. Hence, we can use the flux of Lyα photons received at the surface of Type B worlds as a reasonable proxy for measuring the energy available for prebiotic synthesis. The flux received by the Type B world depends not only on its orbital distance a but also on the mass of the host star $M_\star$. A heuristic expression for the UV energy flux (denoted by $\Phi_{\rm UV}$) is given by

$$\Phi_{\rm UV} \sim 10^6\,{\rm J\,m^{-2}\,yr^{-1}} \left(\frac{a}{1\,{\rm AU}}\right)^{-2} \left(\frac{M_\star}{M_\odot}\right)^{\nu}, \tag{7.21}$$

where ν is a power-law exponent that is empirically determined for different stars. For example, Lingam and Loeb (2018d) roughly estimate that $\nu \approx 1.2$ for $M_\star \lesssim M_\odot$ and $\nu \approx 6.8$ for $M_\odot < M_\star \lesssim 2\,M_\odot$. This exponent is not constant for all stars because a significant portion of the Lyα emission from low-mass stars (particularly M-dwarfs) is from stellar chromospheric activity, in contrast to Solar-type stars. Although the above expression only applies to Type B worlds, any Type U worlds passing in the vicinity of O / B-type

stars would receive high, but transient, doses of UV radiation. Type U worlds that fall under this category may consequently witness the rapid formation of biologically relevant molecules, albeit over a short period of time.

As we are primarily interested in the feasibility of prebiotic synthesis, we need to select a particular class of biomolecules and compute the total yields for different pathways. We shall focus on amino acids since they are the building blocks for peptides and proteins and are quite commonly observed as end products in laboratory experiments. By making use of (7.21) and the quantum yield of $\sim 10^{-4}$ for the pathway studied in Muñoz Caro et al. (2002),[10] the mass of amino acids ($\mathcal{M}_{\rm UV}$) produced per year is

$$\mathcal{M}_{\rm UV} \sim 10^{10}\,{\rm kg/yr} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a}{1\,{\rm AU}}\right)^{-2} \left(\frac{M_\star}{M_\odot}\right)^{\nu}. \qquad (7.22)$$

Before proceeding further, some caveats pertaining to UV radiation should be described. Most laboratory experiments operate under the assumptions that the requisite feedstock organic molecules are already available and that the reactions take place at temperatures of a few tens of Kelvin. However, there is no guarantee that the same reactions would occur in an uncontrolled environment at higher temperatures or that the feedstock molecules are sufficiently abundant. In the case of methanol ices, for instance, the yield of prebiotic compounds has been shown to be sensitive to both these factors.

Next, it should be noted that UV photons as well as energetic particles influence not only the formation of organic molecules but also their destruction. At a temperature of 100 K, the half-life of amino acids situated in the uppermost meter of a pure ice layer is of order ten years (Orzechowska et al. 2007). Thus, it is imperative that any prebiotic molecules synthesized on the surface must be briskly transported into the icy crust and eventually into the ocean prior to their widespread decomposition. Lastly, some of the prebiotic pathways studied in the laboratory entailed the existence of liquid water, but this substance will not be present on the surface over long intervals. With that said, the possibility of transient water due to phenomena like impact events (on Type B worlds) and cryovolcanism cannot be ruled out.

---

10. In this context, the quantum yield represents the ratio of the number of biomolecules synthesized to the number of photons absorbed during the chemical reactions.

The next energy source that we consider is the CMB radiation. The CMB energy flux $\Phi_{\rm CMB}$, which is redshift dependent, is expressible as

$$\Phi_{\rm CMB} \sim 4.6 \times 10^2 \ (1+z)^4 \ {\rm J\,m^{-2}\,yr^{-1}}. \tag{7.23}$$

Considering $z \sim 30$, we find that $\Phi_{\rm CMB}$ will increase by almost six orders of magnitude compared to its value in the current epoch. Yet, even at high redshifts, the peak wavelength of the CMB radiation lies in the far-infrared (far-IR). At these relatively low energies, it seems unlikely that this energy will suffice for enabling prebiotic synthesis since the photons must carry enough energy to cleave and form chemical bonds. On the other hand, as we have seen, the surface temperature (7.20) is governed by the CMB radiation and can therefore facilitate the prevalence of thin ice shells at high redshifts.

For some of the widely studied energy sources on Earth, such as electrical discharges and shock waves from impacts, the prebiotic synthesis of organic compounds takes place in the atmosphere. Hence, we do not need to take them into consideration because our subsurface worlds are assumed either to lack an atmosphere altogether or to have a very tenuous one. It is less clear how to deal with prebiotic chemistry involving cryovolcanism; hence we will not deal with this source further in our analysis. The next source that is worth considering is radioactivity, whose surficial flux is denoted by $\Phi_{\rm rad}$. Upon utilizing (7.3) and (7.5), we end up with

$$\Phi_{\rm rad} \sim 2.7 \times 10^6 \, {\rm J\,m^{-2}\,yr^{-1}} \ \Gamma \left(\frac{R}{R_\oplus} \right)^{\gamma - 2}. \tag{7.24}$$

Not much work has been undertaken with regard to understanding how naturally occurring reactors and other radioactive environments on *icy* surfaces can drive the formation of organic compounds. One important point worth contemplating is the potential existence of microbes on Type B and U worlds, akin to *Desulforudis audaxviator* on Earth, that derive their energy from radioactive decay.

To date, we are not aware of any laboratory experiments that have yielded the desired G-values for the synthesis of amino acids in icy environments through natural radioactivity. The G-value represents the number of molecules of the product formed per 100 eV of energy supplied. Hence, we will need to utilize an indirect strategy for computing the yield of amino

acids. To begin with, recall that alpha and beta particle decay mechanisms produce energetic helium nuclei (comprising only protons and neutrons) and electrons, respectively. Therefore, we will posit that the efficiency of prebiotic pathways involving natural radioactivity is somewhat similar to those involving irradiation with energetic protons and electrons. Many lab experiments have been undertaken in connection with the synthesis of organic compounds via energetic particles endowed with energies in the range of KeV to MeV. The organic molecules thus produced include hydrogen cyanide, aldehydes, formamide, amino acids, and nucleosides.

We will consider the pathway investigated by Kobayashi et al. (1995), in which amino acids were obtained (after acid hydrolysis) by irradiating cometary ice analogs at 77 K with 3 MeV protons, and the corresponding G-value was $\sim 10^{-4} - 10^{-5}$. We will employ the lower bound since the efficiency of prebiotic chemistry mediated by natural radioactivity is anticipated to be lower than irradiation by energetic protons. The energy from radiogenic heating in the ice layer ($Q_{\rm ice}$) can be estimated by assuming that the volumetric heating rate is approximately constant and that the volume of the ice shell is $4\pi R^2 H$. With these simplifications, we end up with

$$Q_{\rm ice} \sim Q\left(\frac{4\pi R^2 H}{4\pi R^3/3}\right) \sim Q\left(\frac{3H}{R}\right). \tag{7.25}$$

Hence, upon combining this with the G-value of $\sim 10^{-5}$, we conclude that the mass rate of amino acids synthesized ($\mathcal{M}_{\rm rad}$) is given by

$$\mathcal{M}_{\rm rad} \sim 6.8 \times 10^4\,{\rm kg/yr}\,\Gamma\left(\frac{R}{R_\oplus}\right)^{\gamma-1}\left(\frac{H}{1\,{\rm km}}\right). \tag{7.26}$$

If we wished to calculate the amount of amino acids produced in the ocean, we would need to replace H by the ocean depth $\mathcal{H}$ and potentially use a different G-value for amino acid synthesis because the environmental factors are different.

A point worth emphasizing is that the yields (G-values) depend on the radiation dose and are therefore not truly constant. Furthermore, high-energy particles contribute not only to the formation but also to the destruction of organic molecules in a manner analogous to UV radiation. Lastly, (7.26) should be perceived as an upper bound since we have assumed that all of the energy liberated from radiogenic heating in the ice shell is

available for prebiotic synthesis. In actuality, amino acids and other biogenic molecules are likely to be produced in significant quantities through radiolysis only in localized environments where radioactive elements are present in unusually high concentrations.

In light of the preceding discussion, we pivot to another energy source that we encountered just before: energetic particles. As noted, they produce a wide variety of biologically relevant compounds. There exist at least three distinct sources of energetic particles for Type B worlds, but just one for Type U worlds. While stellar energetic particles (SEPs) and high-energy particles emanating from the magnetospheres of gas giant planets are unique to Type B, Galactic Cosmic Rays (GCRs) are common to both Type B and Type U. Of course, in order for particles from Jovian-type planetary magnetospheres to become a significant energy source (as in Europa), it follows by definition that the Type B world must be a moon.

Let us commence our analysis by estimating the particle flux contributed by planetary magnetospheres and incident on exomoons orbiting Jupiter-analogs ($\Phi_{\rm GP}$); if we were to consider a Saturn-analog instead, the value of $\Phi_{\rm GP}$ becomes much lower and depends on complex magnetospheric physics not considered here. From the data provided in Cooper et al. (2001), we obtain

$$\Phi_{\rm GP} \sim 4 \times 10^6 \,{\rm J\,m^{-2}\,yr^{-1}} \left(\frac{a_m}{4.5 \times 10^{-3}\,{\rm AU}} \right)^{-2}, \qquad (7.27)$$

where a_m is the orbital distance of the moon from the giant planet, the normalization factor 4.5×10^{-3} AU represents the Europa-Jupiter distance, and the energetic particle flux is expected to obey an inverse square-law behavior. Using the higher range of the G-values from Kobayashi et al. (1995) in conjunction with (7.27), we find

$$\mathcal{M}_{\rm GP} \sim 1.9 \times 10^8 \,{\rm kg/yr} \left(\frac{R}{R_\oplus} \right)^2 \left(\frac{a_m}{4.5 \times 10^{-3}\,{\rm AU}} \right)^{-2}, \qquad (7.28)$$

with $\mathcal{M}_{\rm GP}$ representing the mass rate of amino acids synthesized during the bombardment of the icy moon's surface by energetic particles from the giant planet's magnetosphere.

The SEP energy flux is not straightforward to evaluate for a number of reasons. First, the physics behind the generation of SEPs is complex and the integrated fluence will depend on both the particle acceleration mechanisms

and the sites of origin. Second, the SEP flux is highly variable because it depends on the stellar age, mass, and rotation. This flux is expected to be significant for low-mass stars with high activity and close-in planets but may not be a dominant source for middle-aged Solar-type stars. As a consequence, one cannot find a ready parameterization for the SEP energy flux in terms of basic stellar and planetary parameters.

We turn our attention to GCRs, another important source of energetic particles. Their energy flux $\Phi_{\rm CR}$ near Earth is known to be

$$\Phi_{\rm CR} \sim 4.6 \times 10^2 \,{\rm J\,m^{-2}\,yr^{-1}}. \tag{7.29}$$

By comparing the above equation with (7.23), we see that the GCR energy flux is approximately equal to the CMB energy flux at $z = 0$. Although (7.29) comes across as being rather small, it is worth recalling that the cosmic-ray flux increases toward the Galactic center. More importantly, it constitutes one of the few energy sources that is universally accessible to Type B and U planets. The corresponding mass $\mathcal{M}_{\rm CR}$ of amino acids produced per unit time is

$$\mathcal{M}_{\rm CR} \sim 1.7 \times 10^4 \,{\rm kg/yr} \left(\frac{R}{R_\oplus}\right)^2, \tag{7.30}$$

where we have utilized the upper bound for the G-value presented in Kobayashi et al. (1995).

The next major source of prebiotic compounds worth considering is the exogenous delivery of organic molecules via interplanetary dust particles (IDPs), comets, and meteorites. A classic overview of this pathway and the corresponding organic yields on Earth can be found in Chyba and Sagan (1992). The delivery rates of organics when it comes to IDPs are ~ 3 and ~ 5 orders of magnitude greater than comets and meteorites, respectively, although the latter are rich in organic compounds—particularly those meteorites belonging to a class called carbonaceous chondrites. Furthermore, it is evident that the exogenous delivery of prebiotic compounds via comets and meteorites will not be applicable to Type U planets.

Hence, we shall restrict our attention to the exogenous delivery of organics via IDPs. A crucial point worth emphasizing is that the "soft landings" of IDPs on the planetary surface typically requires a moderately thick atmosphere, unlike our worlds with atmospheres that are either absent or rarefied. Hence, the efficiency, and even the viability, of this pathway

remains an open question. However, for the sake of completeness, it would be instructive to estimate the mass of organic compounds delivered by IDPs. To do so, we must first determine the rate at which Type B and U worlds accrete IDPs. Clearly, formulating a universal mass accretion rate for all Type B and U worlds is not readily achievable. Instead, we propose that the following expression for the mass accretion rate $\dot{M}$ serves as a reasonable approximation:

$$\dot{M} \approx 4\pi R_{\max}^2 \rho_d \sigma, \tag{7.31}$$

where $\sigma = \sqrt{V^2 + c_s^2}$, ρ_d denotes the density of the ambient dust particles, c_s is the sound speed, and V represents the relative velocity between the object and the dust. Note that $R_{\max} = \max\{R, R_B\}$, with R_B denoting the modified Bondi radius defined as $R_B = GM/\sigma^2$. The accretion of IDPs will occur even when the dust grains and bolides are decoupled from the gas after the solar nebula has been dispersed. In the event that there is no gas left, σ must be replaced with V, and R_B becomes the Hoyle-Lyttleton radius in this regime.

As per this formula, for $R_{\max} = R$, we obtain the geometric mass accretion rate. In contrast, when $R_{\max} = R_B$, we arrive at the Bondi-Hoyle-Lyttleton accretion rate used in many fields of astronomy. To proceed further, we shall assume that the accretion rate of organics from IDPs, represented by $\mathcal{M}_{\rm DP}$, has a linear dependence on $\dot{M}$. To find the normalization factor, we must choose an organic deposition rate for Earth, but this task is complicated by the fact that $\mathcal{M}_{\rm DP}$ has varied by three orders of magnitude during the history of the Earth. Hence, we choose the fiducial value of $\sim 10^8$ kg / yr, which conceivably represents the deposition rate of intact organics (via IDPs) at ~ 4 Ga, as there are some grounds to believe that life may have originated during this period. Consequently, we find that $\mathcal{M}_{\rm DP}$ becomes

$$\mathcal{M}_{\rm DP} \sim 10^8 \,{\rm kg\,/\,yr} \left(\frac{R}{R_\oplus}\right)^{2\beta} \left(\frac{\sigma}{26\,{\rm km\,/\,s}}\right)^{-3} \left(\frac{\rho_d}{2\times 10^{-22}\,{\rm kg\,/\,m^3}}\right), \tag{7.32}$$

in the event that $R_{\max} = R_B$ is valid. The dust density near the heliosphere is specified to be $\sim 2 \times 10^{-22}$ kg / m^3; this choice is about two orders of magnitude higher than the present-day value to reflect the greater spatial density of interplanetary dust during this epoch. The relative inflow velocity,

which dominates over the sound speed, of the dust is ~ 26 km / s in the Solar neighborhood. Although the values for σ and ρ_d in the interstellar medium will be quite different, we anticipate that the radius will play the most dominant role in governing the magnitude of $\mathcal{M}_{\rm DP}$ on account of its high power-law exponent. In contrast, for the case $R_{\rm max} = R$, we find that $\mathcal{M}_{\rm DP}$ is

$$\mathcal{M}_{\rm DP} \sim 10^8 \,{\rm kg / yr} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\sigma}{26\,{\rm km / s}}\right) \left(\frac{\rho_d}{2 \times 10^{-22}\,{\rm kg / m^3}}\right). \quad (7.33)$$

Before proceeding further, it must be emphasized that $\mathcal{M}_{\rm DP}$ quantifies the *total* amount of organics delivered (not just amino acids), and therefore it constitutes an upper bound.

The last major source that we take into consideration is the abiotic synthesis of amino acids from alkaline hydrothermal vents. These environments have attracted much attention as putative sites of abiogenesis, and we discussed them in detail in Section 2.7.1, on the origin of life. One of the desirable features associated with hydrothermal vents is that they may enable the abiotic synthesis and polymerization of prebiotic compounds (W. Martin et al. 2008; Sojo et al. 2016). We will, therefore, restrict ourselves to only computing the putative yield of amino acids at these sites.

Before doing so, let us recall that hydrothermal vents are contingent on the existence of water-rock interfaces. As we have stated earlier, this is a rather nontrivial requirement because high pressures at the bottom of the ocean will result in the formation of high-pressure ices. As a result, the subsurface ocean would become sandwiched between two ice layers, and this outcome is conventionally regarded as having a negative impact on the habitability of such worlds. Yet, such an internal structure need not necessarily spell doom and gloom as per certain studies (Choblet, Tobie, Sotin, Běhounková et al. 2017; Kalousová et al. 2018). Despite the presence of high-pressure ices at the ocean floor, a combination of convection and melting could enable the slow transport of salts and nutrients, and this mechanism has been predicted to operate on Ganymede.

However, we will assume henceforth that water-rock interactions are feasible and that hydrothermal vents are present at the floor of the subsurface ocean. To compute the abiotic yield of amino acids is not an easy task, since there are a large number of environmental factors that influence it. One such example is the water activity, since laboratory experiments have

indicated that the rate of serpentinization (which produces H_2) decreases with an increase in the salinity. Another difficulty stems from the fact that nucleobases and amino acids have short half-lives at high temperatures and pressures (Levy & Miller 1998; Aubrey et al. 2009); therefore the production rates computed below do not account for the decomposition of organics. With these caveats in mind, the first quantity that we seek to determine is the amount of molecular hydrogen produced through serpentinization.

Let us denote the production rate of H_2 by $\mathcal{N}_{H_2}$ in units of mol / yr. We make use of the simple model outlined in Vance et al. (2016) and E. L. Steel et al. (2017), wherein $\mathcal{N}_{H_2}$ is given by

$$\mathcal{N}_{H_2} = \epsilon \, V_R, \tag{7.34}$$

where V_R is the volume of the region undergoing serpentinization and ϵ represents a conversion factor that accounts for how many moles of H_2 is produced per unit volume. Among other factors, ϵ will depend on properties of the mineral(s) subject to serpentinization. The volume V_R can be further expressed as

$$V_R = \frac{4\pi}{3}\left[R_c^3 - (R_c - \langle z \rangle)^3\right] \approx 4\pi R_c^2 \langle z \rangle, \tag{7.35}$$

provided that $\langle z \rangle \ll R_c$. Here, R_c is the radius of the core region (comprising silicates and / or metals) and $\langle z \rangle$ represents the width of the serpentinization front. If we assume that the front spreads through diffusion, it can be approximated via the root mean square diffusion distance,

$$\langle z \rangle \approx \sqrt{2 D_s t}, \tag{7.36}$$

where D_s is the diffusion constant for the advancement of the serpentinization reaction front, and t stands for the elapsed time. We use the scaling $R_c \propto R$ because of its validity for terrestrial and icy worlds. Upon combining all of the above equations, we can see that the scaling $\mathcal{N}_{H_2} \propto R^2$ follows as a result, although $\mathcal{N}_{H_2}$ also depends on several other factors. If we invoke the assumption that all of the parameters on which $\mathcal{N}_{H_2}$ depends, barring R, are similar to the icy worlds in our Solar system (Vance et al. 2007), we arrive at

$$\mathcal{N}_{H_2} \sim 2.7 \times 10^{11} \text{ mol / yr} \left(\frac{R}{R_\oplus}\right)^2. \tag{7.37}$$

We can undertake some simple calculations to assess the accuracy of this formula. In the case of Enceladus, we obtain $\mathcal{N}_{H_2} \sim 4 \times 10^8$ mol / yr, and it matches closely with the value $\mathcal{N}_{H_2} \sim 10^9$ mol / yr determined from *Cassini* observations of the Enceladus plume (Waite et al. 2017). In the case of Europa, we find $\mathcal{N}_{H_2} \sim 1.6 \times 10^{10}$ mol / yr, which is in excellent agreement with the theoretical value of $\mathcal{N}_{H_2} \sim 10^{10}$ mol / yr obtained by Vance et al. (2016). Lastly, the corresponding value for the Earth obtained from (7.37) is approximately equal to the production rate of H_2 from water-rock reactions ($\sim 10^{11}$ mol / yr), based on measurements at multiple Precambrian sites (Sherwood Lollar et al. 2014).

Next, we wish to calculate the mass rate of *abiotic* amino acids produced ($\mathcal{M}_{\rm HV}$) from hydrothermal vents. In most serpentinization experiments, glycine is the most commonly produced amino acid at lower temperatures; it is produced via the following hydrothermal reaction:

$$2\,CO_2 + NH_4^+ + 3\,H_2 \rightarrow \text{glycine} + H^+ + 2\,H_2O \tag{7.38}$$

By utilizing the fact that $\mathcal{M}_{\rm HV}$ will be proportional to $\mathcal{N}_{H_2}$ and making use of the data from E. L. Steel et al. (2017) for determining the proportionality constant, we arrive at

$$\mathcal{M}_{\rm HV} \sim 6.7 \times 10^8\,\text{kg / yr} \left(\frac{R}{R_\oplus}\right)^2. \tag{7.39}$$

7.4.2 Further steps toward abiogenesis

Hitherto, our focus has only been on documenting the various energy sources and the yields of certain organic molecules (amino acids in our case) for different prebiotic pathways. Yet, even after biomolecules such as amino acids and nucleobases are formed, there is a long way to go before proteins and nucleic acids, respectively, can be assembled. Even after these biopolymers have been synthesized, many steps still lie ahead before the first lifeforms emerge. In view of the fact that most of these steps are poorly understood even on Earth, making quantitative statements for subsurface ocean worlds is very difficult. With the exception of organics produced at hydrothermal vents and through radiolysis, all of the other energy sources described in Section 7.4.1 result in the formation of organic molecules on, or near, the icy surface. If these molecules are not transported

into the interior, they will be subject to decomposition by sputtering, electromagnetic radiation, and charged particles. Hence, we will briefly describe how the delivery of the aforementioned organic molecules to the subsurface ocean may occur and the potential transformations that might be feasible along the way.

To begin with, the vertical transport of the organics into the ocean could take place via three different channels. The first entails impact gardening, a nonlinear process that unfolds through the bombardment of the surface by micrometeorites and enables the burial of organics into deeper layers. The next two mechanisms rely on vertical mixing driven by plate tectonics and volcanism, but these phenomena are rarely present on icy worlds. Europa is an intriguing exception, since it appears to display evidence of both subduction and cryovolcanism (Sparks et al. 2017). Another point worth highlighting is that gardening would be largely absent on Type U worlds because the flux of micrometeorites is negligible in comparison to Type B worlds, unless the former are located in unusually dense interstellar environments.

One of the most significant challenges of prebiotic chemistry is that the appropriate organic compounds must occur in sufficiently high concentrations to undergo chemical reactions, as explained in Chapter 2. Even if the concentrations of these molecules are high enough, they must undergo polymerization to eventually yield peptides and nucleic acids, among other biopolymers, without forming "tar" in the process (Budin & Szostak 2010). A wide range of environments have been proposed on Earth (e.g., beaches, intermountain valleys), as summarized in Section 2.7, but most of them are absent on subsurface worlds. Thus, prima facie, it would appear as though we have encountered an insurmountable snag in enabling the origin of life.

However, as we encountered in Section 2.4.2, ice is an unusual, and unusually effective, medium when it comes to prebiotic chemistry. Several laboratory experiments have demonstrated that ice furnishes a viable environment for concentrating organic compounds (Trinks et al. 2005). In addition, freeze-thaw cycles are helpful in the formation of RNA polymerase ribozymes because ice possesses certain stabilizing properties. However, what remains unclear is whether these phenomena are sufficiently rapid on a global scale to facilitate the concentration and polymerization of these molecules. We have also implicitly assumed here that the requisite minerals (for catalysis) and bioessential elements are present in reasonable amounts, but neither of them can be taken for granted. Yet, despite these

potential issues, we ought not rule out the possibility that ice is one of the viable sites for abiogenesis on these worlds.

To quantify the requirements for polymerization, we evaluate the conditions under which the desired chemical reactions become exergonic—namely, the criterion $\Delta G < 0$ must be valid, where ΔG denotes the Gibbs free energy of formation. For a more detailed discussion of such an approach, the reader may consult Kimura and Kitadai (2015), where the constraints on the polymerization of amino acids and nucleotides are evaluated. It was found that ΔG depends weakly on the pressure and becomes negative for $T_p \sim 50 - 110$ K, depending on the polymerization reaction under consideration. If we assume $T_s < 50$ K to be on the safe side, we can solve for the depth H_p below the icy surface at which T_p will be attained by employing (7.7). By solving for H_p, the final solution is given by (7.9), except for the fact that Λ must be replaced by $\Lambda_p = T_p/T_s$. For subsurface ocean worlds not greatly smaller than the Earth, the following approximation is valid:

$$\frac{H_p}{H} \approx \frac{\ln \Lambda_p}{\ln \Lambda}. \tag{7.40}$$

As an example, let us consider a Type U world with a geothermal heat flux and radius similar to the Earth's. For these parameters, we end up with $\ln \Lambda_p \approx 0.36 - 1.15$ and $H_p \approx 2.7 - 8.6$ km. Hence, at a depth of a few kilometers, the formation of peptides would become favorable on thermodynamic grounds. An obvious implication of the above equation is that H_p decreases as one moves toward the central regions of the Galaxy, where T_s could be higher. Although the fact that these reactions are exergonic should signify that polymerization will occur, the rate at which they occur is a completely different matter. As per the Arrhenius equation, the reaction rates are proportional to the Boltzmann factor $\exp(-E_a/k_B T)$, where E_a is the activation energy and T is the ambient temperature. Hence, accounting for the datum that low values of T are commonplace on icy worlds, polymerization may only take place at very low rates.

Laboratory experiments suggest that a monomer concentration of ~ 0.1–1 mM (mmol / L) might be necessary for initiating prebiotic self-assembly processes (Budin & Szostak 2010).[11] Let us choose the lower

11. One liter of water has a mass of approximately 1 kg, owing to which mM is loosely equivalent to mmol/kg in the literature. Later, we will use the second definition for the sake of convenience.

bound as the critical value ($\zeta \sim 0.1$ mM) and ask ourselves what the net delivery rate of amino acids (denoted by $\mathcal{M}_t$) ought to be for ensuring that ζ is achieved in a timescale t_g that is on the order of Gyr. Using the fact that $\mathcal{M}_t \sim (0.1\text{kg/mol})\, \zeta\, V_{\text{ocean}}/t_g$,[12] where V_{ocean} is the total volume of the subsurface ocean, we end up with

$$\mathcal{M}_t \sim 3.8 \times 10^6 \,\text{kg}\,/\,\text{yr} \left(\frac{\zeta}{0.1\,\text{mM}}\right)\left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\mathcal{H}}{1\,\text{km}}\right)\left(\frac{t_g}{1\,\text{Gyr}}\right)^{-1}. \tag{7.41}$$

By comparing this estimate with the rates presented in Section 7.4.1, it might seem plausible that amino acid concentrations reach the desired level within Gyr timescales. However, a crucial caveat is that we do not know the magnitudes of the decomposition rates, as they will counteract the production rates of monomers. The timescale t_g for achieving sufficiently high concentrations can be found by rearranging (7.41) if an estimate for $\mathcal{M}_t$ is available. We point out, however, that the actual value of t_g could be lower than the theoretical estimate due to the operation of thermodynamic cycles and the presence of mineral catalysts that are sensitive to the nature of the available microenvironments. Hence, given our analysis thus far, it may be possible for prebiotic compounds to be concentrated, polymerized, and delivered to the subsurface ocean underneath the ice envelope, where they undergo subsequent prebiotic evolution and perhaps result in abiogenesis.

We will round off the analysis of prebiotic chemistry with a couple of points worth bearing in mind as we progress ahead.

- Ice may not only provide advantages for abiogenesis to occur but also provide a habitat for certain microbes. On Earth, specialized microbes (called psychrophiles) have evolved many unique genotypic and phenotypic characteristics in order to inhabit a wide range of sea-ice habitats (A. Martin & McMinn 2018). As a result, it has even been suggested that some of these sea-ice microbes could be capable of surviving on icy worlds. Hence, in the future, sampling the sea-ice in Europa and Enceladus might reveal the existence of microorganisms.

---

12. The conversion factor of 0.1 kg / mol has been included to convert the delivery rate that would be expressed in terms of mol / yr into kg / yr instead.

- There are innumerable factors that we have not dealt with here: one such example is the self-organization of amphiphilic compounds into vesicles. In light of the complexity of the transition from nonlife to life, it seems plausible that no single microenvironment will have all of the ingredients required for abiogenesis from the beginning. Instead, what may be essential is the collective action of multiple environments and mechanisms in order to facilitate the origin of life on these worlds with subsurface oceans.

7.5 ECOSYSTEMS IN PLANETS WITH SUBSURFACE OCEANS

Life is not expected to have access to sunlight in subsurface ocean worlds, because of their very nature. On Earth, solar radiation is widely considered to be the predominant energy source, and this is reflected in the fact that the majority of Earth's biomass is dependent, either directly or indirectly, on photosynthesis. Hence, we may be predisposed to believe that the biological potential of subsurface worlds is extremely low relative to Earth (Jakosky & Shock 1998; Gaidos et al. 1999). While this surmise is conceivably valid, we shall discover that a wide range of ecosystems are nevertheless possible on subsurface ocean worlds when viewed purely through the prism of energetics (Chyba & Hand 2001; R. M. Jones et al. 2018). On the other hand, if we take biogeochemical considerations into account, the status quo might change greatly, as elucidated further in Section 7.6.

Before embarking on our discussion, let us recall that there are multiple constraints enforced by the likes of pH, salinity, and pressure on aquatic habitats (Kargel et al. 2000; Jebbar et al. 2020). For instance, it is suspected that Europa's ocean has a high supply of oxidants delivered from the surface, eventually resulting in a highly acidic ocean with a pH of ~ 2.6 when they react with sulfides (Pasek & Greenberg 2012). At this level of acidity, many factors including biomineralization, which enables the formation of eggshells and bones, among other things, could end up becoming severely disrupted. On the other hand, current models suggest that the pH may range between ~ 2 and ~ 6, based on the available empirical and theoretical constraints (Bouquet et al. 2019).

The first energy source that we consider is the delivery of oxidants from the icy surface to its ocean, and this point is rendered particularly important for moons in intense radiation environments (e.g., Europa). Oxygen

has had a transformative effect on Earth's biosphere (Lane 2002), and there are fairly sound theoretical and empirical grounds for believing that aerobic metabolism releases about an order of magnitude more energy than anaerobic metabolism given the same supply of raw materials. Thus, several scientists have posited that the rise in molecular oxygen (O_2) levels constitutes a vital rate-limiting step in the evolution of extraterrestrial life (Catling et al. 2005; Knoll 2015). Although the importance of oxygen is well established, it is also worth recalling that it forms peroxides and superoxides that can destroy organic compounds. As a result, bacteria had to evolve special evolutionary adaptations in order to offset the negatives.

In Section 7.4.1, we concluded that there will be a steady flux of energetic particles to the surface from giant planetary magnetospheres, SEPs, and GCRs. Let us denote the total flux, i.e., the sum of the individual three fluxes—by Φ_{T}. For Type B moons around Jupiter-like planets, we anticipate that $\Phi_{\mathrm{T}} \approx \Phi_{\mathrm{GP}}$, while $\Phi_{\mathrm{T}} \approx \Phi_{\mathrm{CR}}$ for Type U worlds. The basic setup for the delivery of oxidants to the subsurface ocean is as follows. High-energy particles from previously delineated sources facilitate the formation of clathrate hydrates (crystalline icelike solids) of oxidants, e.g., H_2O_2, O_2, and CO_2, through radiolysis on the surface (R. E. Johnson et al. 2003). Subsequently, gardening and geological activity enable these compounds to be thrust into the lower layers and eventually delivered to the subsurface ocean. In the latter environment, it has been hypothesized that they can sustain an (indirect) radiation-driven ecosystem (Chyba & Philips 2001).

As there are several stages involved in this process, there is no guarantee that these oxidants will ultimately reach the ocean. The rates of sputtering and gardening are subject to uncertainties, and this is also true with regard to the transport rate of oxidants due to geological activity. Owing to all these factors, estimates for the depth of the oxygenation layer and the concentration of surficial radiolytic products vary considerably, as do the theoretical rates of O_2 delivered to the Europan ocean—the latter span six orders of magnitude, from $\sim 10^5$ mol / yr to $\sim 10^{11}$ mol / yr. We can estimate the delivery rate $\mathcal{N}_{O_2}$ by adopting the premise that gardening is primarily responsible for the transport of oxidants and using the scaling relations in Hand et al. (2007), which leads us to

$$\mathcal{N}_{O_2} \sim \frac{4\pi R^2 d_g C_0}{\tau_d}. \tag{7.42}$$

Here, d_g is the gardening depth, C_0 is the molar concentration of oxidants, and τ_d is the delivery time. The ansatz $C_0 \propto \Phi_{\rm GP}$ appears to be reasonable since a higher particle flux should lead to the deposition of more oxidants on the surface. We also make use of the empirical scaling $d_g \propto \tau_d^{1/2}$ (Cooper et al. 2001). Substituting these expressions into the above equation and using the fiducial values of $\mathcal{N}_{O_2} \sim 10^9$ mol / yr and $\tau_d \sim 50$ Myr for Europa, we obtain

$$\mathcal{N}_{O_2} \sim 1.7 \times 10^{10}\,\text{mol / yr} \left(\frac{\tau_d}{50\,\text{Myr}}\right)^{-1/2} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a_m}{4.5 \times 10^{-3}\,\text{AU}}\right)^{-2} \tag{7.43}$$

after making use of (7.27). On the other hand, if the icy worlds possess either active tectonics or volcanism (or both), it is more transparent to work with the variable $\delta_g = d_g/\tau_d$ that equals 4 m / Myr for Europa, and the characteristic value of $\mathcal{N}_{O_2}$ is $\sim 10^{11}$ mol / yr (Greenberg 2010). Using this information in (7.42), we end up with

$$\mathcal{N}_{O_2} \sim 4.3 \times 10^{11}\,\text{mol / yr} \left(\frac{\delta_g}{1\,\text{m / Myr}}\right) \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a_m}{4.5 \times 10^{-3}\,\text{AU}}\right)^{-2}. \tag{7.44}$$

It is, however, crucial to recognize that (7.43) and (7.44) as well as the subsequent estimate of $\mathcal{N}_{O_2}$ in (7.46) may represent an overestimate for Europa by ~ 2 to 3 orders of magnitude as per the latest numerical models (Oza et al. 2019). In this scenario, the corresponding steady-state concentration of oceanic O_2 would have to be scaled downward accordingly.

Next, we remarked earlier that long-lived radioactive elements are responsible for contributing to radiogenic heating. They also play an important secondary role: a combination of alpha, beta, and gamma decay processes powered by ^{40}K, ^{232}T, ^{235}U, and ^{238}U enables the radiolytic dissociation of water, thereby leading to the formation of oxygen and hydrogen. On our planet, the collective radiolysis of water is ostensibly responsible for the production of $\sim 2 \times 10^{10}$ mol / yr of H_2, based on data collected from the Precambrian crust (Sherwood Lollar et al. 2014), and approximately half this amount of oxygen. For the purpose of our order-of-magnitude estimates, we invoke the potentially reasonable assumption that the amount of H_2 produced, the total mass of radioisotopes, and the mass of the ocean are linearly proportional to one another. We reintroduce the radionuclide

enhancement factor Γ to account for the variation in radioisotope concentrations. In reality, the situation is more complex because $\mathcal{N}_{H_2}$ will depend on the extent and properties of water-rock interactions (e.g., rock porosity and density) that are not precisely known, even for Enceladus and Europa. Nevertheless, by proceeding with our simpler ansatz outlined above, the radiolytic production rate of H_2 can be expressed as

$$\mathcal{N}_{H_2} \sim 5.4 \times 10^9 \,\mathrm{mol\,/\,yr}\ \Gamma \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\mathcal{H}}{1\,\mathrm{km}}\right), \tag{7.45}$$

and the corresponding production rate of oxidants $\mathcal{N}_{O_2}$ through radiolysis becomes

$$\mathcal{N}_{O_2} \sim 2.7 \times 10^9 \,\mathrm{mol\,/\,yr}\ \Gamma \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\mathcal{H}}{1\,\mathrm{km}}\right). \tag{7.46}$$

We have used $H/R \ll 1$ and $\mathcal{H}/R \ll 1$ to derive the above equations. These two inequalities are likely to be less accurate when dealing with small worlds like Enceladus. Using the datum that $\mathcal{H} \sim 30$ km for Enceladus, we end up with $\mathcal{N}_{H_2} \sim 2.5 \times 10^8$ mol / yr. This estimate displays excellent agreement with $\mathcal{N}_{H_2} \sim 1\text{–}3 \times 10^8$ mol / yr, derived by means of a detailed radiolysis model for Enceladus (Waite et al. 2017).

Hence, we have delineated two distinct channels by which H_2 can be synthesized: hydrothermal vents and water radiolysis, whose production rates are given by (7.37) and (7.45), respectively. Along similar lines, we have identified two mechanisms for the production of oxidants, i.e., the delivery of oxidants from the surface and water radiolysis. The former is determined from (7.43) or (7.44), depending on the context, while the latter is given by (7.46). As there are several free parameters and processes, it becomes apparent that we cannot easily gauge the characteristic values for these pathways. However, inspecting the preceding formulae reveals that the rates of production of oxygen and hydrogen are not very dissimilar, i.e., they differ by less than a couple of orders of magnitude in many cases. In turn, an approximate redox balance may exist on some of these worlds, analogous to Earth and possibly Europa (Vance et al. 2016). It will thus be necessary for future studies to take the long-term redox history into account to properly assess the habitability of Type B and U worlds.

We are now in a position to model the biomass that can be supported by the various energy sources reported hitherto. In the case of oxidants, we will draw on the procedure outlined in Chyba and Philips (2001). The putative microbes considered in this analysis are assumed to resemble terrestrial methanotrophs that oxidize methane to produce carbon dioxide by way of methanol (CH_3OH), formaldehyde (HCHO), and formate ($HCOO^-$). The oxidation of HCHO by methanotrophs yields 4.7 eV per molecule. Multiplying the energy per molecule with the total number of molecules produced per year—we choose a fiducial value of $\mathcal{N}_{O_2} \sim 10^{10}$ mol / yr—leads us to the total energy of $\sim 4.5 \times 10^{15}$ J / yr. We multiply this number with the efficiency of biomass production, which is typically $\sim 2.4 \times 10^{-7}$ kg / J. Hence, assuming that all of the energy is used in biomass production, we end up with a total biomass production rate of 1.1×10^9 kg / yr. Since this estimate depends linearly on $\mathcal{N}_{O_2}$, we obtain

$$\frac{dm_0}{dt} \sim 1.1 \times 10^9 \,\mathrm{kg / yr} \left(\frac{\mathcal{N}_{O_2}}{10^{10} \,\mathrm{mol / yr}} \right), \tag{7.47}$$

with dm_0/dt denoting the rate of production of biomass. Assuming a turnover time of $\sim 10^3$ yr, based on experimental studies of Earth's deep biosphere (Hoehler & Jørgensen 2013; Lever et al. 2015),[13] the steady-state biomass (m_c) is given by

$$m_c \sim 1.1 \times 10^{12} \,\mathrm{kg} \left(\frac{\mathcal{N}_{O_2}}{10^{10} \,\mathrm{mol / yr}} \right). \tag{7.48}$$

Alternatively, we can compute the corresponding number of cells and their rate of production (in cells / yr) by using the datum that the dry mass of each cell is $\sim 2 \times 10^{-17}$ kg.

We estimate the amount of biomass that can be sustained through the influx of reductants in a similar fashion. The microbes under consideration are known as methanogens, which are reliant on the reduction of carbon dioxide to produce methane in the following fashion:

$$CO_2 + 4\,H_2 \rightarrow CH_4 + 2\,H_2O. \tag{7.49}$$

---

13. However, in extreme environments (e.g., ice and permafrost) at ~ 230 K, the turnover time could prove to be several orders of magnitude higher.

Although evident, it is worth remarking that the total biomass computed for oxidants and reductants is clearly contingent on the actual existence of analogous methanotrophs and methanogens endowed with these particular metabolic pathways in the subsurface oceans. To leading order, since the biomass produced per year is proportional to $\mathcal{N}_{H_2}$ (E. L. Steel et al. 2017), we arrive at

$$\frac{dm_0}{dt} \sim 2 \times 10^7 \,\mathrm{kg\,/\,yr} \left(\frac{\mathcal{N}_{H_2}}{10^{10}\,\mathrm{mol\,/\,yr}} \right), \tag{7.50}$$

and the corresponding steady-state biomass is expressible as

$$m_c \sim 2 \times 10^{10}\,\mathrm{kg} \left(\frac{\mathcal{N}_{H_2}}{10^{10}\,\mathrm{mol\,/\,yr}} \right). \tag{7.51}$$

Apart from these putative ecosystems, other forms of life that employ alternative sources of energy (e.g., thermal, electrical, or gravitational) might also exist on Type U and B worlds, but we shall not analyze them here since their total biomass is probably lower (Reynolds et al. 1983; Schulze-Makuch & Irwin 2018). However, another ecosystem worth contemplating is based on organisms that derive energy from electrical currents by means of electron capture compounds that can enable electron transfer reactions. Although this mechanism has not been documented for any organisms on our planet, electron capture compounds suitable for this purpose appear to exist in terrestrial biology. The secondary electrons required for powering these reactions are assumed to be produced by the magnetospheres of giant planets; this process has been dubbed "direct electrophy" (Stelmach et al. 2018). As a consequence, it is apparent that this near-surface ecosystem would be feasible only for (Type B) moons around Jovian planets.[14] Using the data tabulated in Stelmach et al., the maximum steady-state biomass is

$$m_c \sim 10^{11}\,\mathrm{kg} \left(\frac{R}{R_\oplus} \right)^2 \left(\frac{\Phi_e}{10^{10}\,\mathrm{m^{-2}\,s^{-1}}} \right), \tag{7.52}$$

with Φ_e denoting the electron number flux received at the surface of the moon, assuming that the average energy of the particles is approximately 0.5 MeV.

14. This ecosystem must be situated close to the icy surface because the energetic electrons will not penetrate very deep into the ice layer.

As a point of comparison, we observe that the global net primary production (NPP) of the Earth is $\sim 10^{14}$ kg / yr and the total biomass is $\sim 2 \times 10^{15}$ kg.[15] Thus, by comparing the latter quantity with any of the above biomass estimates, the biological productivity on subsurface ocean worlds appears to be orders of magnitude lower than that of the Earth. The steady-state concentration of cells (η_c) in the ocean can be estimated by dividing m_c with the mass of an individual cell and the volume of the ocean. Therefore, we end up with

$$\eta_c \sim 10^9 \,\text{cells/m}^3 \left(\frac{m_c}{10^{10}\,\text{kg}}\right)\left(\frac{R}{R_\oplus}\right)^{-2}\left(\frac{\mathcal{H}}{1\,\text{km}}\right)^{-1}, \tag{7.53}$$

and this value is similar to the cellular densities observed in other extreme habitats on Earth, some of which are delineated below.

- Subglacial lakes and other icy environments (e.g., Vostok and Grímsvötn) are estimated to have concentrations of $\sim 10^9 - 10^{10}$ cells / m^3, although most of these values are based on indirect experimental evidence. It should, however, be noted that densities as low as $\sim 10^6 - 10^7$ cells / m^3 are cited in the literature for Lake Vostok.
- The microbial concentrations in granitic rock groundwater, located at the depth of a few kilometers below the surface, are typically on the order of $\sim 10^{10}$–10^{12} cells / m^3.
- In highly oligotrophic habitats, such as the sub-seafloor sediments of the North Pacific Gyre, microbial densities of $\sim 10^9$ cells / m^3 have been documented.

Until this juncture, our discussion has been centered on the potential existence of microbial ecosystems. Yet, the possibility of larger organisms (macrofauna) cannot be dismissed outright. An inspection of (7.43), (7.44), and (7.46) reveals that $\mathcal{N}_{O_2} \sim 10^9$–$10^{11}$ mol / yr for Earth-sized subsurface ocean worlds. Next, we note that the level of dissolved oxygen in Earth's ocean is around 0.25 mM. We can therefore compute the time t_{O_2} that would be required for the oxidant levels to reach this value in the subsurface

15. The NPP is a measure of the total amount of biomass on our planet, and it quantifies the difference between the rates of C fixed by photosynthesis and used up by respiration.

ocean, which is given by

$$t_{O_2} \sim 10^7 \text{ yrs} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\mathcal{H}}{1\,\text{km}}\right) \left(\frac{\mathcal{N}_{O_2}}{10^{10}\,\text{mol}\,/\,\text{yr}}\right)^{-1}. \tag{7.54}$$

Thus, depending on the values of R, $\mathcal{H}$, and $\mathcal{N}_{O_2}$, some of the subsurface ocean worlds could attain oxygen levels that are comparable to the Earth over a period of time < 1 Gyr. Now, in order to compute the biomass of macrofauna that are sustainable in subsurface oceans, we shall assume that they possess energetic requirements similar to fish. Using the fact that a single fish requires about ~ 100 mol yr^{-1} kg^{-1} (Greenberg 2010), the total steady-state mass of macrofauna (M_{bio}) that are supportable by the delivery of oxidants is

$$M_{\text{bio}} \sim 10^8\,\text{kg} \left(\frac{\mathcal{N}_{O_2}}{10^{10}\,\text{mol}\,/\,\text{yr}}\right), \tag{7.55}$$

assuming that all of the available O_2 is consumed by the putative macrofauna and that none of the oxygen is utilized by microbes. Interestingly, if we compare the above estimate with (7.48), the latter is roughly four orders of magnitude higher, thereby implying that most of the biomass in subsurface ocean worlds may consist of microbes. The mass density of macrofauna Σ_{bio} can be estimated via $\Sigma_{\text{bio}} \sim M_{\text{bio}}/(4\pi R^2)$ after applying the reasonable approximation $H/R \ll 1$; Σ_{bio} is measured in terms of mass per unit area (and not volume). Solving this equation for Σ_{bio} after invoking (7.55) yields

$$\Sigma_{\text{bio}} \sim 2 \times 10^{-7}\,\text{kg}\,/\,\text{m}^2 \left(\frac{\mathcal{N}_{O_2}}{10^{10}\,\text{mol}\,/\,\text{yr}}\right) \left(\frac{R}{R_\oplus}\right)^{-2}, \tag{7.56}$$

and we can further convert the expression into the average number density of macrofauna by dividing it with the typical mass of an individual organism. However, since this mass remains completely unknown, we shall bring our analysis to a close and delve into the next topic.

7.6 BIOESSENTIAL ELEMENTS AND SUBSURFACE OCEAN WORLDS

Most studies of habitability begin and end with the prospects for hosting liquid water on the surface. In contrast, we have investigated the likelihood of life in Sections 7.4 and 7.5 by adopting a "follow the energy" approach

(Hoehler et al. 2007; R. M. Jones et al. 2018). Yet, an important point that merits highlighting at this stage is that life requires much more than an energy source and a solvent. Among other things, it needs a suitable supply of chemical elements for carrying out the basic functions of metabolism and replication. In particular, the elements carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur are regarded as essential ingredients of life-as-we-know-it. On account of this reason, they are often grouped under the category of *bioessential* elements.

One of the unique features of life is that it has found ways to exploit nearly all of the chemical elements available on Earth for a variety of biological processes (Anbar 2008). Consider, for example, the element boron: borate minerals were conceivably a major component of the pathway(s) leading to abiogenesis on our planet through stabilizing ribose, the carbohydrate that forms the backbone of RNA. In addition, it could aid in preventing the formation of organic "tar" by selectively inhibiting the reactivity of different organic compounds. Next, let us turn our attention to the element molybdenum, which can occur in the form of molybdate minerals. The latter have been argued to play an essential role in the formation of ribose by functioning as catalysts. Hence, if the RNA world hypotheses outlined in Section 2.5 are proven to be correct, the availability of borates and molybdates may have constituted key (albeit not irreplaceable) desiderata for the origins of life on Earth and elsewhere, especially perhaps on Mars (Benner & Kim 2015). At the risk of digressing, in accordance with the preceding discussion, we believe space missions for assessing the habitability of the clouds of Venus should also attempt to quantify the abundances of bioessential elements such as molybdenum.

Thus, the above examples illustrate that the availability of the appropriate chemical elements at the desired concentrations is an inescapable constraint on the origin and sustenance of biospheres on all planets and moons, not just ocean worlds. In what follows, we will specialize to the case of subsurface oceans and examine the role of select bioessential elements. An unusual aspect of our subsequent analysis is that most of the conclusions are equally applicable to planets with only oceans on the *surface* (see Section 5.5.1). These worlds are probably more prevalent when the radius of the planet exceeds $\sim 1.5\ R_{\oplus}$ (Rogers 2015; Zeng et al. 2019), and there are theoretical grounds for presuming that many worlds within the HZs of low-mass stars consist of high water inventories (Tian & Ida 2015). Given the low densities inferred for some of the planets orbiting

the ultracool dwarf star TRAPPIST-1, these worlds are believed to possess a combination of thick atmospheres, oceans, and ice (Grimm et al. 2018).

7.6.1 Ocean worlds and their productivity

The amount of biomass within the oceans is limited by the availability of nutrients, of which some belong to the list of bioessential elements described earlier; in addition, it could be further constrained by other factors such as the abundance of electron donors. However, the question of which nutrient regulates ocean productivity is difficult to answer. Nutrient limitation is manifested through multiple avenues and across contrasting timescales. For example, nutrient availability regulates the growth rates of individual cells and imposes constraints on the maximum amount of sustainable biomass: these two factors correspond to the Blackman and Liebig limitations, respectively. Recent research has also highlighted the significance of nutrient co-limitation, wherein two or more nutrients collectively limit ocean productivity. Another key point worth noting is that different nutrients may serve as the limiting factors depending on the timescales under consideration. Hence, this led to the distinction between the proximate and the ultimate limiting nutrients, with the former and latter governing ocean productivity over short and long timescales, respectively.

The net primary productivity (NPP) on Earth is directly proportional to the availability of nutrients through the factor γ_g that quantifies the effect of nutrient availability on the maximum growth rate of marine organisms. The most common model for γ_g is the Monod equation,

$$\gamma_g = \frac{\phi_E}{\mathcal{K}_E + \phi_E}, \tag{7.57}$$

where ϕ_E denotes the concentration of the limiting element E, and $\mathcal{K}_E$ is the corresponding Monod constant (at which half-saturation occurs). The choice of this function is well justified on observational grounds (Sarmiento & Gruber 2006). An immediate consequence of (7.57) is that $\mathrm{NPP} \to 0$ when $\phi_E \to 0$ for this model. Hence, in this regime, putative biospheres would become highly oligotrophic, i.e., with very low concentrations of nutrients.

As noted earlier, answering the question of what constitutes the limiting nutrient is difficult. We are primarily interested in the ultimate limiting nutrient, since we seek to probe habitability across long (typically Gyr) timescales. Despite the inherent uncertainties, the three most prominent candidates for the limiting element on Earth are nitrogen (N), phosphorus (P), and iron (Fe). As per convention, P (in the form of phosphates) constitutes the limiting element from the standpoint of geochemists, whereas N (in the form of nitrates) is considered to be the limiter by biologists.[16] The support for N as the limiting nutrient is partly based on the facts that nitrate is depleted slightly prior to phosphate in laboratory experiments and that the addition of nitrate stimulates growth in many nutrient-limited environments, while the addition of phosphate does not produce an equivalent effect.

From the geochemical perspective, there are no biogenic sources and sinks analogous to nitrogen fixation and denitrification when it comes to phosphates. Hence, the availability of P is entirely contingent on its delivery from external sources, owing to which it is regarded as the ultimate limiting nutrient (Tyrrell 1999). On Earth, there is mounting evidence that the rise in oxygen levels and the diversification of animals may have coincided with a fundamental shift in phosphorus cycling and those of other elements during the second half of the Neoproterozoic era, sometime between 800 and 541 Ma (Reinhard, Planavsky, et al. 2017; Laakso et al. 2020). Prior to this period, the abundance of dissolved P was ostensibly much lower than today, consequently suppressing oceanic productivity (Hao et al. 2020). If the emergence of animals was indeed linked with the P cycle, the availability of nutrients (especially phosphorus) might be directly intertwined with the widespread emergence of complex multicellularity on Earth, as well as on other exoplanets.

Furthermore, the importance and versatility of phosphorus—particularly in the form of phosphates—is well established in biology, and it plays a vital role in myriad functions ranging from energy transfer to protein synthesis (Kamerlin et al. 2013). To elaborate further, (1) it forms the backbone of RNA and DNA through sugar phosphates; (2) it is an essential

16. In our subsequent discussion, "P" should be viewed as shorthand notation for phosphates and "N" for nitrates. This point is relevant because it is not actually phosphorus (or nitrogen) that is being either destroyed or created but actually phosphates (or nitrates).

component of adenosine triphosphate (ATP)—to wit, the "unit of currency" in intracellular energy transduction; and (3) it represents a major ingredient of cellular membranes (in the form of phospholipids). Overall, its importance to terrestrial life, especially inasmuch as marine ecosystems are concerned, is so profound that it has even been dubbed the "staff of life" (Karl 2000). For each of these reasons, we shall henceforth focus mostly on the sources and sinks of P and explore the ensuing implications.

A strong case could be made for iron as the limiting nutrient on modern Earth (Tagliabue et al. 2017). Iron is a central component of proteins used in respiration and photosynthesis and seems to be necessary for many enzymes involved in fixing nitrogen and using nitrates. As a consequence, lower levels of Fe can lead to reduced growth rates and uptake of other nutrients (such as nitrogen) and thereby set limits on the total productivity. However, the actual availability of Fe is contingent on its solubility in water; in turn, this depends on the amount of oxidants and reductants present in the ocean. Apart from Fe, other trace metals such as molybdenum (Mo), manganese (Mn), and cobalt (Co) may also serve as the limiting nutrients during certain epochs (Anbar 2008), but the number of empirical and theoretical studies involving these elements is limited when compared to those dealing with N, P, and Fe.

7.6.2 Sources and sinks for the limiting nutrients

Broadly speaking, nutrients can either be supplied (via sources), depleted (via sinks), or recycled (without loss or gain). The relative contribution from sources and sinks varies over time, as does the amount of nutrients that are recycled. Although these dynamical variations make it complex to analyze any of the oceanic nutrient cycles, it appears reasonable to suppose that the net inflow (sources) must balance the net outflow (sinks) eventually. In this limit, the concentrations of the nutrients would approach their steady-state values.

Hence, we will utilize the following idealized model to describe how nutrient availability evolves over time:

$$\frac{d\mathcal{C}_E}{dt} = \mathcal{S}_E - \mathcal{L}_E \mathcal{C}_E, \tag{7.58}$$

where $\mathcal{C}_E$ (in mol) is the total amount of the limiting nutrient in the ocean, $\mathcal{S}_E$ denotes the net influx (sources) of the nutrient (in mol / yr), and $\mathcal{L}_E$

represents the net outflux (sinks) of the nutrient (in yr^{-1}); as before, the subscript E refers to the element under consideration. The steady-state value of $\mathcal{C}_E$ that will be finally attained is found by equating the right-hand side to zero, which leads us to

$$\mathcal{C}_E = \frac{\mathcal{S}_E}{\mathcal{L}_E}. \tag{7.59}$$

Thus, we will delve into the major sources and sinks on Earth and determine whether they would be operational on worlds with subsurface oceans. Subsequently, we can use (7.59) to obtain the steady-state concentration. Most of our analysis will be centered on P for the reasons outlined earlier.

7.6.2.1 Phosphorus: Its sources and sinks

On Earth, the major sources of P to the oceans are riverine (continental weathering and transport by rivers) and atmospheric (deposition via aerosols and dust) in nature, although glacial weathering also plays an important role (Schlesinger & Bernhardt 2013). Neither of the first two major sources is anticipated to be operational on worlds with subsurface oceans.

An important abiotic sink of P on Earth is hydrothermal activity. Hence, this mechanism is likely to function on subsurface ocean worlds, provided that they are endowed with water-rock interactions and hydrothermal vents. Note that Enceladus and perhaps Europa fall under this category. The major hydrothermal processes that are responsible for P removal include sedimentation caused by low-temperature alteration of the oceanic crust (by reacting with the seawater) and scavenging driven by Fe oxyhydroxide particles in hydrothermal plumes. In order to calculate $\mathcal{L}_P$ for hydrothermal vents, it is instructive to consider how this estimate is carried out on Earth, then subsequently generalize this result to other worlds. Our approach closely mirrors the methodology outlined in Wheat et al. (1996).

On Earth, the majority of P depletion occurs through low-temperature ridge-flank hydrothermal interactions. The corresponding heat flow in this region is $Q_{HV} \sim 8 \times 10^{12}$ W, and the temperature of the water is raised by $\Delta T \sim 10$ K. Thus, if the entirety of Q_{HV} is used up for raising the temperature of the water flowing through the hydrothermal systems, the total circulation F_h becomes

$$F_h = \frac{Q_{HV}}{c_w \Delta T} \sim 6.3 \times 10^{15}\,\mathrm{kg\,/\,yr}, \tag{7.60}$$

where $c_w \approx 4 \times 10^3$ J kg^{-1} K^{-1} denotes the heat capacity of water at this temperature. In the idealized limit, all of the P present in the seawater will be precipitated and removed when it passes through the vent. On Earth, the seawater concentration of P, depleted through hydrothermal vent activity, is around 2 μM,[17] and multiplying this factor with F_h yields a total P removal of 1.2×10^{11} mol / yr, which differs from the more recent empirical estimate only by a factor of 2.5 (Wheat et al. 2003). Hence, we will adopt the same formalism for determining the total P removal for other worlds. Lastly, for the purpose of our order-of-magnitude calculations, we can estimate the average concentration in the ocean via $\phi_P \approx \mathcal{C}_P / M_{\rm oc}$, where $M_{\rm oc}$ is the mass of the ocean. Now, we make use of the relation $F_h \cdot \phi_P \equiv \mathcal{L}_P \cdot \mathcal{C}_P$, on account of the fact that the right-hand side and left-hand side are merely different representations of the total loss rate (in mol / yr), thereby yielding

$$\mathcal{L}_P \approx \frac{F_h}{M_{\rm oc}}. \tag{7.61}$$

Thus, with the various ingredients assembled together, we find ourselves in a position to compute $\mathcal{L}_P$ for an arbitrary subsurface ocean world.

From (7.60), we see that $F_h \propto Q_{HV}$ is a reasonable assumption provided that ΔT is the same on other ocean worlds. Since we have no way of knowing this information, we will proceed with this simplification. We also know that $M_{\rm oc} \propto R^2 \mathcal{H}$ is a fairly accurate scaling even for small worlds like Enceladus. We are now confronted with Q_{HV}, which still lacks a straightforward scaling in terms of basic physical parameters. Although this value is clearly bound to depend on the specific characteristics of the ocean world under consideration, it is ultimately derived from internal heating. For the latter, we will use the ansatz (7.5) with $\alpha = 1$ but leave Γ as a free parameter because it encapsulates the effects of other heating sources (e.g., tidal heating). Hence, we end up with $Q_{HV} \propto \Gamma M \propto \Gamma R^{3.3}$, with the last scaling following from the mass-radius relationship for icy worlds smaller than the

17. One of the common units for molarity is 1 μM and is defined as 10^{-6} moles of solute per one liter of solvent (water in our case); similarly, by definition, 1 mM $= 10^3$ μM and 1 nM $= 10^{-3}$ μM.

Earth. By substituting the above scalings into (7.61), we obtain

$$\mathcal{L}_P \sim 1.7 \times 10^{-5}\,\mathrm{yr}^{-1}\,\Gamma \left(\frac{R}{R_\oplus}\right)^{1.3} \left(\frac{\mathcal{H}}{1\,\mathrm{km}}\right)^{-1}. \tag{7.62}$$

We will now turn our attention to abiotic source(s) for P. The primary source is expected to be the weathering of the ocean floor (submarine), but one crucial difference exists. On Earth, the weathering of continents (subaerial) occurs by means of rainwater with a pH of around 5.6, whereas the equivalent submarine weathering is through seawater, whose pH is assumed to be approximately 8.0. The dissolution rate per unit area (Υ) of phosphate-producing minerals is estimated to be

$$\log \Upsilon = \log k_{H_+} - n_{H_+}\mathrm{pH}, \tag{7.63}$$

where $\log k_{H_+} \approx -4.6$ denotes the logarithmic value of the intrinsic rate constant, and $n_{H_+} \approx 0.9$ is the reaction order for combination of the minerals chlorapatite, merrillite, whitlockite, and fluorapatite (Adcock et al. 2013). Let us represent the dissolution rates on Earth and subsurface worlds by Υ_E and Υ_{SO} respectively, and introduce the variables $\Delta = \mathrm{pH}_{SO} - \mathrm{pH}_E$ (where $\mathrm{pH}_E = 5.6$) and $\delta = \Upsilon_{SO}/\Upsilon_E$. Using these definitions and simplifying (7.63), we end up with

$$\log \delta = -0.9\Delta, \tag{7.64}$$

and choosing $\Delta \approx 2.4$ on the basis of the above considerations leads us to $\delta \approx 7 \times 10^{-3}$. The dissolved pre-anthropogenic P input from continental weathering is a challenging task for geochemists, and we adopt the value $\sim 1.3 \times 10^{10}$ mol / yr for the Earth, which is consistent with the upper bound delineated in Paytan and McLaughlin (2007). We will also suppose that the area of weathered regions is linearly proportional to the total area of the subsurface world. Using the above data, the net delivery of P to the ocean becomes

$$\mathcal{S}_P \sim 1.3 \times 10^{8}\,\mathrm{mol\,/\,yr} \left(\frac{\delta}{0.01}\right) \left(\frac{R}{R_\oplus}\right)^2, \tag{7.65}$$

in which we have normalized δ by its characteristic value estimated above; setting $\delta = 1$ will yield the estimate for Earth specified in the previous paragraph. In actuality, (7.65) will be raised by a factor of order unity

since continental weathering spans about 30 percent of the Earth's surface, whereas submarine weathering in ocean worlds would transpire over the entire area due to the complete absence of continental masses. Moreover, in deriving (7.65), we have implicitly presumed that the mean abundance of phosphates within rocks from the oceanic crust is comparable to those from the continental crust. Inasmuch as the former is, on average, much lower than the latter (Maruyama et al. 2013), we caution that (7.65) serves as an upper bound.

We are now in a position to use (7.62) and (7.65) for determining $\mathcal{C}_P$ by means of (7.59), which yields

$$\mathcal{C}_P \sim 7.6 \times 10^{12}\,\mathrm{mol} \left(\frac{1}{\Gamma}\right)\left(\frac{\delta}{0.01}\right)\left(\frac{R}{R_\oplus}\right)^{0.7}\left(\frac{\mathcal{H}}{1\,\mathrm{km}}\right). \tag{7.66}$$

As a consistency check, choosing $\delta = 1 = \Gamma$ and $\mathcal{H} \approx 3.7$ km for the Earth leads us to $\mathcal{C}_P \sim 2.8 \times 10^{15}$ mol, which is almost identical to the empirically estimated value of 3.2×10^{15} mol in Earth's oceans currently (Benitez-Nelson 2000). Although $\mathcal{C}_P$ has been calculated, we are more interested in the average concentration $\phi_P \approx \mathcal{C}_P / M_{\mathrm{oc}}$ introduced earlier. Hence, using this definition in conjunction with (7.66) leads us to

$$\phi_P \sim 20\,\mathrm{nM} \left(\frac{1}{\Gamma}\right)\left(\frac{\delta}{0.01}\right)\left(\frac{R}{R_\oplus}\right)^{-1.3}. \tag{7.67}$$

A couple of interesting features concerning this formula are the absence of any dependence on the average ocean depth and the monotonic decrease with respect to the radius. Another point to note is that $\Gamma \gg 1$ is possible, which would lead to lower values of ϕ_P. At first glimpse, ϕ_P appears to be around two orders of magnitude lower than the typical P concentrations observed in Earth's oceans, but it must be recognized that it exhibits a very strong (i.e., exponential) dependence on the pH through δ. To illustrate this point further, let us evaluate ϕ_P for Europa and Enceladus. To err on the side of caution, we will use $\Gamma = 1$ here, although it is truly a few times higher for Europa and $\mathcal{O}(10)$ for Enceladus. Hence, the concentrations we obtain below are best viewed as upper bounds.

When it comes to Europa, we pointed out in Section 7.5 that oxidants formed on the surface via radiolysis are presumably delivered to the

subsurface ocean, where they react with sulfides and consequently give rise to a highly acidic ocean with a pH of 2.6. In this case, we obtain $\delta \approx 5 \times 10^2$, thereby leading us to $\phi_P \sim 6.1$ mM. Hence, at least when viewed solely from the standpoint of P availability, Europa does not seem to pose any issues for habitability. However, there are other detrimental effects arising from ocean acidification to this degree, which shall not be addressed here. It is worth noting, however, that recent models are compatible with higher pH values of $\lesssim 6$ that could reduce ϕ_P by orders of magnitude but may concomitantly pose fewer challenges to aquatic lifeforms.

The situation is rendered very different when we consider Enceladus. As per certain theoretical models, the ocean is believed to comprise a Na-Cl-CO_3 solution with a typical pH of ~ 11 or 12 (Glein et al. 2015). The alkanine (high pH) nature of the ocean was argued to stem from the serpentinization of chondritic rock. If we choose a pH of 11.5, we find that $\delta \approx 4.9 \times 10^{-6}$, and using (7.67) leads us to $\phi_P \sim 0.65$ nM. The concentrations of total dissolved phosphorus (TDP), conventionally defined as the sum of dissolved inorganic phosphorus (SRP) and dissolved organic phosphorus (DOP), in many oligotrophic biospheres (e.g., the North Pacific Gyre) are typically ~ 10 to 100 nM. However, there are some environments on Earth where oligotrophes survive even at sub-nM concentrations of dissolved P (Karl & Björkman 2015). Hence, if Earth's extremophiles do constitute genuine extraterrestrial analogs, there are tenable grounds for contending that life-as-we-know-it may exist on Enceladus even at the very low concentrations predicted by our model.

On the other hand, if we adopt $\Gamma \sim 30$ for Enceladus, which preserves consistency with the latest models (Choblet, Tobie, Sotin, Kalousová & Grasset 2017), we end up with $\phi_P \sim 0.02$ nM. In case this steady-state concentration is accurate, basing our analysis on the oligotrophs documented on Earth, the chances of Enceladus supporting a thriving global biosphere are rather slim prima facie. However, this conclusion must be viewed with due skepticism because current theoretical modeling based on the composition of Enceladus's plumes suggests that the pH is ~ 9 or 10 (Glein et al. 2018; Glein & Waite 2020). Thus, by selecting a pH of 9.5, we arrive at $\phi_P \sim 41.3$ nM after employing (7.67) and setting $\Gamma = 1$; on the other hand, choosing $\Gamma \sim 30$ yields $\phi_P \sim 1.38$ nM. As the revised values are nearly two orders of magnitude higher than our prior estimates, the prospects for an aquatic biosphere within Enceladus are considerably

enhanced, although the putative organisms are still likely to be oligotrophic in nature.

A case could be made for testing the validity of our hypothesis and the steady-state concentration by sampling the Enceladus plume or ocean. While H_2 and complex organic molecules above 200 atomic mass units, including those with potentially nitrogen-bearing species, have been detected in the Enceladus plume by the *Cassini* spacecraft, there is no evidence of phosphorus thus far. Recently, the ROSINA (Rosetta Orbiter Spectrometer for Ion and Neutral Analysis) mass spectrometer was able to detect phosphorus (presumably in the form of PH_3) in the coma of a comet (67P / Churyumov-Gerasimenko) for the first time. Even though this discovery does open up the possibility of carrying out similar observations of the Enceladus plume, a major difference is that the predicted abundance of PH_3 relative to water was quite high ($\sim 10^{-3}$) in the coma. On the basis of our predictions that have yielded a much lower steady-state concentration of P in the Enceladus ocean, the mass spectrometer aboard a future mission will require a high degree of sensitivity to account for this scenario. In contrast, if in situ analysis of the ocean can be carried out, even sub-nM concentrations are readily measurable by systems with miniaturized spectrophotometers, i.e., instruments that analyze molecules on the basis of the amount of light that is reflected or transmitted.

We may also ask the question of whether there is a source analogous to glacial weathering on subsurface ocean worlds. The complicated interplay of resurfacing processes, especially if the world is geologically active, can certainly result in some fraction of the ice being melted into the ocean and supplying nutrients. The difficulty with this approach is that we have no way of knowing the abundance of soluble P in the ice layer. As it would be depleted over time by deposition into the ocean, it will need to be replenished by external sources such as meteorites and comets, and the flux of impactors remains unknown for extrasolar systems. Lastly, if the ice layer is in equilibrium, an equal amount of water (along with the dissolved P) must be refrozen in the same period, thereby removing P. In summary, it seems plausible that this mechanism would be neither a net source nor a sink. Thus, we will not seek to quantify this source in this chapter.[18]

---

18. A related discussion of this issue is presented in Lingam and Loeb (2019g).

The last sink that we consider for P is an important one: the burial of organic sediments. Estimating the magnitude of this biotic sink is difficult because it depends on biological factors and the burial rate is subject to major spatiotemporal variability and dependent on a number of environmental factors. Of equal importance is the datum that there exist multiple channels by which P is lost through this sink. An overview of this complex sink can be found in Benitez-Nelson (2000). However, there is a crude, but fairly effective, method for calculating the amount of P depleted through this sink. Let us begin by considering the process of sedimentation. In the idealized limit, the sedimentation process is well approximated by Stokes' Law. In this regime, the settling velocity U of the particle is given by

$$U = \frac{g\Delta\rho\mathcal{D}^2}{18\mu_f}, \tag{7.68}$$

where $\Delta\rho$ is the density difference between the particle and the fluid, $\mathcal{D}$ is the diameter of the particle, and μ_f is the viscosity of the fluid. Naturally, we cannot estimate $\Delta\rho$ and especially $\mathcal{D}$ since these depend on organismal properties, and we therefore assume their characteristic values are similar to those on Earth. Before rushing ahead, it should be realized that an important criterion was implicitly held to be valid: the (downward) settling velocity exceeds the (upward) vertical velocity w in the deep ocean. In quantitative terms, this relation is expressible as follows:

$$\left(\frac{\Delta\rho}{200\,\mathrm{kg/m^3}}\right)\left(\frac{\mathcal{D}}{10^{-5}\,\mathrm{m}}\right)^2\left(\frac{R}{R_\oplus}\right)^{1.3} > 0.01\left(\frac{w}{10^{-7}\,\mathrm{m/s}}\right) \tag{7.69}$$

The left-hand side and right-hand side have been normalized on the basis of the characteristic values for phytoplankton and the average deep ocean vertical velocity on Earth, respectively. Hence, from (7.69), it is conceivable that sedimentation of organic material could occur only when $R/R_\oplus \gtrsim 0.03$. In this case, adopting the above normalizations for $\Delta\rho$ and $\mathcal{D}$, we find $U \sim 1.2 \times 10^{-5}\,(R/R_\oplus)^{1.3}$, where the scaling follows from $g \sim GM/R^2 \propto R^{1.3}$.

To illustrate how our heuristic method for computing the magnitude of this sink works, let us consider the Earth first. We have $\mathcal{H} \approx 3.7$ km and $U \sim 10^{-5}$ m / s. Thus, the rate at which the particle will be transported to the bottom, assuming a constant velocity, is $\tau_s^{-1} \approx U/\mathcal{H}$. The variable τ_s

could be envisioned as the sedimentation time, and using the above data leads us to $\tau_s \approx 9.8$ yr for the Earth. Now, we will assume that the total depletion rate of P (in mol / yr) through this sink is the product of the sedimentation rate and the amount of P (in mol) that is lost. In using the word *lost,* we mean that the organic P in the ocean is not perfectly recycled—that is, there exists a fraction ε that will be buried through our sink mechanism. The parameter ε is highly variable, both across space and time, but most studies indicate that $\varepsilon \ll 1$ for present-day Earth (Sigman & Hain 2012). Thus, let us work with the fiducial value of $\varepsilon \sim 3 \times 10^{-4}$ to match the depletion rate of P via this sink on Earth (as shown below). The burial efficiency documented in productive regions of the ocean (e.g., the area encompassing Antarctic Circumpolar Current in the Southern Ocean) may lie between 2×10^{-4} and 8×10^{-4}.

We are now in a position to compute the depletion rate of P due to burial of organic matter from $\varepsilon \cdot \mathcal{C}_P \cdot \tau_s^{-1}$. Given that the total P inventory in Earth's oceans is $\mathcal{C}_P \sim 3.2 \times 10^{15}$ mol, we end up with the P loss rate of 9.8×10^{10} mol / yr that is consistent with the empirically determined lower bound of 9.3×10^{10} mol / yr for our planet (Paytan & McLaughlin 2007). However, a hidden assumption permitted us to obtain this result: we have supposed that all of the dissolved P in the ocean was incorporated into organisms—namely, dissolved P is solely prevalent in organic form. Despite this supposition being patently invalid on Earth, it may be relatively accurate for worlds with oligotrophic biospheres where organisms maximize their consumption of the available nutrients. In view of this caveat, it seems more appropriate to reinterpret the fudge factor ε as the effective fraction of the *total* dissolved P that is subjected to sedimentation.

Using the fact that the net output of P, by definition, is $\mathcal{L}_P \cdot \mathcal{C}_P$ and equating this with $\varepsilon \cdot \mathcal{C}_P \cdot \tau_s^{-1}$ leads us to the loss rate

$$\mathcal{L}_P \sim \frac{\varepsilon U}{\mathcal{H}}. \tag{7.70}$$

Combining (7.70) with our prior scaling for U yields the final result,

$$\mathcal{L}_P \sim 3.8 \times 10^{-5}\,\mathrm{yr}^{-1} \left(\frac{\varepsilon}{10^{-4}}\right) \left(\frac{R}{R_\oplus}\right)^{1.3} \left(\frac{\mathcal{H}}{1\,\mathrm{km}}\right)^{-1}, \tag{7.71}$$

and choosing $\mathcal{H} \approx 3.7$ km and $\varepsilon \sim 3 \times 10^{-4}$ will yield the appropriate value of $\mathcal{L}_P$ for modern Earth. Upon comparing (7.62) and (7.71), two points

stand out. First, both these equations display the same dependencies on R and $\mathcal{H}$ and are comparable in magnitude for Earthlike worlds. Second, they both have free parameters, Γ and $\varepsilon/\left(10^{-4}\right)$, that can become much greater than unity. Recent estimates indicate that during the Precambrian era, the amount of P recycled was lower due to a lack of oxidants (Kipp & Stüeken 2017). If this hypothesis is correct, the value of ε would have been higher, owing to which the steady-state concentration ϕ_P also undergoes a reduction. Hence, this qualitatively explains why the primary productivity of the oceans was much lower during this epoch.

7.6.2.2 Sources and sinks for other nutrients

Let us briefly consider the sources and sinks for N. We have seen that hydrothermal vents serve as sinks for P, and the same is true for several other trace metals and rare earth elements. However, hydrothermal systems are ostensibly not an abiotic sink for N. As a result, two phenomena are left: submarine weathering and burial of organic material, which function as a source and a sink for N, respectively.

Studies that estimate the dissolution rates for nitrate-producing minerals for different pH values are rare. Hence, constructing the analog of (7.63) is not easy, and the issues with estimating the organic burial rate were delineated earlier. However, a potential way for calculating the latter is to assume that the sedimentation rate $\mathcal{L}_N$ will be the same as (7.71). This is reasonable because all particles of a given size and composition would undergo sedimentation at a constant rate. On the other hand, note that $\mathcal{L}_P \mathcal{C}_P \neq \mathcal{L}_N \mathcal{C}_N$ even if $\mathcal{L}_P \approx \mathcal{L}_N$ since the concentrations of P and N differ in organic matter. On Earth, most organisms in the ocean are characterized by the famous Redfield (stoichiometric) ratio of N:P = 16:1. Owing to the pH dependence that we have not quantified, the production rate of N due to submarine weathering is unknown; however, there is a strong possibility that it will be lower than the corresponding value for continental weathering by rainwater; if we consider the latter as approximated by the riverine input, it equals $\sim 10^{12}$ mol / yr for an Earth-sized world.

Next, let us turn our attention to Fe. Although the major sources of dissolved Fe in the oceans are mineral dust and sediments from subaerial (aboveground) continental weathering, neither are directly present on subsurface worlds. Instead, the submarine weathering of Fe will need to be considered, which also depends on the pH and will not be considered further herein. However, an interesting aspect of the iron cycle is that

hydrothermal vents actually serve as a *source* of dissolved Fe (Tagliabue et al. 2017). The total amount of Fe produced per unit time ($\mathcal{S}_{\rm Fe}$) depends on the hydrothermal circulation F_h that was argued to be proportional to ΓM. On the basis of this ansatz, we obtain

$$\mathcal{S}_{\rm Fe} \sim 9 \times 10^8 \,{\rm mol/yr}\, \Gamma \left(\frac{R}{R_\oplus}\right)^{3.3}, \tag{7.72}$$

where the normalization has been chosen to preserve consistency with the Earth, although recent studies suggest that this value may be lower by about 20 percent. Turning our attention to the sinks of Fe, the burial of organic sediments plays a major role. It goes without saying that estimating the magnitude of this sink is not an easy task because of its biotic nature. However, if the loss rate can be approximated by our model in Section 7.6.2.1, then $\mathcal{L}_{\rm Fe} \approx \mathcal{L}_P$ is potentially valid for the reasons outlined earlier when discussing the N sinks.

A few general remarks are in order at this stage. The solubility in water undoubtedly plays a major role in determining the concentrations of dissolved nutrients. In many instances, phosphates are relatively insoluble, whereas nitrates are soluble (Schlesinger & Bernhardt 2013). Iron is of more interest in this regard, since it happens to be insoluble when oxidized but is highly soluble in reduced form; this property is also evinced by manganese. In other words, on subsurface ocean worlds like Enceladus, where H_2 production has been documented, there is a high likelihood that Fe would be soluble. This leads us to conclude that iron has a low probability of serving as the limiter of ocean productivity on such worlds. Future analyses of the limiting nutrient on subsurface ocean worlds should therefore take the solubility in water into account.

7.6.3 What can stellar spectroscopy reveal about bioessential elements?

One of the difficulties with studying the habitability of extrasolar systems, particularly those with subsurface oceans, is that very few direct measurements are possible. Hence, even deducing limited information can assist in selecting targets for more detailed observations by state-of-the-art telescopes. One method that we shall describe below is to use stellar spectroscopy and constrain the *stellar* abundances of the bioessential elements.

Some observational studies have already been devoted to finding the P abundances for stars with varying values of [Fe / H], either via the near-ultraviolet P I doublet at 2135 / 2136 Å or through the weak P I lines in the infrared at 10500–10820 Å. Most of these analyses obtained an average value of [P / Fe] of ~ 0.1 for stars whose metallicity was in the range $-1.0 < [\text{Fe} / \text{H}] < 0.2$. An interesting point worth noting here is that [P / H] appears to be roughly proportional to the metallicity [Fe / H] across several orders of magnitude (Jacobson et al. 2014).

When it comes to heavier elements in biology, such as molybdenum, they are synthesized through a variety of mechanisms including the slow and rapid neutron-capture processes (the s- and r-process, respectively) and the proton-capture process (p-process). The Laser Interferometer Gravitational-Wave Observatory (LIGO) detection of gravitational wave GW170817 and follow-up electromagnetic observations have revealed that neutron star mergers may play a dominant role in the production of r-processed elements. Since these events are spatially very localized, the stellar abundances of elements produced through the r-process should vary to a considerable degree, and this hypothesis is borne out by observations (Delgado Mena et al. 2017). For example, observations of metal-poor stars using the Mo I 3864 and 5506 Å spectral lines have illustrated the datum that the Mo / Fe ratio varies by more than two orders of magnitude (Hansen et al. 2014). Putting these facts together, it becomes quite evident that habitable exoplanets with bioessential element abundances very different from terrestrial values are bound to exist.

Although we have argued in favor of stellar spectroscopy as an important diagnostic tool, this method does have its share of limitations insofar as the habitability of planets and moons is concerned. To begin with, the stellar and planetary abundances (of bioessential elements) need not be similar since terrestrial planets are expected to vary widely in terms of their composition. Next, it does not automatically follow that the concentrations of biogenic elements in the oceans will be proportional to their crustal abundances. As we have seen, this is because the concentrations in the oceans depend on a number of biotic and abiotic sources and sinks that are evolving with time. Despite these caveats, we believe that measuring the stellar abundances of P and other bioessential elements is a path worth pursuing, especially in light of our current observational limitations and modest theoretical understanding of habitability (Hinkel et al. 2020).

7.7 EVOLUTIONARY TRAJECTORIES ON SUBSURFACE OCEAN WORLDS

An admittedly speculative, but fundamental, question immediately presents itself when we contemplate ocean worlds, whether on the surface or beneath it. We have become so attuned to the complexity of life on land that we lose sight of the richness of life in Earth's oceans. Nevertheless, the question remains: Is it possible for intelligent life to exist in ocean worlds? To answer this question, we must recall from Section 3.7 that we have deployed the word *intelligence* to broadly encompass four traits that have, in the historical past, been identified as distinctively human: (1) self-awareness, (2) tool construction and usage, (3) culture, and (4) language. We have chosen these four traits since they are relatively easy to define and analyze. In contrast, questions such as "Are animals conscious?" are tremendously important yet also very difficult to approach because of the inherent ambiguities in defining and testing concepts like consciousness (Griffin 2001; de Waal 2016). It goes without saying that these four traits are not mutually exclusive, and there are plenty of other characteristics that have been put forward as being unique to humans (e.g., foresight), but an in-depth treatment of this vast topic falls outside the scope of this chapter.

7.7.1 Evolutionary innovations on the land and in the sea

Before embarking on an in-depth scrutiny of (1)–(4), it is worth examining a related question: Are there any major evolutionary innovations that have a higher likelihood of transpiring on land as opposed to in the seas? In order to answer this question, we must begin by contemplating the evolutionary history of the Earth, our only life-bearing sample. Upon doing so, we find that many of the earlier fundamental evolutionary breakthroughs possibly occurred in the sea. One such notable development was the onset of predation in the Precambrian, a breakthrough that forced prey to adapt accordingly and vice versa, thereby sparking off an evolutionary arms race. Some of the other major innovations that arose in the sea during the Cambrian epoch (or earlier) include biomineralization, mixing of sediments by animals (bioturbation), development of different modes of locomotion like undulation and jet propulsion, and internal fertilization.

However, once life had established itself on the land after the end of the Ordovician period at ~ 440 Ma, most of the subsequent major evolutionary

breakthroughs seem to have occurred first on land: to be more precise, at least eleven out of thirteen appear to fall under this category (Vermeij 2017). Let us denote the difference in the time periods when a particular evolutionary innovation appeared on land and the sea by Δt. Some of the major breakthroughs on land, along with their corresponding values of Δt, are adumbrated below:

- Vascular structure in plants ($\Delta t \approx 345$ Ma): The chief advantages of vascular tissue in plants (although complex) include enhanced transport of water and nutrients and mechanical support. The development of vascular tissue enabled plants to get bigger, increase their photosynthetic yield, and encompass a wider range of environmental habitats.
- Nutrient mining in plants ($\Delta t \approx 117$ Ma): The development of roots enabled plants to gain greater access to nutrients; the correlation between the depth (or breadth) of root systems and the sizes of plants has been documented.
- Aerial locomotion ($\Delta t \approx 94$ Ma): Several advantages are associated with aerial locomotion, such as achieving faster velocities, escaping from predators, and reaching new habitats. We compare the speeds of flying and swimming a few paragraphs below.
- Animal endothermy ($\Delta t \approx 222$ Ma): The many benefits of endothermy include the ability to function in niches inaccessible to ectotherms, mitigated vulnerability to environmental fluctuations, increased muscle power, and higher levels of activity.
- Eusociality ($\Delta t \approx 116$ Ma): Eusocial animals possess several advantages, including access to greater resources (e.g., food and territory) due to living in large groups and better protection (owing to cooperative care). Eusociality is predominantly terrestrial, and it may be possible that this has something to do with the fact that the construction, protection, and stability of nests is higher on land with respect to the sea.
- Animal-mediated dispersal of seeds and spores: The importance of land animals as pollinators has been established beyond doubt, given that nearly 90 percent of flowering plants in certain latitudes rely on them. Despite the centrality of this phenomenon on land and its continual existence since at least 425 Ma, it has never been documented in the sea.

- Communal construction: A wide range of animals are known to build nests, burrows, and other shelters. The most striking example in this category is an animal that evolved quite recently and goes by the unassuming scientific name of *Homo sapiens,* which has successfully built the likes of the Great Pyramid of Giza, the Eiffel Tower, and the Great Wall of China. Communal nests increase the odds of offspring survival due to improved defenses against predators, decreased parental costs, and perhaps polyandry. No marine animals appear to have evolved communal construction, although their counterparts on land evolved this feature around 180 Ma.

Collectively, this points to an irreversible and genuinely significant shift in the evolutionary trajectory of life on our planet. Due to this, the colonization of land was identified by Knoll and Bambach (2000) as one of the six major evolutionary steps on Earth, and possibly on other worlds as well. From the above evidence, we also have a partial answer to our question: yes, many innovations seem to originate with greater ease on land. For some of the evolutionary breakthroughs, the reason behind this discrepancy between land and sea stems at least partly from physical considerations. Most high-performance evolutionary innovations entail a high degree of activity, and it may be easier to achieve them in the less viscous and dense medium of air than in water. We will consider the example of aquatic locomotion (swimming) and demonstrate how it engenders lower speeds in comparison to aerial locomotion (flying).

In carrying out this analysis, which essentially mirrors the approach laid out in Bejan and Marden (2006), we will get rid of all constant numerical factors since we are only interested in the algebraic expressions—to wit, the scaling relationships. We commence our study by considering the case of swimming animals. If a generic aquatic animal wishes to stay above the water, it must expend energy, and the required amount can be calculated by estimating the work done against gravity. This leads us to $E_p \sim M_b g L$, where $M_b \sim \rho_b L^3$ denotes the mass of the animal, with ρ_b and L representing its density and scale length, respectively. At the same time, the animal must spend energy to combat the effect of drag.

The exerted force can be derived from the drag equation, which is approximately $F_d \sim \rho_w L^2 V_s^2$, where $\rho_w \sim \rho_b$ is the density of

water,[19] V_s is the swimming velocity, and we have neglected a constant that is typically of order unity. In order to determine the energy expended, we are free to employ the relation $E_d \sim F_d L'$, where $L' \sim V_s \left(L/g\right)^{1/2}$. The factor $\left(L/g\right)^{1/2}$ follows from the postulate that the time it takes to lift the animal above water is roughly the same as the time that it would take to descend due to gravity, where the latter is encapsulated by the time of descent.[20] We are interested in calculating the speed at which the energy required for swimming a unit distance will be duly minimized (Alexander 1996). In mathematical terms, this statement is equivalent to calculating the extrema of the function $E' = (E_p + E_d)/L'$. By setting $dE'/dV_s = 0$, we end up with

$$V_s \sim g^{1/2} \rho_b^{-1/6} M_b^{1/6}, \tag{7.73}$$

where we have dropped a constant factor of order unity in presenting the final result. A similar analysis can be carried out for locomotion in air, and the procedure is virtually identical, except for the fact that ρ_w must be replaced by ρ_a, the density of air, in the expression for the drag force. By carrying out the same optimization procedure, we find

$$V_f \sim \left(\frac{\rho_b}{\rho_a}\right)^{1/3} g^{1/2} \rho_b^{-1/6} M_b^{1/6}, \tag{7.74}$$

where V_f denotes the flying velocity. Upon comparing the optimal velocities for flying and swimming, it can be seen that $V_f \sim 10 V_s$ because $\rho_b \sim 10^3 \rho_a$. This serves to illustrate that the flying speeds are about an order of magnitude higher than the swimming speeds. We have not worked out the optimal running speed here, but it turns out to be identical to (7.73), except for an additional factor χ^{-1} on the right-hand side, where $\chi \in (0, 1)$ denotes the coefficient of friction. Another noteworthy point here is that the speeds are monotonically increasing functions of g (albeit somewhat weakly), indicating that smaller worlds could possess organisms with lower optimal speeds.

---

19. In this treatment, the effects of buoyancy are not explicitly included.

20. The situation is analogous to throwing a ball vertically in the air, after which it falls back to the original position. The times taken for ascent and descent are approximately equal.

7.7.2 The likelihood of intelligence on ocean worlds

> The more I consider this mighty tail, the more do I deplore my inability to express it. At times there are gestures in it, which, though they would well grace the hand of man, remain wholly inexplicable. In an extensive herd, so remarkable, occasionally, are these mystic gestures, that I have heard hunters who have declared them akin to Free-Mason signs and symbols; that the whale, indeed, by these methods intelligently conversed with the world.
>
> —Herman Melville, *Moby-Dick; or, The Whale*

We will outline the evidence for intelligence in marine animals on Earth and draw on these results to discuss the prospects for intelligent life in extraterrestrial oceans; as always, such extrapolations must be viewed with a healthy dose of skepticism due to the plethora of unknowns involved.

7.7.2.1 Self-awareness and theory of mind

To put it simply, self-awareness constitutes the propensity of individuals to possess an *autobiographical sense*—that is, the sense of "I." One of the most widely used experiments in this context is mirror self-recognition (MSR). Species that are capable of MSR can recognize themselves in the mirror. The difficulty with most tests of animal cognition, however, is that they must simultaneously account for the materialization of false positives (i.e., unconsciously influencing the animals to achieve the desired outcome) and false negatives (which arise when the tests are not properly oriented toward the animals being studied).

A well-known study by Reiss and Marino (2001) presented convincing evidence that dolphins appear to successfully pass the MSR test. Markings were placed on a couple of dolphins, after which they were situated in front of a mirror. It was found that the dolphins spent much more time in front of the mirror when the markings were present and that they responded to the type of markings. Similarly, when the mirror was removed (or covered up), it was found that the dolphins spent fewer time at the same location. Subsequent experiments also provided further evidence in favor of dolphins using mirrors to investigate their own bodies, thereby passing the MSR test on the surface.

Another important trait that reflects on self-awareness is metacognition. The most common definition is "thinking about thinking," but it is

more instructive to envision metacognition as the ability to monitor, assess, and adaptively control one's learning and behavior. We refer the reader to Hampton (2009) for a detailed overview of metacognition in nonhuman animals. Dolphins have been shown to display evidence of metacognition: when subjected to multiple tests, a dolphin adapted to undertake easier tests to maximize its rate of reward and declined to take on difficult tests. This behavior is akin to that of humans when it comes to exams, where they may choose to spend more time on certain portions of the syllabus in order to maximize their chances of getting a high grade.

Despite these promising results, we should bear in mind that assessing the full extent of self-awareness in animals is very difficult, and tests like the MSR shed light only on certain narrow aspects. In some quarters, there has been a tendency to loosely conflate self-awareness and the theory of mind (ToM), despite the two being quite distinct. The latter is defined as the ability to recognize and comprehend the mental states of another individual (or oneself) and to act in accordance with this knowledge. There exists some evidence, albeit controversial, in favor of the hypothesis that the great apes possess a theory of mind (Krupenye et al. 2016). However, there appears to be an absence of similar evidence, which is not necessarily tantamount to evidence of absence, for dolphins and whales.

7.7.2.2 Tool making and use

The dexterity of human hands (especially the opposable thumbs) when it comes to undertaking complex tasks is a fact that we often take for granted. If one scrutinizes images of the *cortical homunculus*—a heuristic neurological map (of sorts) that illustrates the extent of the brain dedicated to operating the different parts of the human body—the disproportionately high weight attached to our hands becomes patently obvious (F. R. Wilson 1998). Clearly, on our planet no other species, be it aquatic or terrestrial, has become as proficient at manufacturing advanced tools as humans have. Thus, if we were to restrict ourselves to tool construction comparable to the level of humans, this would superficially bring us to the end of the discussion, insofar as the Earth is concerned.

Although the usage of tools by land animals has received the most attention, there are many aquatic animals that fall under this category (Mann & Patterson 2013). For instance, octopodes have been found to carry coconut shells with them (despite being an encumbrance), and assemble

them in the future to construct shelters from predators. This has been argued to be an example of not only tool usage but also goal-directed behavior. In addition, octopodes modify their dens extensively to accord better protection, and they use water jets as a means of burrowing deeper and camouflaging themselves from predators; neither of these activities may fall strictly under the domain of tool use, but they reveal the capacity for sophisticated object manipulation. Here, we have selectively emphasized octopodes to bolster the conjecture that adept tool users with limbs can evolve and thrive in the oceans. Moreover, octopodes as well as other cephalopods (i.e., members of the class *Cephalopoda*) evince other remarkable traits—such as recognizing and remembering other individuals of the same species—and manifest multiple signatures of social complexity, high general intelligence, and consciousness (Godfrey-Smith 2016; Hanlon & Messenger 2018; Amodio et al. 2019).

Cetaceans (*Cetacea*), comprising whales, dolphins, and porpoises, are classic examples of tool-using aquatic animals sans arms. Bottlenose dolphins (*Tursiops* sp.) have been shown to break sponges (possibly even conch shells) and wear them over their rostrums (beaks) for protection when foraging for bottom-dwelling fish. Humpback whales (*Megaptera novaeangliae*) have been shown to utilize a complex method called bubble-netting, expelling bubbles to carefully construct a cylindrical "net" that entraps their prey; bottleneck dolphins utilize a similar procedure called mud ring feeding. An unusual feature exhibited by these cetaceans is that certain tool-use methods are culturally transmitted—that is, these practices are vertically transmitted from mother to female offspring, and that too only within a fraction of the total population.

Although it appears that some degree of tool usage could evolve on ocean worlds, it might seem unlikely that sophisticated technology comparable to the level of humans is attainable, especially close to (or at) the surface. To an extent, this notion is attributable to the inherent difficulty of extracting raw materials from the ocean seafloor and transporting them to higher altitudes, unless the putative macrofauna live permanently on the ocean floor. It should also be said that the likelihood of evolving dexterous appendages for sophisticated tool making and usage in ocean worlds remains unclear. At least insofar as Earth is concerned, instances of land animals using tools are more numerous than their aquatic counterparts, and land dwellers outwardly display a greater degree of sophistication and range when it comes to deploying tools.

7.7.2.3 Cultural transmission

Culture refers to the vertical transmission of behavioral traits that originated in local populations across multiple generations by virtue of social learning. In other words, it facilitates the flow of information in ways that are akin to genetic information, thereby driving innovation and evolution. As pointed out in Section 3.8.3, rapid advances in our understanding of cultural transmission have revealed the fundamental role of culture in shaping and accelerating human evolution. The reader is directed to *Not by Genes Alone: How Culture Transformed Human Evolution* by Richerson and Boyd (2005) for a thoughtful, if somewhat dated, account of this subject; additional references and descriptions are expounded in Section 3.8.3.

Bearing the above discussion in mind, we ought to inquire whether cultural transmission is manifested in aquatic animals, especially cetaceans. Owing to many factors—the complexity of the question, the semantic difficulties in defining *culture,* the possibility of anthropomorphic bias, and checkered history until the dawn of the twenty-first century, to name a few—this question is not straightforward to answer. However, in view of the cumulative advances made over the past couple of decades, there are cogent grounds for contending that whales and dolphins are characterized by intricate cultures and exhibit compelling signatures of gene-culture coevolution (Foote et al. 2016; Whitehead 2017). The reader may consult *The Cultural Lives of Whales and Dolphins* by Whitehead and Rendell (2015) for a meticulous overview of this subject. This comprehensive book furnishes a number of examples, of which we shall point out merely a few cases.

One example, which we have already encountered, is that both whales and dolphins take part in the vertical social transmission of information about tool usage. The bottlenose dolphins in Shark Bay that use sponges in foraging have been shown to descend from a single matriline—that is, they descended from one dolphin named Sponging Eve. The feeding behavior of killer whales (*Orcinus orca*) also displays very specialized patterns that appear to stem from social learning. Another major line of evidence comes from the analysis of humpback whale songs. Noad et al. (2000) demonstrated that the particular song used by males along the east coast of Australia underwent a radical change after the apparent introduction of a few foreigners from the west coast. Both the rapidity and extent of this change in vocalization were highly unusual and were interpreted as strong evidence in favor of social learning and whale culture. On a more speculative note, it has

even been suggested that cetaceans may possess a sense of morality (a word indubitably loaded with both meaning and ambiguity).[21] While most of the behavioral traits outlined above rest on a fairly sound empirical footing, they do lend themselves to alternative interpretations. Nonetheless, if we adopt the definition of *culture* introduced previously, it seems plausible that whales and dolphins do indeed lead cultural lives and are consequently shaped by gene-culture coevolution. Hence, one may be predisposed to believe that extraterrestial life in ocean worlds could evolve culture, although this putative culture will probably materialize in forms very different from those documented in humans.

7.7.2.4 Language

Although language is an everyday part of human life, its origin and evolution remain shrouded in mystery. We will not delve into this topic in detail, as we have already elaborated on it in Section 3.8.2.

There are two general statements worth recalling at this stage. If we restrict ourselves to communication, many species on Earth have the ability to transmit information to varying degrees. The second point that might be made with a reasonable degree of confidence is that only humans possess the capability for recursive strategies (cf. Abe & Watanabe 2011; Ferrigno et al. 2020), which allows them to convey abstract concepts via symbolic communication. This forms the heart of the famous hypothesis outlined by Hauser et al. (2002), where they distinguish between FLB and FLN, the faculties of language in the broad and narrow senses, respectively. While the roots of FLB may exist in other nonhuman animals, the authors posit that FLN (recursion) is a recent evolutionary innovation that is unique to humans.

Another point that deserves to be highlighted is that cetaceans reportedly evince many of the basic skill sets associated with acquiring languages in humans (Janik 2014). A couple of examples serve to illustrate the advanced vocal learning abilities of cetaceans. Dolphins have been shown to apparently understand the intent behind the pointing gestures used by

---

21. Of course, this perspective is hardly novel, as illustrated by the following perceptive statement from *The Descent of Man* by Charles Darwin (1871, p. 75): "Besides love and sympathy, animals exhibit other qualities connected with the social instincts, which in us would be called moral. . . ."

humans, and they also have the capacity to utilize such gestures themselves; this is presumed to be of high importance given the close links that are believed to exist between language ability and pointing. In addition, dolphins and killer whales appear to use learned signals for maintaining social relationships and recognizing other individuals. Lastly, the songs of humpback whales can undergo rapid changes due to the effects of cultural transmission, consequently revealing their aptitude for social learning.

Given that humans are seemingly possessed of unique linguistic abilities, it remains unknown whether this faculty could develop in the oceans of Earth and other worlds. In this context, let us compare the rate of information transfer (denoted by $\mathcal{B}$) for humans and humpback whales, with the proviso understanding that these rates represent uncertain estimates at best and are therefore subject to variations depending on the methodology used. In the case of humans, the communication rate during speech appears to be $\mathcal{B}_H \approx 40$ bit / s across languages (Reed & Durlach 1998; Coupéet al. 2019). When it comes to humpback whales, the information content has been tentatively estimated to be ~ 130 bit / song (Suzuki et al. 2006). If we adopt a characteristic length of ~ 15 min for whale songs, we obtain an information rate of $\mathcal{B}_W \sim 0.14$ bit / s. Thus, the corresponding value of the information transmission rate for humans might be conceivably two orders of magnitude higher than humpback whales.

However, it would be a mistake to assume that the information transfer rate for *all* aquatic organisms (terrestrial or extraterrestrial) must be equally low. Instead, we propose that an idealized upper limit for $\mathcal{B}_{\max}$ in water could be obtained by assuming that it will be approximately equal to Shannon's channel capacity—a classical result from information theory.[22] Pursuing this line of reasoning, we have

$$\mathcal{B}_{\max} \approx \Delta\nu \log_2 \left(1 + \frac{S}{N}\right), \tag{7.75}$$

where $\Delta\nu$ is the bandwidth of the communication channel, and S/N is the signal-to-noise ratio (SNR). Although this formula essentially has only two free parameters, determining either of them is not an easy endeavor. In

22. Channel capacity was derived for additive white Gaussian noise (AWGN), but this is not likely to be valid in water. AWGN represents external white noise that is characterized by a normal distribution in the time domain.

the context of the SNR, the attenuation of the signal over distance must be taken into account. Furthermore, several components contribute to the total noise: for example, turbulence noise, thermal noise, and wave noise. In spite of these complications, it seems reasonable to argue that $S/N \sim 1$ is necessary for effective two-way communication, as otherwise the signals would be swamped by the noise. For this choice, we end up with $\mathcal{B}_{\max} \sim \Delta\nu$. On the one hand, humpback whale songs span more than 20 kHz and therefore have a remarkably high bandwidth, but on the other hand, most of the other whale sounds have bandwidths ranging between $\lesssim 100$ and ~ 1000 Hz (Thompson et al. 1986). Thus, in principle, information transmission rates much higher than $\mathcal{B}_W$, or even $\mathcal{B}_H$, might be attainable for extraterrestrial aquatic lifeforms. It must be reiterated that $\mathcal{B}_{\max}$ is dependent on numerous environmental factors as well as the bandwidth, thereby making it difficult to derive a generic upper bound.

We will wrap up our discussion by pointing out a couple of features making communication in water possibly more advantageous than in air. For starters, the speed of sound in water is higher. A fairly general expression for the speed of sound (denoted here by c_m) is

$$c_m = \sqrt{\left(\frac{\partial P_m}{\partial \rho_m}\right)_S} = \sqrt{\frac{K_m}{\rho_m}}, \tag{7.76}$$

where P_m, K_m, and ρ_m are the pressure, bulk modulus, and mass density of the medium, respectively; the subscript S on the right-hand side indicates that the expression is evaluated at constant entropy. Hence, even though the density of water is higher than air by a factor of $\sim 10^3$, its bulk modulus is about 2×10^4 higher at standard temperature and pressure conditions. As a result, the sound speed in water is about $\sqrt{20} \approx 4.5$ times higher compared to air. The second point that we wish to highlight is that sound is less attenuated in water compared to air. To illustrate this, we consider the idealized attenuation coefficient α_m computed by Sir George Stokes, whom we encountered in Section 7.6, for low-frequency sound waves in a fluid medium,

$$\alpha_m \propto \frac{\nu_m \omega^2}{c_m^3}, \tag{7.77}$$

where ω represents the angular frequency of the wave and ν_m is the kinematic viscosity of the medium. Thus, for a fixed frequency, we infer that

the attenuation coefficient for water is smaller than air by a factor of approximately 1000. The attenuation coefficient has units of inverse length, implying that sound will be subject to much less attenuation in water relative to air, and the signal can therefore travel over longer distances. However, the converse is true for electromagnetic radiation (light) indicating that visual sensing and communication are probably limited on ocean worlds. In particular, the theoretical visibility range in pure water is around 70 to 80 m when the photon wavelength is 550 nm, while the corresponding value for clear air is more than three orders of magnitude higher.

While none of the above reasons are by themselves sufficient to argue that language—or some equivalent version entailing symbolic communication—is capable of evolving on ocean worlds, water arguably offers more advantages than air in certain respects, from the standpoint of physics. Hence, one may contend that the putative absence of language in aquatic species on Earth does not imply that language is ineluctably a forbidden phenotype in marine environments.

7.8 NUMBER OF SUBSURFACE OCEAN WORLDS AND THE IMPLICATIONS FOR DETECTION

We will briefly discuss the implications concerning the commonality of worlds with subsurface oceans and comment on the prospects for their detection.

7.8.1 Number of planets with potential subsurface oceans

Let us begin by introducing some notation for later use. The variable N denotes the number and $\mathcal{P}$ represents the probability. We shall employ the subscripts "HZ" and "SO" to distinguish between rocky planets and moons located in the habitable zone versus those that could have subsurface oceans—that is, the Type B and U worlds introduced in Section 7.3.

We begin by estimating N_{HZ}, the total number of rocky planets in the HZ within our Galaxy. For calculating N_{HZ}, we rely on the data from the *Kepler* mission. Statistical analyses have yielded somewhat disparate results that are dependent on the spectral type of star considered, the limits of the HZ, and other factors. A summary of these findings for main-sequence stars can be found in Kaltenegger (2017). For our purposes, we adopt an estimate of ~ 0.1 rocky planets in the HZ per host star. This value errs slightly on the

side of caution, because it is approximately 50 percent of the corresponding fraction for M-dwarfs, which are the most common type of stars in our Galaxy. Then, using the fact that there exist $\sim 10^{11}$ stars in the Galaxy, we find

$$N_{\mathrm{HZ}} \sim 0.1 \times 10^{11} \sim 10^{10}. \tag{7.78}$$

Next, we seek to estimate N_{SO}, the total number of worlds in our Galaxy that could host subsurface oceans. This can be done by splitting this number into two components as follows: $N_{\mathrm{SO}} \approx N_{\mathrm{B}} + N_{\mathrm{U}}$, where N_{B} and N_{U} are the number of bound and free-floating worlds, respectively. However, because of the relative scarcity of data, it is rather difficult to estimate how many potential worlds with subsurface oceans exist outside the HZ of the host star.

Hence, at this stage, we will invoke a loose variant of the Copernican Principle, also referred to as the Principle of Mediocrity, for calculating N_{B}. We will therefore assume that our Solar system is not highly atypical and that the number of *potential* subsurface ocean worlds per star is roughly comparable to that of the Solar system. Here, it should be reiterated that we are not concerned with the actual existence of subsurface ocean worlds; instead, we are primarily interested in the possibility that a given world could theoretically have subsurface oceans. On a similar note, not all of the worlds within the HZ are automatically guaranteed to actually have liquid water on their surfaces; instead, the HZ should be regarded as the theoretical zone where surficial liquid water can exist.

As per the above reasoning, we calculate the total number of worlds within the range of $200 < R < 6400$ km in the solar system outside the HZ. Note that the lower and upper bounds reflect the radius of Enceladus and Earth, respectively, with the lower bound being chosen in light of the fact that Enceladus—and perhaps Mimas, which is virtually the same size as Enceladus—has a subsurface ocean. In carrying out this count, we should remember that not all Trans-Neptunian Objects (TNOs) have been detected. On the basis of our count, we find that there exist ~ 100 "planets" within the above range. Instead, if we increase the lower bound to a more reasonable ~ 500 km, the number drops to ~ 25. Barring Enceladus and possibly Mimas, most of the worlds that may possess subsurface oceans in our Solar system have $R \gtrsim 500$ km (Lunine 2017); therefore it seems more prudent to use this cutoff. With this set of assumptions, we arrive at

$$N_{\mathrm{B}} \sim 25 \times 10^{11} \sim 2.5 \times 10^{12}. \tag{7.79}$$

Next, turning our attention to $N_{\rm U}$, two strategies are open to us. The first relies on results from detailed numerical simulations, whereby the ejected number of worlds (with different masses) is computed for a wide range of initial planetary and debris disk configurations. At this juncture, we must impose a lower cutoff for free-floating worlds that have the potential to host subsurface oceans. Provided that the water inventories are sufficiently high, Type U worlds similar in size to Europa (or equivalently, our Moon) ought to be capable of maintaining oceans over Gyr timescales purely through radiogenic heating (Spohn & Schubert 2003). The simulations by Barclay et al. (2017) indicate that ~ 100 worlds with sizes greater than, or comparable to, the Moon are ejected from planetary systems with giant planets; the corresponding number drops to ~ 10 worlds in the absence of giant planets. Given that giant planets orbit ~ 20 percent of all stars, we find that ~ 30 worlds on average are ejected per star, thus yielding

$$N_{\rm U} \sim 30 \times 10^{11} \sim 3 \times 10^{12}. \tag{7.80}$$

The second method for inferring $N_{\rm U}$ relies on the recent discovery of the interstellar object 'Oumuamua by the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) in 2017.[23] The detection of this object enabled an empirical determination of free-floating objects in the Solar neighborhood. Current models suggest that the number density of objects may be $\sim 10^{14} - 10^{15}$ pc^{-3} (Do et al. 2018), although this estimate is subject to variability. For the purposes of our calculation, we will assume that the lower value of the number density ($\sim 10^{14}$ pc^{-3}) is typical throughout the Galaxy, which leads us to the conclusion that there exist $\sim 7 \times 10^{25}$ objects in the Milky Way. Although 'Oumuamua is very elongated and therefore has multiple length scales, the characteristic size can be chosen as ~ 0.1 km. We are interested in determining how many worlds with sizes greater than (or similar to) the Moon exist; that is, we must calculate the number of worlds with diameter > 3000 km. As per the collisional cascade model that describes the size distribution of objects taking part in inelastic collisions and fragmentation, the number density of objects with sizes greater than $2R$ (the diameter) exhibits power-law behavior with an

23. We will revisit this object and its successor, the interstellar comet 2I / Borisov, in Section 10.2.2.2.

exponent of −2.5 (Dohnanyi 1969). However, it is important to recognize that the populations of comets, asteroids, and KBOs display different power-law exponents. Nonetheless, proceeding with our analysis, we obtain

$$N_{\mathrm{U}} \sim 7 \times 10^{25} \left(\frac{3000}{0.1} \right)^{-2.5} \sim 4 \times 10^{14}, \tag{7.81}$$

and this is evidently higher than (7.80) by two orders of magnitude. Note that (7.81) also translates to $\sim 4 \times 10^3$ Type U worlds per star that are larger than, or comparable to, the Moon. Rather interestingly, predictions based on quasar microlensing indicate that there exist $\sim 2 \times 10^3$ worlds per star that are bigger than the size of the Moon (Dai & Guerras 2018). Despite this excellent agreement, we will proceed with the more conservative value for N_{U} computed in (7.80).

Finally, upon adding (7.79) and (7.80), we see that the total number of potential subsurface ocean worlds is $N_{\mathrm{SO}} \sim 5.5 \times 10^{12}$. From (7.78), it becomes apparent that N_{SO} is $\sim 10^3$ times higher than N_{HZ}. Thus, there are compelling grounds for conjecturing that the total number of worlds with subsurface oceans may dwarf the number of rocky planets with liquid water on the surface in our Galaxy by orders of magnitude. Hence, if this surmise is correct, we find ourselves immediately confronted with the following question: If worlds with subsurface oceans are much more common than rocky planets in the HZ, why do we find ourselves on the latter? There are a number of potential resolutions for this question. Clements (2018) touched on a few candidates, but we shall adopt a slightly different tack.

To begin with, it is quite conceivable that the probability of life on these subsurface worlds (denoted by $\mathcal{P}_{\mathrm{SO}}$) is selectively suppressed. For instance, this outcome is feasible if our analysis in Section 7.6 concerning the paucity of bioessential elements on these worlds is correct. Furthermore, a subtle distinction is inherent in the question posed above. In using the word *we* therein, we are actually referring to observers that can be placed in the same class as humans. In other words, we are dealing with intelligent, conscious, and technologically sophisticated species. Hence, even if $\mathcal{P}_{\mathrm{SO}}$ happens to be relatively high on subsurface ocean worlds, it could very well be that the emergence of species that share the same attributes as *H. sapiens* is rare, as we discussed in Section 7.7. The explanations offered here are by no means exhaustive, especially given that our knowledge of Earth-based intelligent marine life is far from complete.

7.8.2 On the likelihood of lithopanspermia

Lithopanspermia refers to the transfer of life by means of rocky material from one object to another. As it constitutes one of the primary subjects of Chapter 10, we will not delve into this topic in detail at this juncture. Most analyses are concerned with prospects for either interstellar or interplanetary panspermia: the former refers to the transfer of life between different stellar systems via rocks, and the latter represents the transport of organisms between two worlds orbiting the same star. Here, we will briefly expound a regime that overlaps both these domains.

We will tackle a two-step process wherein (1) a free-floating world is temporarily captured by a star by means of gravitational interactions, and (2) it subsequently seeds other planets or moons orbiting that star. The total probability $\mathcal{P}_{\rm tot}$ for this process can be estimated by means of the following equation, which shares some similarities with the Drake equation in SETI,

$$\mathcal{P}_{\rm tot} = \mathcal{P}_{\rm cap} \cdot \mathcal{P}_{\rm planet} \cdot \mathcal{P}_{\rm SO} \cdot \mathcal{P}_{\rm PS}, \tag{7.82}$$

where $\mathcal{P}_{\rm cap}$ denotes the capture probability of a free-floating world by a star in its lifetime, $\mathcal{P}_{\rm SO}$ is the probability that the captured world already has life, $\mathcal{P}_{\rm planet}$ is the number of worlds that could host life around that star but are not necessarily located within the HZ, and $\mathcal{P}_{\rm PS}$ represents the probability of successful interplanetary panspermia. $\mathcal{P}_{\rm planet} \sim 0.1$ if we restrict ourselves only to rocky planets inside the HZ. However, allowing for the possibility of seeding other subsurface ocean worlds, we set $\mathcal{P}_{\rm planet} \sim 10$, based on the discussion surrounding $N_{\rm B}$ in Section 7.8.1. It is not easy to calculate the typical value of $\mathcal{P}_{\rm cap}$ as it depends on the relative velocity of the approaching object, stellar mass, the inclination angle, and other factors. Numerical simulations undertaken by Goulinski and Ribak (2018) seem to suggest that $\mathcal{P}_{\rm cap} \sim 0.01$ is a reasonable estimate, but this depends on multiple parameters.

Using these values, we find $\mathcal{P}_{\rm tot} \sim 0.1 \cdot \mathcal{P}_{\rm SO} \cdot \mathcal{P}_{\rm PS}$, but we encounter an immediate stumbling block since the magnitudes of the remaining two variables are highly uncertain. Although the captured world will not ineluctably stay bound to the planetary system for long, this may be offset by the stance that interplanetary panspermia is conventionally regarded as being more likely than its interstellar counterpart. For instance, within our Solar system, several authors have concluded that Mars and Earth had a high likelihood of having exchanged microbial life in the past (Mileikowsky

et al. 2000). Similarly, the likelihood of interplanetary panspermia in densely packed planetary systems that are akin to TRAPPIST-1 might be orders of magnitude higher with respect to the transfer of life from Earth to Mars (Lingam & Loeb 2017a). In contrast, the viability of interstellar panspermia is poorly understood, but it appears to be less likely. Finally, we are left with the variable $\mathcal{P}_{\mathrm{SO}}$, which remains virtually unconstrained given that we do not fully understand how abiogenesis occurred even on the Earth. However, if we discover subsurface exolife on Europa, Titan, or Enceladus, it may improve our understanding of $\mathcal{P}_{\mathrm{SO}}$.

In order to assess the total number of stellar systems that have been seeded this way through panspermia, we must multiply $\mathcal{P}_{\mathrm{tot}}$ with 10^{11}, which is clearly a large number. Hence, if we choose $\mathcal{P}_{\mathrm{SO}} \sim 10^{-3}$ and $\mathcal{P}_{\mathrm{PS}} \sim 10^{-2}$, we find that $\sim 10^5$ stellar systems might have been seeded with life. Needless to say, these values have been chosen arbitrarily and the actual value could be much higher ($\lesssim 10^{10}$ stars) or lower (close to zero). Future statistical surveys may permit us to assess the likelihood of panspermia processes by looking for signs of clustering (Lin & Loeb 2015). Alternatively, we can carry out in situ explorations of interstellar objects that have been trapped by the Sun-Jupiter system, which serves as an effective "fishing net". We will not delve into this aspect further, as it is analyzed in Lingam and Loeb (2018a) and in Section 10.3.1.

7.8.3 Prospects for detecting subsurface ocean worlds

We will confine ourselves to Type U worlds because there are already several missions and proposals underway for in situ exploration of Type B worlds within our own Solar system, as noted in Section 7.1. One of the possible differences that could exist between Type U worlds and Type B planets and moons within our Solar system is that the initial conditions for their formation were possibly different—for example, formation in gas-starved versus gas-rich disks. These distinctions, in turn, have ripple effects and influence the subsequent geological, chemical, and biological evolution of these worlds, thereby providing a motive for the detection and study of Type U worlds.

On the basis of our analysis leading up to (7.80) and (7.81), the number of free-floating worlds (with $R \gtrsim 0.3R_{\oplus}$) is about 30 to 4000 times higher than the number of stars in the Milky Way. As we know that the nearest star, Proxima Centauri, is ~ 1 pc away, we suggest that the nearest such object

might be located at a distance of $\langle r \rangle \sim 0.01 - 0.1$ pc from the Earth using the scaling $\langle r \rangle \propto N_{\rm U}^{-1/3}$. This distance translates to $\langle r \rangle \sim 2 \times 10^3 - 2 \times 10^4$ AU, and the lower bound is actually comparable to the inner edge of the Oort cloud ($\sim 2{-}5 \times 10^3$ AU) and the aphelion of certain TNOs such as Sedna (~ 900 AU). By adopting the lower bound for $\langle r \rangle$, we can compute the thermal flux that would be received on Earth via

$$S_{\rm max} \approx 1.5\,{\rm mJy} \left(\frac{T_s}{40\,{\rm K}}\right)^3 \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\langle r \rangle}{2000\,{\rm AU}}\right)^{-2}. \tag{7.83}$$

The flux density $S_{\rm max}$ has been computed at the blackbody peak (Wien maximum); therefore, we have $\lambda_{\rm max} \approx 126$ μm for the characteristic temperature of ~ 40 K for Type U worlds that are roughly Earth-sized. Hence, it can be seen that the value of $\lambda_{\rm max}$ lies within the far-IR range, and several telescopes are operational at such wavelengths.

The maximum distance at which Earth-sized free-floating planets can be detected is ~ 830 AU for both the Vera C. Rubin Observatory and Pan-STARRS (Abbot & Switzer 2011), suggesting that the characteristic distances of the nearest Type U world fall below the detection threshold. The maximum sensitivity of the Photoconductor Array Camera and Spectrometer (PACS) built for the Herschel Space Observatory is a few milliJansky, indicating that it may not be capable of detecting such worlds. In contrast, the upcoming Cerro Chajnantor Atacama Telescope (CCAT) has been predicted to reach a sensitivity of ~ 0.36 mJy for a wavelength of 200 μm,[24] which is slightly lower than the value of $S_{\rm max}$ obtained above. Hence, it seems plausible that upcoming telescopes could, in principle, detect such free-floating planets, although the chances for success are sensitive to the number density of objects in the solar neighborhood.

Even if we do detect such worlds, the chances for finding any tangible biomarkers, and determining the interior composition and structure, are very slim. The major difficulty stems from the datum that life would be located deep beneath the surface, and scientists would not be able to search for atmospheric biosignatures like oxygen and methane as these worlds are not likely to have sufficiently dense atmospheres. If these worlds emit plumes akin to those of Enceladus, and perhaps Europa, it may be possible to search for biomarkers therein, but a significant difficulty arises from the fact that

24. See CCAT-prime, http://www.ccatobservatory.org/index.cfm

the photon flux (from internal heating) received at Earth scales as $\langle r \rangle^{-2}$; for reflected light from the Sun, the scaling becomes $\langle r \rangle^{-4}$. Hence, ceteris paribus, the flux from the nearest Type U world would be 10^{-5} or 10^{-6} that of Enceladus—unless we serendipitously discover the object at a much closer distance to our planet, in which case the plumes or cometary tails produced by solar heating can be analyzed spectroscopically for biomarkers. Furthermore, as noted in Section 7.8.2, undertaking in situ explorations of Type U worlds captured by the Jupiter-Sun fishing net also appears to be a practical endeavor.

Hence, from a long-term standpoint, perhaps the most likely means of exploring these worlds will be through small spacecrafts that undertake flyby missions. These spacecraft could be powered by light-sail technology, along the lines of the recently announced *Breakthrough Starshot* project.[25] If such technology becomes feasible, traveling at 20 percent the speed of light, such a spacecraft might be able to reach the nearest Type U world in a span of about one year.

7.9 CONCLUSION

If, Ocean, you could grant, out of your gifts and dooms,
some measure, fruit or ferment for my hands,
I'd choose your distant rest, your brinks of steel,
your furthest reaches watched by air and night,
. . . . . .
Your petals throb against the world,
your submarine crops tremble,
the smooth algae brood like a menace,
the schools navigate and propagate.

—Pablo Neruda, *Canto General,* Canto XIV, *El Gran Oceano*

When it comes to looking for life beyond Earth in our Solar system, three of the best candidates (Europa, Titan, and Enceladus) are moons with subsurface oceans of liquid water. Motivated by this fact, in this chapter we have evaluated the prospects for life on worlds with subsurface oceans. However, before embarking on the précis, we caution that we currently remain in the

25. See Starshot, http://breakthroughinitiatives.org/initiative/3

dark regarding multifarious aspects of aquatic biospheres. To put it more whimsically, we know very little of the "smooth algae" and "submarine crops" that inhabit extraterrestrial oceans, to quote Pablo Neruda's poetic phrases.

These worlds can be classified into two broad categories: those orbiting a host star and those traversing the cold reaches of interstellar space. Collectively, our analysis indicates that they may outnumber Earth-sized planets in HZs by a factor of almost a thousand. If subsurface ocean worlds are capable of giving rise to life, regardless of its complexity, studying such worlds is of paramount importance. We commenced this study by calculating the temperature profile of such worlds and demonstrated that oceans underneath icy shells of moderate thickness (on the order of a few tens of kilometers) could exist in a wide range of habitats ranging from the Galactic center to the high-redshift Universe.

Although we are far from having identified all the requirements for life, some the essential criteria are known to us. One of them is the availability of a liquid solvent, which is automatically satisfied on worlds with subsurface oceans, although the water itself may not be habitable (e.g., due to high acidity). Life also requires sources of free energy, sensu lato, for powering prebiotic chemistry and sustaining ecosystems. We contend that a wide range of energy sources are theoretically suitable for stimulating the formation of various organic compounds that constitute the backbone of biology on Earth. Some of the notable ones include ionizing radiation, exogenous delivery of organics, and radiogenic heating.

For each of these sources, we computed the corresponding energy fluxes and presented heuristic estimates for the quantity of amino acids that could be produced. However, most of the pathways delineated drive the formation of organic compounds on the icy surface, where they are susceptible to destruction by radiation and energetic particles. Hence, we briefly sketched methods whereby they may be transported across the icy shell into the ocean. In this context, we also examined how the unique properties of ice can play a beneficial role in the concentration and polymerization of these molecules, thereby playing a potentially invaluable role in bringing about abiogenesis.

Subsequently, motivated by energetic considerations, we evaluated the biological potential of these worlds by calculating the maximum amount of biomass that could be supported by different sources. A diverse array

of mechanisms have the capacity to support biospheres by virtue of the free energy derived from (1) the formation and delivery of oxidants from the surface, (2) hydrothermal vents, and (3) the radiolysis of water. Under certain circumstances, it is conceivable that a redox balance (as documented on Earth) may persist over geological timescales, although its likelihood is probably low for generic icy worlds. For most of these energy sources, we concluded that the rate of biomass production was likely to be sizable but ostensibly several orders of magnitude lower than on Earth. Over a wide spectrum of subsurface ocean worlds, we found that the dissolved oxygen levels might approach that of the Earth in sub-Gyr timescales and that they seem capable of supporting large numbers of macrofauna (akin to fish).

Apart from the presence of energy sources and water, the availability of elements like phosphorus (P), in the form of phosphates, is essential for life. By formulating a simple mathematical model for the availability of P in the oceans, we showed that its steady-state concentration is conceivably a few orders of magnitude below that of Earth, provided that these worlds are relatively large and have alkaline oceans (with $\mathrm{pH} \gtrsim 8.0$) and hydrothermal activity. Even though our results indicate that the putative biospheres have low nutrient concentrations (oligotrophic), and are perhaps sparsely populated, they do not rule out the prospects for life altogether. While our analysis was applicable to subsurface oceans, similar conclusions are plausible for ocean planets with surface oceans, as noted in Section 5.5. As a consequence, the limiting factor for complex biospheres on ocean planets might not be energy—due to the plenitude of electromagnetic radiation from the host star—but the lack of bioessential elements such as phosphorus.

Thereafter, we undertook a qualitative exploration of the evolutionary trajectories feasible on ocean worlds (either surface or subsurface). Considerable empirical evidence on Earth indicates that most of the recent major evolutionary innovations arose in water much later than on land (or never). Hence, there are some grounds for supposing that these breakthroughs have a lower probability of transpiring on ocean worlds. We attempted to gauge whether some of the traits usually associated with humans are capable of manifesting in aquatic species. It is our belief that extraterrestrial life with self-awareness and culture can, at least in theory, evolve in oceans. In contrast, the likelihood of evolving language with syntactic structure (despite certain advantages inherent to communication in water) is probably very small, and the attainment of human-level technology could be even rarer. With that said, as the old adage goes, "never say never." One day in the

future, however distant, we may come across an intelligent aquatic species with whom goal-oriented communication and information exchange is rendered practical, perhaps even on our own planet, where they might have remained hidden in plain sight all along.[26]

If we were to summarize our entire analysis in a sentence, it would be this: life on worlds with subsurface oceans is likely to confront sui generis challenges not prevalent on Earth, but none that are so severe to preclude it from existing altogether. Some of the intrinsic difficulties, especially when it comes to complex multicellular life, include the absence of abundant energy sources (akin to solar radiation) and the limited availability of bioessential elements. Collectively, they may engender the presence of patchy oligotrophic biospheres with low biomass densities, a prospect that could seem unappealing at first sight, given the richness of our own biosphere. Nonetheless, it should be recalled that these worlds are, in all probability, far more common than rocky planets in the HZs of stars.

Hence, even if we exclude all other reasons aside from the last one, there is a genuine necessity for devoting more efforts toward theoretical modeling, laboratory experiments, and carrying out in situ explorations of subsurface ocean worlds in our Solar system. Laboratory experiments, in particular, that seek to simulate environmental conditions akin to Europa and Enceladus will play a major role in enhancing our understanding of the most optimal habitats for the genesis and maintenance of extraterrestrial life in subsurface oceans (Taubner et al. 2020). After all, in doing so, we will take one further step vis-à-vis determining whether the Universe is teeming with life, virtually devoid of it, or somewhere in between.

---

26. In this respect, one is immediately reminded of Douglas Adams's famous science-fiction series, *The Hitchhiker's Guide to the Galaxy,* wherein dolphins (and mice) were humorously posited as being far more intelligent than humans.

PART 3

ASPECTS OF EXTRATERRESTRIAL TECHNOSPHERES

Chapter 8

THE DRAKE EQUATION AND FERMI'S PARADOX

> Now while the great thoughts of space and eternity fill me I will
> measure myself by them,
> And now touch'd with the lives of other globes arrived as far
> along as those of the earth,
> Or waiting to arrive, or pass'd on farther than those of the earth,
> I henceforth no more ignore them than I ignore my own life,
> Or the lives of the earth arrived as far as mine, or waiting to arrive.
>
> —Walt Whitman, *Night on the Prairies*

It has become something of a truism nowadays to state that humans have looked up at the night sky since times immemorial and wondered, "Are we alone?" Yet, the relevance of this question has not diminished despite (or perhaps because of) eons of speculation. In the current epoch, we have witnessed heightened interest in resolving this fundamental question, as we are approaching a critical juncture when the detection of extraterrestrial life is no longer in the realms of science fiction.

However, in referring to *extraterrestrial life,* humanity would, in all likelihood, attach much more significance to the discovery of "intelligent" life than to finding microbes. While this viewpoint can be ascribed to anthropocentrism, as well as manifold complex sociological and theological factors that will not be discussed here,[1] it is worth recalling what the eminent paleontologist George Gaylord Simpson had to say concerning the former in his well-known work *The Meaning of Evolution* (1967, p. 286):

---

1. A detailed account of the historical debate over extraterrestrial life and the accompanying sociocultural implications can be found in *The Biological Universe* (1996) by Steven J. Dick.

> And even if he were the lowest animal, the anthropocentric point of view would still be manifestly the only proper one to adopt for consideration of *his* place in the scheme of things and when seeking a guide on which to base *his* actions and his evaluations of them.

From an anthropocentric perspective, one might therefore be inclined to argue that, even if intelligent life (possibly, but not ineluctably, endowed with consciousness) was rare in the Universe compared to "primitive" (i.e., non-intelligent life) life, this would be counterbalanced by the fact that its detection would carry a much greater significance insofar as humans are concerned, although whether the cumulative impact on humanity's collective psyche will be positive or negative has been the source of much debate.

There is, perhaps, a more interesting and pragmatic reason why the search for intelligent extraterrestrial life may deserve to be accorded a higher weight compared to primitive life in some respects. In 1990, the *Galileo* spacecraft undertook a flyby of Earth en route to Jupiter and its moons. A team of astronomers, led by Carl Sagan, had the innovative idea of envisioning Earth as a target in the search for life, both primitive and intelligent. They discovered that atmospheric methane was in a state of high thermodynamic disequilibrium and found evidence for a surface pigment (chlorophyll) that displayed a distinctive absorption edge in the red wavelengths of the electromagnetic spectrum. While both of these factors were seen as being strongly indicative of life on Earth, the most unambiguous evidence was manifested in the form of "narrow-band, pulsed, amplitude-modulated radio transmission" (Sagan et al. 1993). This result has often been invoked by proponents of the search for extraterrestrial technological intelligence (SETI), a broad topic that comprises the bedrock of Chapter 9, to argue that the pursuit of technosignatures—to wit, signatures of advanced extraterrestrial technology—is inherently advantageous, since they allegedly have a much lower likelihood of originating from nonengineered (natural) phenomena. In our discussion of Fermi's paradox in Section 8.2, we will reconsider this assumption and argue that its universality merits a certain degree of skepticism.

Irrespective of the actual motivation and the relative importance each of us ascribes to carrying out SETI, what cannot be denied is that the field has witnessed a renaissance of sorts in the last few years. While this upswing is partly attributable to renewed sources of funding, the improved technology

of the twenty-first century has played an important role as well. It has not only permitted scientists to undertake searches for technosignatures with greater reach and precision, but it has also empowered us to imagine and conceive new technosignatures (e.g., photovoltaic cells, industrial pollution) on the basis of humanity's own technological developments. Yet, it must be borne in mind that both the practitioners and the field of SETI endured many "slings and arrows" (à la Shakespeare's Hamlet) toward the end of the twentieth century. A major reason behind this downturn in fortunes and federal funding was the apparent absence of any positive results after decades of searching for radio signals. In order to understand the reasons underpinning this outwardly null result, we must begin our discussion by seeking to quantify the number of extraterrestrial technological intelligences (ETIs) that are currently active and detectable.

It behooves us now to take a detour and address the notation employed herein. In the older SETI literature, it was common to conflate *species* with *civilization,* but the former and latter have very specific connotations in the natural and social sciences, respectively. With regard to the latter, one might unknowingly fall into the *unilineal evolution* trap and presume that all extraterrestrial societies will pass through the so-called savage and barbaric phases before entering the civilized era. Moreover, while humans *are* a technological species, not all human civilizations (present or past) have attained the same level of technology. In fact, by equating these two primary concepts, we would be implicitly subscribing to the premise that all members of a given technological species are part of a planet-spanning single civilization: although this situation is certainly feasible, it is by no means guaranteed to be valid for every species at every point in its evolutionary history. We could expostulate likewise at length, but we shall not belabor the point further. Hence, we will employ *technological species* or *ETIs* hereafter, although it must be acknowledged that these phrases are not always apropos and are laden with their own limitations.

8.1 THE DRAKE EQUATION

8.1.1 A history of the Drake equation and its classical form

In 1959, the physicists Giuseppe Cocconi and Philip Morrison authored a seminal paper in which they argued that (1) extraterrestrial species with technology comparable to, or more advanced than, our own exist and

(2) they may broadcast interstellar communications at the radio frequency of 1.42 GHz (with the wavelength of $\lambda = 21$ cm) since it corresponds to the famous emission line of neutral hydrogen. Shortly thereafter, in 1960, Frank Drake conducted a search, whimsically christened Project Ozma,[2] of the two nearby Sunlike stars Tau Ceti and Epsilon Eridani for a total duration of about two hundred hours at the National Radio Astronomy Observatory (NRAO) situated in Green Bank, West Virginia (F. Drake 1960, 1961).

The positive scientific and public reception of Project Ozma, in conjunction with a lecture given earlier by Drake, led to an informal conference being sponsored by the National Academy of Sciences on November 1–2, 1961, at the NRAO Green Bank facility. As the central theme of the conference was "Extraterrestrial Intelligent Life," Drake came up with a heuristic equation to estimate the current number of species that were capable of interstellar radio communication (1965). Drake's equation, which continues to bear his name, has appeared in a multitude of forms, with the canonical version usually expressed as

$$N_s = R_\star \cdot f_p \cdot N_e \cdot f_l \cdot f_i \cdot f_c \cdot L, \tag{8.1}$$

where the various terms are delineated as follows:

- N_s represents the number of species in our Galaxy that are expected to possess the necessary technology for undertaking interstellar communication.
- $R_\star$ is the rate of star formation in our Galaxy.
- f_p denotes the fraction of stars that have planetary systems.
- N_e is the number of planets per planetary system that possess environmental conditions that are potentially amenable to the origin and evolution of life.
- f_l represents the fraction of planets on which life actually arises at some point.
- f_i is the fraction of life-bearing planets on which intelligence actually emerges.

2. The project adopted its name from Princess Ozma, a central character in the famous Oz novels authored by L. Frank Baum.

- f_c denotes the fraction of intelligent species that develop a sufficiently high level of technology that enables them to take part in interstellar communication.
- L is the duration over which these species produce detectable signals.

Although the Drake equation represented the first systematic, albeit simplified, framework for assessing the number of detectable ETIs, a few important papers that tackled these issues appeared around the same time. For example, Ronald Bracewell utilized a graphical method to infer the probability of finding technological species per star (Bracewell 1960), while Sebastian von Hoerner employed a probabilistic approach to calculate the average lifetime of ETIs, concluding it was $\sim 10^4$ yrs (von Hoerner 1961).

An inspection of (8.1) reveals that the first three factors are astrophysical in nature, the fourth and fifth factors are biological, and the last two factors are dependent on sociology. We have now progressed to the stage where we can calculate fairly robust estimates for the first three factors, but the other four remain highly uncertain, notwithstanding major advances in our understanding of the origin of life, evolutionary biology, and ecology. With this caveat in mind, we will next attempt to summarize the current estimates for these factors.

8.1.2 The factors in the Drake equation

For a detailed analysis of the various factors in the Drake equation from both historical and scientific standpoints, the reader is directed to the monograph *The Drake Equation* by Vakoch and Dowd (2015).

The first factor, the star formation rate $R_\star$, was reasonably well-known even in 1961. A simple method for estimating it is to divide the total number of stars in the Milky Way ($\sim 10^{11}$) by its age (10^{10} yr). Upon doing so, we find $R_\star \sim 10$/yr. It should be noted that $R_\star$ calculated using this process only represents a mean value. In reality, $R_\star$ varies both spatially and temporally, thereby making it difficult to assign it a characteristic value. The magnitude of $R_\star$ is estimated through a combination of direct and indirect techniques. The former often relies on main-sequence fitting of luminous stars, a procedure essentially equivalent to counting the number of stars (of different masses) in a given age and thereby calculating the average of $R_\star$. The indirect strategies are many and range from determining the frequency

of core-collapse supernovae (through gamma ray observations) to observations of UV radiation arising from the photospheric emission of young stars. Collectively, these approaches have concluded that $R_\star \sim 1\, M_\odot/\mathrm{yr}$ in the Milky Way, corresponding to $R_\star$ being a few stars per year. We shall adopt the slightly conservative value of $R_\star \sim 1/\mathrm{yr}$ henceforth.

The next quantity that we consider is the fraction of stars with planets f_p. This parameter remained in the realm of speculation until quite recently (i.e., about ten years ago) but the situation changed dramatically after the launch of the *Kepler* satellite, which has discovered thousands of exoplanets. *Kepler* uses the transit method, explained in Section 6.1, thus measuring tiny variations in the host star's brightness as the planet passes in front of it and facilitating the detection of exoplanets and their sizes. Statistical surveys that use gravitational microlensing—where planets are identified due to additional spikes in brightness of a background star that they produce during the lensing event—have provided another line of inquiry to determine f_p. All of these techniques have independently converged on the expected value of $f_p \sim 1$.

In order to gauge N_e, the number of planets that might possess the right conditions for hosting life, it is necessary to recall the concept of the circumstellar habitable zone (HZ). The HZ is the region around the host star that is theoretically capable of supporting liquid water on the surface of a rocky planet. However, at this stage, some important points should be recognized. First, the location and width of the HZ is heavily dependent on both planetary and stellar parameters and also evolve over time. Furthermore, not all planets within the HZ are guaranteed to actually have liquid water on their surfaces, and other factors (e.g., tidal locking of planets around low-mass stars) can have a detrimental effect on planets within the HZ. Lastly, planets and moons outside the HZ could still have conditions amenable to life. Two quintessential examples that fall under this category in our Solar system are Europa and Enceladus, both of which possess subsurface oceans, raw materials, and multiple energy sources to power prebiotic chemical reactions. Bearing these factors in mind, analyses of the *Kepler* data have enabled us to determine the fraction of stars that have rocky (approximately Earth-sized) planets in their HZs. If we interpret this fraction as being roughly equivalent to N_e, most studies suggest that $N_e \sim 0.1$, give or take a factor of a few (Kaltenegger 2017; Zink & Hansen 2019).

We have completed our review of the astronomical factors in the Drake equation. The next two factors, f_l and f_i, are notoriously hard to determine.

We shall discuss these two factors in greater detail when we turn our attention to Fermi's paradox in Section 8.2. The fact that life originated on Earth relatively quickly—less than 0.8 Gyr after the Earth became habitable, while the Earth's total habitability interval is predicted to be ~ 6 Gyr—has been interpreted as evidence by some authors that abiogenesis is relatively common in the Universe. However, this claim is very sensitive to the choice of prior—to wit, the probability distribution for abiogenesis that one assumes a priori (Spiegel & Turner 2012). Hence, in a nutshell, we are not yet at the stage where we can rigorously claim that either one of $f_l \sim 1$ or $f_l \ll 1$ is valid (cf. Kipping 2020).

Next, we turn our attention to the fraction of life-bearing planets on which intelligence arises. Estimations of f_i readily encounter potential difficulties because the word *intelligence* could be interpreted in manifold ways, as elucidated in Chapter 3. Many practitioners and opponents of SETI alike have implicitly, and even explicitly in a few cases, subscribed to the premise that there exists an unambiguous ladder of increasing complexity from "lower" organisms to "higher" ones, finally culminating in *Homo sapiens* on Earth. Scientists who argue in favor of a high value of f_i have, at times, suggested that the emergence of intelligence is both inevitable and frequent, predicated on the above line of reasoning. On the other hand, many distinguished evolutionary biologists have contended that the appearance of humanlike intelligence is extremely rare (i.e., effectively amounting to $f_i \ll 1$) by appealing to evolutionary contingency, thereby asserting that the large number of improbable events that led to the advent of humans has a low probability of occurring elsewhere. We will return to this subject in Section 8.2.4, when discussing Fermi's paradox.

At first glimpse, both camps appear to tacitly accept the concept that human intelligence is special. In reality, this situation is arguably more complex for a number of reasons. Advanced tool-making abilities and language are ostensibly unique to humans, but the situation is less clear when it comes to both intelligence and culture, as explicated in Section 3.7. We will not delve into this topic further, as we already covered it in the Preface when we distinguished between the simplistic notions of *primitive* and *intelligent* forms of extraterrestrial life. Essentially, this debate comes down to the differences between human and nonhuman minds, which has been the source of fierce debate between those who claim that the divergences are "one of degree and not of kind," à la Darwin, and others who contend that there exist major discontinuities between the two (see Section 3.8).

We now turn our attention to arguably the murkiest two factors: f_c and L. Before tackling them, an important item is worth highlighting. While f_i typically deals with the likelihood of the evolution of *intelligent* species, f_c encapsulates the number of *technological* species that attain a sufficiently high level of (technological) proficiency to be detectable by producing appropriate signatures. Hence, as opposed to many analyses that implicitly equate technology and intelligence, a more suitable moniker for f_i may actually be f_t—to wit, the fraction of species that develop advanced tool-making ability. Although this matter seems to be trivial and purely centered on terminology, it serves to illustrate the inclination toward vagueness in certain discussions of the factors making up the Drake equation.

One of the most peculiar aspects about f_c and L is that these values are equally dependent on the species under consideration as well as on the *observing* species. In other words, determining what constitutes a detectable ETI and the duration over which it remains detectable is very much driven by the technological sophistication of the observers.[3] To illustrate this point more concretely, consider the following example. If we suppose that a powerful FM radio or TV antenna has an effective isotropic power (i.e., the equivalent power radiated by an isotropic source) of $P_a \sim 10^6$ W and operates over a bandwidth of $\Delta\nu \sim 5$ MHz, the average flux density S_p at a distance $d_s \sim 30$ pc (~ 100 light yr) is given by

$$S_p = \frac{P_a}{4\pi\, d_s^2\, \Delta\nu} \sim 2 \times 10^{-12}\,\text{Jy}. \tag{8.2}$$

We have chosen $d_s \sim 30$ pc since it represents a useful fiducial value often used in discussions pertaining to the Drake equation. This value of S_p is about ten orders of magnitude lower than the detection limits of the Low-Frequency Array (LOFAR) radio telescope in the band used by FM radio and television. On the other hand, if we were to use the Arecibo transmitter in the radar mode, it would be detectable by current Earth-level SETI technology at a distance of ~ 1 kpc. Yet another complicating factor arises because the leakage can actually decline with time when the

3. In this respect, one could rightly argue that observers are necessary for f_c and L to even make sense. This position is rather reminiscent of John Wheeler's Participatory Anthropic Principle (PAP), in which the existence of observers was posited as a necessary criterion for the Universe to come into being.

species matures technologically—the adoption of spread-spectrum transmissions and digital technologies contributed to this trend in human history. On the whole, it seems unlikely that a species with human-level technology would be able to detect the leakage of electromagnetic signals from another human-level ETI around 30 pc away. On the other hand, if electromagnetic signals were deliberately broadcast using the most powerful transmitters, a species with human-level technology could probably be discerned by species with a similar level of technological advancement, in theory. As a result of these complications, affixing any particular value to f_c is clearly premature.

Finally, we turn our attention to L, which shares many of the difficulties that plague estimates of f_c. An important difference between our definition of L and those in prior studies deserves mention: the latter invariably refer to *civilizations,* while we have chosen to interpret the Drake equation through the framework of *species*. As a result, one ought not draw on analogies with human civilizations on Earth in a simplistic manner and thereby attempt to infer the value of L from such studies. In most earlier analyses, L was claimed to range between $\sim 10^3$ and $\sim 10^8$ yr, with many favoring values closer to the upper bound because of a handful of alleged supercivilizations that would have attained long-term stability by passing through a transient unstable phase. The reader may wish to consult the classic treatise *Intelligent Life in the Universe* by Shklovskii and Sagan (1966) for succinct arguments in support of a high value of L.

As the reader has gained some appreciation of the myriad complexities and uncertainties involved by now, we will present a possibly *optimistic* estimate for the number of technological communication-capable species using (8.1). In addition to the three astronomical factors, whose values are adequately well constrained, we choose $f_l \sim 0.1, f_i \sim 1, f_c \sim 1$, and $L \sim 10^6$ yrs. Upon multiplying all of the factors, we end up with

$$N_s = 1 \times 1 \times 0.1 \times 0.1 \times 1 \times 1 \times 10^6 \sim 10^4. \tag{8.3}$$

In contrast, choosing a short lifetime of $L \sim 100$ yr would yield $N_s \sim 1$, an estimate that is compatible with the lower bound computed by Westby and Conselice (2020) in a detailed analysis of the Drake equation. The average separation between two species d_s can be found by assuming that each occupies a volume of $4\pi d_s^3/3$ and that there exist N_s of them in the Galaxy. In quantitative terms, this relation is expressible as

$$\frac{4\pi d_s^3}{3} \times N_s \sim \pi R_g^2 H_g, \tag{8.4}$$

where $R_g \sim 15$ kpc and $H_g \sim 0.35$ kpc are the radius and scale height (for the thin disk) of the Milky Way, respectively. Upon further simplification, we obtain

$$d_s \sim 180\,\mathrm{pc} \left(\frac{N_s}{10^4}\right)^{-1/3}. \tag{8.5}$$

Lastly, we observe that many classical analyses of the Drake equation hypothesized that $N_s \sim L/(1\,\mathrm{yr})$. If we assume this relation to be valid and use the values for the three astrophysical factors we have outlined earlier in conjunction with (8.1), we arrive at

$$f_l f_i f_c \sim 10. \tag{8.6}$$

However there is some ambiguity in identifying exactly what constitutes N_e—it could be either much smaller or somewhat greater than our choice of $N_e \sim 0.1$—and the exact value of $R_\star$ is higher than unity. Nevertheless, in order for (8.6) to be satisfied, it becomes apparent that each of the f's must be sizable and at least one of them must be higher than unity; this expectation still holds true as long as the right-hand side is around unity or higher. This seems rather implausible, so we can therefore claim that $N_s \sim L/(1\,\mathrm{yr})$ is very unlikely.

8.1.3 Some critiques of the Drake equation and the proposed solutions

The Drake equation was the first systematic effort to encapsulate the many variables involved, and it should be viewed as a valuable heuristic rather than a comprehensive law. In certain scientific quarters, there has been a tendency to sharply (and sometimes unjustly) criticize and dismiss the Drake equation. On the other hand, it has become a mainstay of popular science and has been subject to much overuse (and abuse) therein. It would therefore seem appropriate for SETI practitioners to recognize the Drake equation's worth, while also undertaking critiques and extensions of this formalism. Here, we will discuss some limitations of the Drake equation and its proposed extensions.

Fraction of stars suitable for life: After Drake's original formulation in 1961, it was recognized soon afterward that the use of $R_\star$ was an oversimplification since it quantified the total star formation rate (either instantaneous or time-averaged). Hence, this formulation ignored the fact that not all stars were necessarily conducive to planets capable of hosting life. The recognition of this limitation was partly spurred by the publication of Stephen Dole's prescient *Habitable Planets for Man* (1964), which pointed out that tidally locked planets would be common around low-mass stars and that this could pose difficulties to habitability. As a result, when Shklovskii and Sagan published *Intelligent Life in the Universe* (1966), their version of the Drake equation had the factor f_g, which encapsulated the fraction of stars that were suitable for life to originate on planets orbiting them.

Determining which stars are conducive to hosting habitable planets is an ongoing area of research. In particular, the habitability of planets orbiting M-dwarfs (which usually have $M_\star \lesssim 0.5\,M_\odot$) is a contentious topic that has witnessed many swings in opinion. The latest evidence appears to suggest that a combination of strong stellar winds, high activity, and stellar flares drives the erosion of planetary atmospheres and their ozone shields over relatively short timescales; we shall not delve into this topic further, as we already addressed it in Chapter 4. If we do assume, as many authors have now suggested, that M-dwarfs are not suitable for having life-bearing planets, then $f_g \sim 0.1$. On the other hand, if M-dwarfs are equally likely to have habitable planets, we would have $f_g \sim 1$. For the former choice, the value of N_s in Section 8.1.2 should be lowered by about an order of magnitude due to the inclusion of $f_g \sim 0.1$.

Fraction of species with whom communication is actually possible: To its credit, the Drake equation yields the total number of species that possess the capacity for detectable interstellar communication. However, the equation remains silent on whether all of these species are actually *accessible* to us insofar as communication is concerned. There are several reasons why a particular ETI may not communicate with us, even assuming that it has the technology to do so. First, the cognitive processes that led to its emergence could be so different from humans' that it is effectively out of the question to participate in any mutual exchange of information. Even on Earth, while it is generally accepted that humans are unique in using symbolic language with syntax and grammar, there is much that remains unknown about animal communication in general. Second, the possibility that some

ETIs do not have any desire for interstellar communication cannot be ruled out. Lastly, if a species is technologically far advanced compared to humans, the chances of effective communication might be diminished because of the vast chasm in knowledge and intelligence (Döbler 2020). Thus, it will be necessary to introduce another factor in the Drake equation if one wishes to calculate the total number of species that not only have the requisite technological expertise for interstellar communication but can also send messages that are detectable and comprehensible.

Sites of abiogenesis are not inescapably planets around stars: In our own Solar system, it has been concluded that there exist a fairly high number of moons outside the HZ that possess liquid water in the form of subsurface oceans; some of them, such as Enceladus and Europa, appear to have the raw ingredients necessary for life as well as suitable energy sources. Another possibility is that there are exomoons in the HZ of giant planets—these moons could possess oceans on their surface because of the warmer temperatures in this region. Owing to the absence of statistics, we cannot yet determine the number of HZ exomoons, but some authors have suggested that they may outnumber HZ exoplanets. As the Drake equation focuses on planets (specifically, rocky ones) as the potential sites for life, we suggest that the factor N_e ought to be more broadly reinterpreted as the number of sites per planetary system that are habitable. As the number of worlds with subsurface oceans is much higher than those within the habitable zone—this point is elucidated in more detail in Chapter 7—the value of N_e is potentially raised by two to three orders of magnitude. Yet, it is equally important to recognize that f_i could be selectively suppressed on these worlds for reasons discussed in Chapter 7. In the above spirit, we can generalize the Drake equation further to allow for the presence of free-floating planets with microbial or technological life.

From the standpoint of detecting interstellar communications, the vast majority of signals might originate from freely roaming interstellar spacecraft and thereby have nothing to do with planets. In this case, the effectiveness of the Drake equation would be diminished since it operates under the assumption that ETIs broadcast signals from planets.

Absence of dynamical processes: Even at the time of its conception and shortly afterward, it was understood that the factors in the Drake equation were time independent—that is, they were either mean values averaged over

the Galactic history or corresponded to instantaneous values at a given time (e.g., when life emerged on Earth). In reality, we observe that all of the factors will vary over time as a result of feedback mechanisms at the planetary, stellar, and Galactic scales.

The limitations of uniformitarianism were encapsulated by Milan Ćirković through a useful toy model (Ćirković 2004), which we discuss below with some modifications. Suppose that f_l can be approximated as

$$f_l = 10^{-9} \quad 0 \leq t \leq t_1, \tag{8.7a}$$

$$f_l = 10^{-6} \quad t_1 < t \leq t_2, \tag{8.7b}$$

$$f_l = 10^{-1} \quad t_2 < t \leq t_p, \tag{8.7c}$$

where $t_1 = 3.5$ Gyr, $t_2 = 8.5$ Gyr, and $t_p \approx 13.8$ Gyr is the approximate age of the Universe. The choice of f_l in the range $0 \leq t \leq t_1$ is inspired by the fact that recent simulations indicate that the majority of terrestrial planets formed at $z < 3$ due to the relatively low metallicity at higher redshifts. When we couple this to the potential lack of bioessential elements and the prevalence of high-energy astrophysical radiation (discussed below) at high redshifts, one could argue that the habitability of the high-redshift universe was largely suppressed, motivating our choice of f_l in this range. Next, we turn our attention to the low value of f_l for $t_1 \leq t \leq t_2$. Many detailed numerical simulations have been carried out to assess the habitability of the Universe over time. For instance, studies have concluded that the higher rates of Gamma Ray Bursts (GRBs), which emit lethal high-energy radiation, in the past may have posed a stringent impediment to biospheres. However, it should be recognized that these models are based on the radiation doses of *current* Earth organisms. If the tolerance levels of extraterrestrial organisms were indeed similar to that of Earth (and this is a nontrivial assumption), then f_l would be very small whenever $z > 0.5$ (e.g., Piran & Jimenez 2014).

An important item to note here is that our choice of the time dependence of f_l is arbitrary: for instance, its discontinuity can be removed by adopting a more realistic (e.g., exponential) dependence. On the other hand, as we have remarked, the possibility that f_l was much lower in the past is not entirely random. For our choices, we discover that the mean value of f_l is $\langle f_l \rangle \approx 0.038$. Hence, we see that $f_l\left(t = t_p\right) / \langle f_l \rangle \approx 2.6$, implying that the mean value of f_l in Drake's equation underestimates the actual present-day

value by a factor of about 3. This qualitative feature remains relatively robust regardless of the exact values of f_l, provided that $f_l\,(t > t_2)\,/f_l\,(t < t_2) \gg 1$. In contrast, we see that $f_l\,(t < t_2)\,/\langle f_l \rangle \ll 1$, thereby demonstrating that the Drake equation considerably overestimates the total number of ETIs in the Galaxy during this epoch.

Along thematically similar lines, Balbi (2018) developed a simple quantitative model to account for the fact that ETIs must be in causal contact with us in order for their transmission to reach us now or at some point in the future. This condition, in conjunction with the temporal distributions of the appearance of ETIs and their longevity, serves to constrain the number of actually detectable ETIs. Balbi determined that the number of detectable ETIs (N_D) with the capacity for communication is

$$N_D \sim N_s f_D \frac{t_{MW}}{\langle L \rangle}, \tag{8.8}$$

where N_s is given by (8.1), $\langle L \rangle$ is the average communicative life span of ETIs until the present epoch, f_D signifies the fraction of ETIs that fulfill the criterion of causal contact, and $t_{MW} \sim 10^{10}$ yr is the age of the Milky Way. The key point here is that $f_D \approx \langle L \rangle / t_{MW}$ under certain conditions, but *not* always. Hence, in general, N_D and N_s ought not be conflated with each other.

Effects of interstellar travel: Through a number of methods, discussed in Chapter 10, ETIs may opt to undertake interstellar travel and populate other worlds with life.[4] This could be done either by traveling to remote habitable worlds and settling there or by sending probes to seed these worlds with life. The notion that probes also represent an efficient method for initiating interstellar communication was first articulated by Bracewell (1960) and elaborated further by many authors, including Freitas (1980) and Hippke (2020).

The fact that interstellar travel is capable of greatly altering the value of N_s can be easily inferred from Michael Hart's well-known analysis of the Fermi paradox, which is discussed further in Section 8.2. One may therefore introduce the "settling" factor denoted by S in the Drake equation. One

4. This process is often referred to as *colonization* in the SETI literature. We will, instead, use comparatively neutral words, such as *settling,* wherever feasible.

of the first comprehensive attempts to write down an expression for S was laid out in the paper by Walters et al. (1980), where it was estimated that $S \lesssim 10$. S can also be estimated by envisioning interstellar travel as isotropic expansion (analogous to cosmological expansion), as noted by G. D. Brin in his detailed review (1983) of Fermi's paradox; a simpler version is presented here:

$$S = 1 + f_e \rho_\star \cdot N_e \cdot 4\pi \int_0^{R_{\max}} r^2 \exp\left[-\frac{(R_{\max} - r)}{vL}\right] dr, \tag{8.9}$$

where f_e is the fraction of species that desire to undertake interstellar travel for the purpose of settling, $\rho_\star$ is the stellar density, $R_{\max}$ is the maximum radius over which the expansion occurs, and v is the effective expansion velocity. If $f_e \to 0$ or $v \to 0$, we see that $S = 1$ along expected lines. Although the above discussion pertains to interstellar travel, it also applies to the cases where life is transferred from one planet to another either naturally via rocky material (lithopanspermia) or artificially through automated probes (directed panspermia).

8.1.4 Statistical variants of the Drake equation

Several forms of the Drake equation have been proposed wherein each of its parameters is modeled as randomly distributed variables instead of being held at fixed values. A summary of the relevant literature can be found in the monograph *The Drake Equation,* edited by Vakoch and Dowd (2015), but we will closely mirror the treatment presented in Maccone (2010).

First, we observe that most of the factors on the right-hand side (8.1) are dimensionless, except for $R_\star$ and L. An alternative form of the Drake equation is obtained by replacing $R_\star \cdot L$ with $N_{MW} \cdot L/t_{MW}$, with N_{MW} denoting the current number of stars (t_{MW} was previously delineated); in other words, L/t_{MW} represents the average fractional lifetime of the Milky Way over which an ETI would be detectable. This replacement supposes that the star-formation rate is constant, which, as we have noted before, is not wholly accurate. However, if we work with this assumption, it can be verified that N_s now comprises seven dimensionless factors.

In reality, we need not even stop here. As noted previously, there are plenty of other factors (e.g., f_g and S, introduced earlier) that could be included in the Drake equation to make it more accurate. In addition,

each of the basic seven factors can be further expanded to break it down into its constituent mechanisms. For example, Scharf and Cronin (2016) presented a Drake-like equation for f_l that incorporated several microscale factors including number of potential building blocks, availability of these molecules within a given time, and so forth. Thus, from a broader perspective, we suggest that

$$N_s = \prod_{i=1}^{n} \xi_i, \tag{8.10}$$

where all of the ξ_i's are dimensionless variables and the exact value of n is unknown, apart from $n \gg 1$. From (8.10), upon taking the natural logarithm, it becomes apparent that

$$\ln N_s = \sum_{i=1}^{n} \chi_i, \tag{8.11}$$

where $\chi_i \equiv \ln \xi_i$. At this stage, we can invoke the Lindeberg Central Limit Theorem (CLT) from mathematics; for a detailed derivation and an enjoyable historical account of the CLT and Lindeberg's seminal contribution, the reader may consult *A History of the Central Limit Theorem* (2010) by Hans Fischer. Loosely speaking, the CLT states that the sum of independent, *arbitrarily* distributed, random variables approaches a Gaussian distribution when the number of variables is very large (which is true in our case). The astute reader will observe that not all of the ξ_i's are independent of each other in reality, but we will work with this mathematical idealization henceforth. Since $\ln N_s$ tends to a Gaussian distribution, it follows that N_s approaches the corresponding log-normal distribution,

$$f(N_s) = \frac{1}{N_s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln N_s - \mu)^2}{2\sigma^2}\right], \tag{8.12}$$

with μ and σ defined as follows:

$$\mu = \sum_{i=1}^{n} \langle \chi_i \rangle, \quad \sigma^2 = \sum_{i=1}^{n} \sigma_{\chi_i}^2 \tag{8.13}$$

The mean $\langle N_s \rangle$ and the standard deviation σ_{N_s} for N_s can be computed accordingly from (8.12), and consequently we obtain

$$\langle N_s \rangle = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{8.14}$$

and

$$\sigma_{N_s} = \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[\exp\left(\sigma^2\right) - 1\right]^{1/2}. \tag{8.15}$$

We have now laid out the groundwork for our central thesis: let us define $\sigma_0^2 = \min\{\sigma_{\chi_1}^2, \sigma_{\chi_2}^2, \ldots, \sigma_{\chi_n}^2\}$. From (8.13), it is clear that $\sigma^2 > n\sigma_0^2$; it must also be noted that $\exp\left(n\sigma_0^2\right) \to \infty$ in the formal limit $n \to \infty$, regardless of the actual value of σ_0. Thus, from (8.14) and (8.15) and these relations, the coefficient of variation, defined as $\sigma_{N_s}/\langle N_s \rangle$, obeys

$$\frac{\sigma_{N_s}}{\langle N_s \rangle} > \exp\left(\frac{n\sigma_0^2}{2}\right), \tag{8.16}$$

for asymptotically large n. In this limit, the distribution (8.12) would be very spread out by virtue of the exponential dependence on n. Even for the case of finite n, it can be shown that the coefficient of variation is rendered greater than unity whenever $n\sigma_0^2 > 1$. Hence, this result shows that the mean value of the Drake equation rapidly loses its significance if the number of factors that enter the expression is high, since the associated distribution becomes increasingly broad.

Yet another implication that emerges from the preceding analysis deserves mention. The probability $\mathcal{P}_1$ of $N_s \leq 1$ is determined by calculating the cumulative distribution function (CDF), thereby yielding

$$\mathcal{P}_1 = \frac{1}{2}\operatorname{erfc}\left(\frac{\mu}{\sqrt{2}\sigma}\right). \tag{8.17}$$

Now, let us tackle the limiting case wherein $\mu/\sigma \ll 1$. In principle, this scenario is feasible when at least one of the variables in the Drake equation exhibits an exceptionally high variance. Theoretical estimates for the likelihood of abiogenesis span numerous orders of magnitude, which has prompted some scientists to assign an uncommonly large standard deviation for the corresponding probability distribution function. In this regime, we find that (8.17) simplifies to

$$\mathcal{P}_1 \approx \frac{1}{2}\left(1 - \sqrt{\frac{2}{\pi}}\frac{\mu}{\sigma}\right), \tag{8.18}$$

implying that $\mathcal{P}_1$ approaches $1/2$. In other words, the likelihood of humans being the sole technological entities in our Galaxy is seemingly around 50 percent. Thus, there is ostensibly a reasonably high chance that no ETIs exist in the Milky Way prima facie. However, we emphasize that this result is enforced by a very low mean and/or high variance. As a counterexample, consider the limit $\mu/\sigma \gg 1$, which is tantamount to positing that the variables in the Drake equation are tightly constrained with nonnegligible mean values. For this choice, we find $\mathcal{P}_1 \propto (\sigma/\mu)\exp\left(-\mu^2/2\sigma^2\right)$, consequently implying that $\mathcal{P}_1 \to 0$. Hence, the likelihood of humans *not* being the only technological entities in our Galaxy is close to 100 percent in this regime.

Our discussion indicates that the probable number of ETIs in the Milky Way is highly sensitive to the magnitudes of μ and σ. As we lack sufficient knowledge of either quantity, it is premature at this juncture to either favor or dismiss a priori the hypothesis that humans are not the sole technological intelligence in our Galaxy or the Universe. The only viable means of falsifying this hypothesis necessitates sustained empirical investigations supplemented by theoretical and numerical analyses.

8.1.5 Drake equation constraints on the probability of life

It is also possible to use the Drake equation approach to study the constraints on the prevalence of technological (or microbial) species in the observable universe. This strategy was outlined by Adam Frank and Woodruff T. Sullivan, and we shall adopt their analysis, with some alterations (Frank & Sullivan 2016).

Let us denote the total number of stars in the observable Universe by N_U. There is a quick way of estimating this by noting that there are $\sim 10^{11}$ stars in the Milky Way and $\sim 10^{11}$ galaxies in the observable Universe. It turns out that this value ($\sim 10^{22}$) is very close to the choice of $N_U \sim 2 \times 10^{22}$ in Frank and Sullivan (2016). Next, we reintroduce the factors $f_p \sim 1$ and $N_e \sim 0.1$ (for planets in the HZ) from Section 8.1.2 along with their associated definitions. Hence, we define A, the number of ETIs in the observable Universe, as follows,

$$A = N_U \cdot f_p \cdot N_e \cdot f_l \cdot f_i, \tag{8.19}$$

and recall that f_i is actually the fraction of life-bearing planets that develop species with intelligence *and* technology. As both f_l and f_i are unknown, we can define $f = f_l f_i$ as a cumulative biological factor that takes into account abiogenesis and biological and cultural evolution. Given that humans exist in this Universe, one may be tempted to set $A = 1$. In reality, even if A was somewhat smaller (say, $A = 0.1$), we cannot consider our existence in this particular Universe a genuine fluke. On the other hand, if A were infinitesimally small (e.g., $A = 10^{-40}$), this would imply that our existence could be categorized as anomalous unless one appeals to either intelligent design or the multiverse. Identifying the boundary between the two cases is not easy, but let us choose $A > 10^{-2}$ to roughly ensure that humanity is not particularly singular. From the above data along with (8.19), we find

$$f = \frac{A}{N_U f_p N_e} > 5 \times 10^{-24}, \tag{8.20}$$

which serves as a heuristic lower bound on the probability of the emergence of an ETI (on a planet in the HZ) in the Universe. Since we have mostly worked with $f_i \lesssim 1$, using this inequality in (8.20) also tells us that the minimum value of f_l must be 5×10^{-24} if we restrict ourselves only to planets within the HZ; note that the inclusion of potentially habitable exomoons in the HZ will not alter the results by more than a factor of a few.

However, there is no need to confine ourselves to habitats within the HZ, especially when it comes to microbial life. When we consider the possibility of life emerging on planets and moons with subsurface oceans, two different classes open up, as discussed in Chapter 7. The number of worlds that have the *potential* for hosting subsurface oceans (n_{SO}) is ~ 50 per star (Lingam & Loeb 2019g). Hence, the total number of possible sites for abiogenesis is given by $N_U \left(f_p N_e + n_{SO}\right) \approx N_U n_{SO}$. Thus, we must replace (8.19) with the equation

$$A = N_U \cdot n_{SO} \cdot f_l, \tag{8.21}$$

where we have not included the factor f_i since we are merely tackling the possibility that the Earth is the only world with a biosphere and not necessarily a technosphere. Using the same value of $A > 10^{-2}$, we arrive at

$$f_l = \frac{A}{N_U n_{SO}} > 10^{-26}. \tag{8.22}$$

Upon contemplating either (8.20) or (8.22), one is readily struck by how small these numbers are. For instance, the former is approximately the probability of rolling a six on a die thirty times in succession. If one adopts the stance that humans, and the multitude of species inhabiting the Earth, are not special, a new problem presents itself. If ETIs are not rare, then why have we not encountered any evidence of their existence? This question represents the heart of Fermi's paradox, which serves as the subject of our subsequent discussion.

8.2 THE GREAT SILENCE: WHERE IS EVERYBODY?

8.2.1 The gist of Fermi's paradox

In the summer of 1950, the physicist Enrico Fermi—while having lunch with Edward Teller, Herbert York, and Emil Konopinski—raised the question, "Where is everybody?" in connection with the apparent absence of extraterrestrials, despite the expectation that interstellar travel should be quite feasible for sufficiently advanced ETIs. This seemingly innocuous question has attracted a great deal of attention since its inception, and a diverse array of hypotheses have been propounded to explain the absence of such extraterrestrials.

The erroneously named Fermi's paradox is not a paradox in the strict logical sense (R. H. Gray 2015). There are a number of reasons why technological ETIs could exist and yet not be "here"; alternatively, they might already be present in our neighborhood, but we have not detected them yet for one reason or another. Furthermore, it remains unclear what Fermi was referring to when he asked his famous question: some have interpreted it as an argument against the existence of extraterrestrial intelligence altogether, while others see it as skepticism against the possibility of interstellar travel (and not ETIs themselves); based on eyewitness accounts, the latter seems more plausible. Lastly, we note that Fermi's paradox, despite its name, was actually articulated by one of the pioneers of modern spaceflight, Konstantin Tsiolkovsky, in 1933. Tsiolkovsky held the belief that spacefaring extraterrestrials did exist and explained their absence by conjecturing that these ETIs did not deem humans ready for first contact.

The first paper to truly articulate the potential correlation between the apparent absence of extraterrestrials and their nonexistence was published by Michael Hart in 1975, and a similar argument was propounded by David Viewing in the same year (Hart 1975; Viewing 1975). Hence, it has been suggested that Fermi's paradox should be called the Tsiolkovsky-Fermi-Viewing-Hart paradox. Hart's argument was further extended by Frank Tipler in 1980, who argued that the existence of self-replicating probes (SRPs), on the basis of von Neumann's universal constructor, sharpened the mystery of Fermi's paradox because these probes could populate the Galaxy in a very short time interval compared to its age (Tipler 1980). Hence, as pointed out by several authors, the premise that the absence of technologically advanced extraterrestrials can be viewed as robust evidence for their absence deserves to be called the Hart-Tipler argument. In the rest of our discussion, we will employ the terminology *Fermi's paradox* to preserve historical continuity with prior publications.[5]

In order to estimate the timescale for spreading throughout the Galaxy, let us suppose that the ships travel at a speed of $v \sim 10^{-4}c$, as this velocity is currently attainable by humans using chemical rockets. The maximum distance between stars in our Galactic disk is $D \sim 30$ kpc. The corresponding timescale t_E for traversing this distance is

$$t_E \sim \frac{D}{v} \sim 10^9 \text{ yr}. \tag{8.23}$$

It goes without saying that this scaling does not take into account the colony ship launch rate, technological advancement, and gravitational slingshot dynamics, to name a few factors. When these are taken into consideration in a more sophisticated treatment, the upper bound on the timescale for settling the Galaxy is $t_E \sim 10^8$–10^9 yr, which is lower than the age of the Galaxy by one to two orders of magnitude. Hence, if only the timescales are considered, it seems plausible that extraterrestrials, or their SRPs, have the capacity to encompass the Galaxy. In fact, provided that the SRPs are capable of successful long-term operability at relativistic speeds, Fermi's paradox becomes glaringly obvious. A detailed analysis of SRPs on

5. Although Fermi's paradox has typically been studied in the context of interstellar travel and the absence of extraterrestrials, it should be understood that the seeming absence of evidence of extraterrestrial technological signatures (e.g., radio signals) also belongs to a similar category.

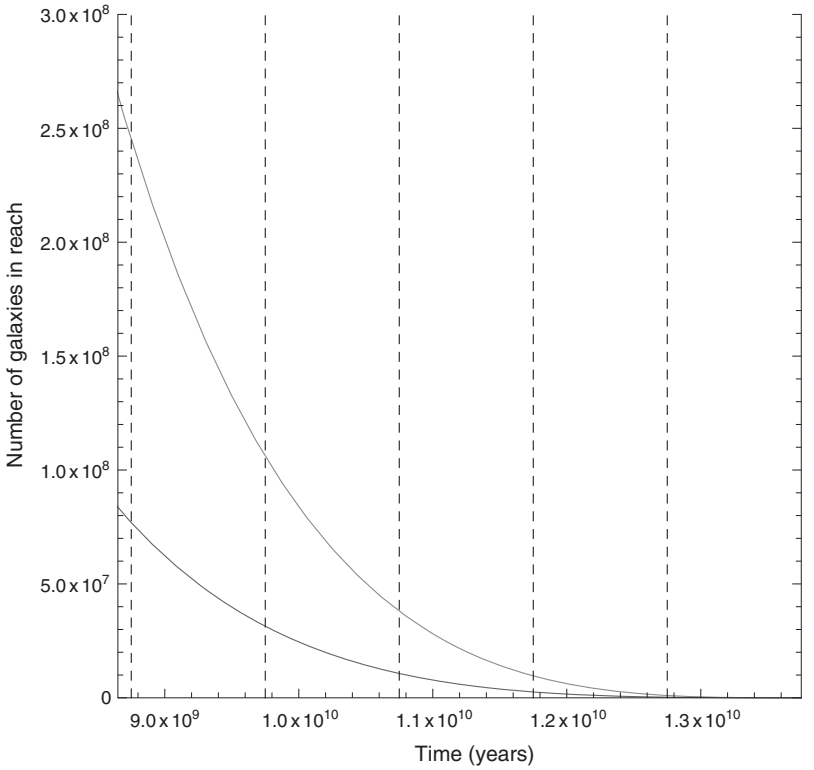


Figure 8.1 The number of galaxies inhabited by putative technological species whose self-replicating probes can reach the Earth as a function of the time at which they commence their journey; here, the clock is initiated when the Big Bang occurred. The curves, viewed from bottom to top, were obtained for SRP velocities of 0.5c, 0.8c, 0.99c, and 1.00c, respectively. The vertical dotted lines are exact intervals of 1 Gyr in the past with respect to the present day. (© 2013 IAA. Published by Elsevier Ltd. *Source:* Stuart Armstrong and Anders Sandberg [2013], Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox, *Acta Astronautica* 89: 1–13, fig. 5.)

a cosmological scale was undertaken by Armstrong and Sandberg (2013), who showed that SRPs (moving at relativistic speeds) from $\sim 10^5$ to 10^8 galaxies should have reached the Earth within the past few Gyr, as seen in Figure 8.1.

We can also estimate the number of self-replicating probes N_{SRP} that would reach each star,

$$N_{SRP} \sim S \cdot N_s \cdot f_e \cdot N_p \cdot N_w \cdot f_w, \tag{8.24}$$

where S and f_e are the expansion factor and the fraction of expansionist species introduced in Section 8.1.3, N_p is the number of SRPs per star sent out during a single expansion wave, N_w is the number of waves, and f_w is the fraction of the Galaxy covered during a single wave. As all these factors are unknown, we must enter the realm of guesswork. We choose $N_s \sim 10^4$ in order to preserve consistency with (8.3) and adopt $S \sim 10$ on the basis of Section 8.1.3. We will err on the conservative side and assume $N_p \cdot N_w \sim 10$, $f_e \sim 0.1$, and $f_w \sim 0.01$. Our choices lead to $N_{SRP} \sim 10^3$; that is, every star ought to have approximately 1000 SRPs in the stellar neighborhood. As this number is clearly quite high, the problem posed by Fermi's paradox still remains unanswered.

The number of solutions that have been proposed as resolutions of Fermi's paradox total at least over a hundred. As we cannot hope to address all of them, we will focus on a few select examples. For comprehensive analyses of this subject, we refer the reader to the monographs by Davies (2010), S. Webb (2015), Ćirković (2018a), and Forgan (2019) as well as the classic review by Brin (1983). The solutions that we will discuss henceforth can be broadly classified into three different categories along the lines of S. Webb (2015):

- Technologically advanced extraterrestrials have, or had, visited the Solar system.
- ETIs exist, but there has been no contact with them so far for various reasons.
- We are the only technological intelligence in the Galaxy, and perhaps the Universe.

8.2.2 ETIs are/were here

The class of theories that suggests that technologically advanced ETIs are (or were) here may seem preposterous. The reader will probably be uncomfortably reminded of unidentified flying objects (UFOs) and ancient astronauts.[6] We will begin by describing how we cannot definitively rule out the existence of ETIs in our Solar system and even the Earth, discuss some

6. The latter refers to the idea that ETIs played a key role in the development of human civilizations and assisted in the construction of Egypt's pyramids, Peru's Nazca Lines, and other phenomena.

of the places where ET probes might be potentially located (Foster 1972; Wertz 1976), and conclude with a discussion of some other hypotheses that fall under the category of "they are / were here."

8.2.2.1 Could there exist extraterrestrial artifacts in the Solar system?

Many proponents of the Hart-Tipler argument take it for granted that ETIs (or their probes and artifacts) are not present in the Solar system, and they use this "fact" to argue that such ETIs do not exist. We will now assess the veracity of the claim that extraterrestrial artifacts (ETAs) do not exist in the Solar system by adopting the methodology and notation outlined in Haqq-Misra and Kopparapu (2012).

Let us suppose that H corresponds to the hypothesis that ETAs are present in a given search volume V, while $\bar{H}$ represents its complement, i.e., ETAs are *not* present in the volume V. However, in referring to the search volume V, it should be recognized that the search will not be exhaustive; it will be limited by the spatial resolution R. Thus, we denote by $\bar{V}_R$ the sampled subvolume of V (at spatial resolution R) that discovers no ETAs. The conditional probability $P\left(\bar{H}|\,\bar{V}_R\right)$ signifies the probability that no ETAs exist in the volume V given that the search over $\bar{V}_R$ has yielded null results. Upon using Bayes' theorem, we end up with the following:

$$P\left(\bar{H}|\,\bar{V}_R\right) = 1 - P\left(H|\,\bar{V}_R\right) \tag{8.25}$$

$$= 1 - P\left(\bar{V}_R|H\right)\frac{P(H)}{P\left(\bar{V}_R\right)} \tag{8.26}$$

$$= 1 - P\left(\bar{V}_R|H\right)\frac{P(H)}{P\left(\bar{V}_R|\bar{H}\right)P\left(\bar{H}\right) + P\left(\bar{V}_R|H\right)P\left(H\right)} \tag{8.27}$$

We introduce the prior odds ratio $\Theta(H) \equiv P\left(\bar{H}\right)/P\left(H\right)$ that measures the likelihood of $\bar{H}$ relative to that of H. If technological intelligence is common in our Galaxy (and therefore in our volume V), then $\Theta(H)$ will be low while the converse is explicated a similar fashion. Next, we observe that $P\left(\bar{V}_R|\bar{H}\right) = 1$ since a null search result (regardless of the resolution R) will automatically hold true, provided that ETAs do not exist in the search volume. Lastly, we make use of the following ansatz for $P\left(\bar{V}_R|H\right)$:

$$P\left(\bar{V}_R|H\right) = 1 - \frac{\bar{V}_R}{V} \qquad \text{for } R \leq d \tag{8.28}$$
$$P\left(\bar{V}_R|H\right) = 1 \qquad \text{for } R > d$$

Here, d is the characteristic length scale of the ETAs. If $R > d$, the ETAs are too small to be found by the search, irrespective of the volume searched. In this event, the probability of not finding ETAs in $\bar{V}_R$ will be unity. On the other hand, if $R \leq d$, the probability of finding ETAs is assumed to be proportional to $\bar{V}_R/V$; consequently, the probability of *not* finding them equals $1 - \bar{V}_R/V$. Upon combining (8.27) and (8.28), we end up with

$$P\left(\bar{H}|\bar{V}_R\right) = \frac{\Theta(H)}{\Theta(H) + 1 - \bar{V}_R/V} \qquad \text{for } R \leq d \tag{8.29}$$
$$P\left(\bar{H}|\bar{V}_R\right) = \frac{\Theta(H)}{\Theta(H) + 1} \qquad \text{for } R > d.$$

From (8.29), a few general inferences can be drawn:

- If we consider the limit $\Theta \to 0$, we find that $P\left(\bar{H}|\bar{V}_R\right) \to 0$. The regime $\Theta \to 0$ corresponds to the case where ETAs have a very high likelihood of being present in our Solar system. In other words, this would mean that the probability of definitively ruling out their existence in the search volume (regardless of its actual size) becomes low. To put it differently, if probes sent out by technological ETIs were common, it would lower our chances of unambiguously ruling out their existence because we would have a high likelihood of encountering them.
- In the limit $\Theta \to \infty$, we find that $P\left(\bar{H}|\bar{V}_R\right) \to 1$ for arbitrary values of $\bar{V}_R/V$. The case with $\Theta \to \infty$ implies that ETAs are highly uncommon, and therefore it should not be surprising that it will be easier to conclusively rule out the existence of such probes in our Solar system; the latter translates to the condition $P\left(\bar{H}|\bar{V}_R\right) \to 1$.
- For the case with $R \leq d$, we find that $P\left(\bar{H}|\bar{V}_R\right)$ increases with the search ratio $\bar{V}_R/V$. This is along expected lines since a null search over a larger volume makes it more plausible that ETAs are indeed absent from our Solar system.

As per this formalism, we ought not definitively rule out the existence of ETAs even on the Earth or in its vicinity. The surface of the Earth has almost completely been mapped out at a spatial resolution of less than 1 m. On the other hand, the same cannot be said of the subsurface. To offer an example, Thomas Gold published an influential paper in which he suggested that Earth's crust was home to microbial life in a "deep, hot biosphere" (Gold 1992). Despite considerable advances in our understanding of subterranean ecosystems, much also remains unknown about them. Let us suppose for the sake of argument that $\bar{V}_R/V \sim 0.8$ for the Earth—in reality, this number is probably lower—and we choose $\Theta \sim 1$ since we have no way to predict its magnitude a priori. Upon using (8.29) and substituting these choices, we end up with $P\left(\bar{H}|\bar{V}_R\right) \sim 0.83$. While this probability is undoubtedly quite high, we cannot be truly certain that no ETAs exist on Earth.

When dealing with other planets and moons—except for Earth's Moon, which has been mapped to almost the same degree of spatial resolution—the value of $P\left(\bar{H}|\bar{V}_R\right)$ becomes even lower, thereby reducing our certainty of the absence of ETAs further. For instance, if we consider the Solar system as a whole, we have $\bar{V}_R/V \ll 1$, thus implying that $P\left(\bar{H}|\bar{V}_R\right) \sim 0.5$ for $\Theta \sim 1$. In Figure 8.2, $P\left(\bar{H}|\bar{V}_R\right)$ has been depicted as a function of Θ and $\bar{V}_R/V$.

We can also tackle the likelihood of ETAs from an alternative standpoint by posing the question: How many ETAs are likely to have been captured by the Earth over its life history? In order to arrive at a potential answer, we will draw on the methodology delineated in Arkhipov (1996, 1998). The number density of ETAs, denoted by ρ_A, is estimated as

$$\rho_A \sim \rho_\star \cdot N_s \cdot f_e \cdot N_A, \tag{8.30}$$

where $\rho_\star \sim 1\,\mathrm{pc}^{-3}$ is the global stellar density (the value in the Solar neighborhood is an order of magnitude smaller); N_s represents the number of detectable ETIs, which we assume is roughly equal to the number of species with technology capable of launching artifacts/probes; f_e is the fraction of ETIs that seek to expand outwards; and N_A is the quantity of artifacts dispersed per such species. The above formula supposes that the artifacts had sufficiently high velocities to be uniformly dispersed throughout the Galaxy. The capture rate of ETAs by the Earth, $\dot{N}_{\mathrm{ETA}}$ in our notation, is given by

$$\dot{N}_{\mathrm{ETA}} \sim \rho_A \sigma_A v_A, \tag{8.31}$$

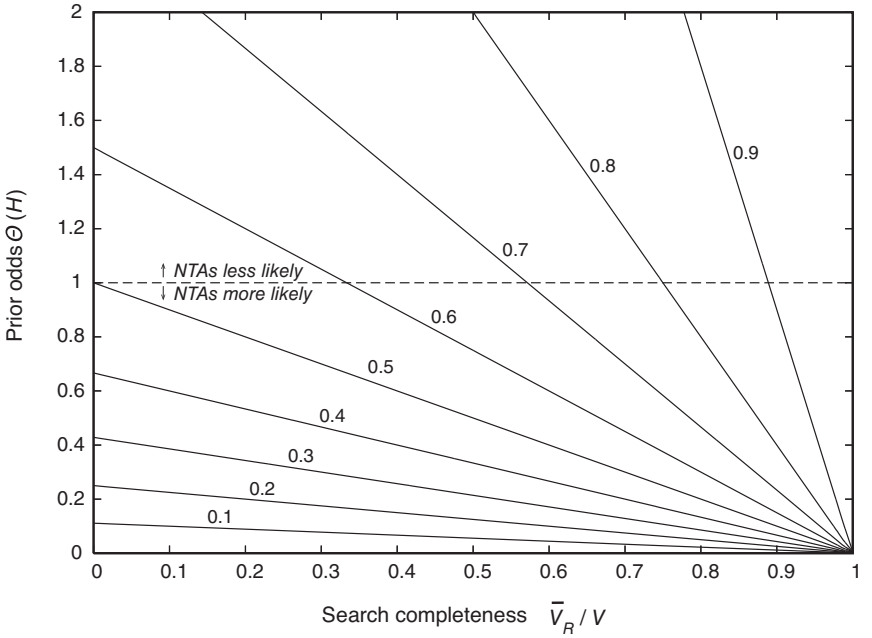


Figure 8.2 The probability that no ETAs exist in the volume V given null search results in the volume $\bar{V}_R$ as a function of the prior odds ratio Θ and the fraction $\bar{V}_R/V$. The dotted line denotes the equal prior odds ratio (i.e., $\Theta = 1$). *NTAs* stands for *non-terrestrial artifacts* and is near-synonymous with *ETAs*. (© 2011 Elsevier Ltd. *Source:* Jacob Haqq-Misra and Ravi Kumar Kopparapu [2012], On the likelihood of non-terrestrial artifacts in the Solar system, *Acta Astronautica* 72: 15–20, fig. 1.)

where v_A is the velocity of the interstellar artifacts relative to the Sun and $\sigma_A \sim \pi R_\oplus^2$ is the capture cross section; we have ignored factors such as gravitational focusing and N-body interactions in calculating σ_A since they ought to yield corrections of order unity. Upon combining (8.30) and (8.31), along with the choice $v_A \sim 2 \times 10^4\,\mathrm{m\,s^{-1}}$, we end up with

$$\dot{N}_{\mathrm{ETA}} \sim 2.4 \times 10^{-12}\,\mathrm{Gyr}^{-1} \left(\frac{N_s}{10^4}\right)\left(\frac{f_e}{0.1}\right) N_A, \tag{8.32}$$

where we have normalized N_s and f_e by the values proposed earlier. Hence, unless each expansionist ETI produces an exceptionally high number of artifacts, it would be very unlikely for the Earth to accrete ETAs; of course, this excludes the possibility in which probes deliberately land on the Earth.

On the other hand, N_A might also be unusually large. For instance, several authors have noted that the asteroid belt represents a promising source of raw materials for manufacturing activities. If a suitably advanced ETI could convert even $\sim 10^{-6}$ of the mass of the Solar system's asteroid belt (which has a mass of $\sim 3 \times 10^{24}$ g) into artifacts of $\sim 10^3$ g, this would amount to $N_A \sim 3 \times 10^{15}$. Upon substituting this value into (8.32), we find that the total number of ETAs accumulated by the Earth is $\sim 3.2 \times 10^4$ for this specific case.

8.2.2.2 Locations of extraterrestrial artifacts in the Solar system

Among the list of potential locations of ETAs, the Earth is a good place to commence our search. Unfortunately, as the Earth is geologically active, most technological signatures are degraded over long timescales (J. T. Wright 2018b). Most notably, Earth's plate tectonics will lead to subduction of the crust on timescales of $\sim 10^8$ to 10^9 yr, implying that ETAs of this age (if present) have a low probability of being discovered. Here, our discussion pertains only to ETAs and not the existence of prior ETIs on our own planet. It is expected that the latter may give rise to more robust signatures than the former, rather akin to how humans have ostensibly left an indelible imprint in the geological record (the Anthropocene). If Earth was indeed visited by ETIs, some of the following signatures analyzed in a different context could serve as the smoking gun evidence (Schmidt & Frank 2019):

- If ETIs carried out nuclear fission, some of the resultant radioactive waste products may be sufficiently long-lived to be detectable. Alternatively, if they visited the Earth more recently, sizable concentrations of short-lived transuranic radionuclides (e.g., plutonium-244) produced during fission, which would otherwise be present only in negligible trace amounts, are potential indicators of ETI activity.
- Carrying out large-scale engineering projects, which require extensive mining or quarrying, is predicted to selectively deplete raw materials and leave long-term "scars". However, the evidence may end up being buried under new layers.
- Other industrial markers such as plastics, fertilizers, heavy metals, and miscellaneous pollutants (e.g., steroids) may end up getting deposited in ocean sediments and surviving over geological timescales.

- Under rare circumstances, ETIs could undergo fossilization and preservation for $\sim 10^8$ yr, but, in most cases, it would require the availability of special substances such as amber. A more exotic, far-fetched possibility is that ETIs undertook genetic engineering of terrestrial genomes to leave behind an artificial message for any subsequent technological species (e.g., humans). Finding concrete evidence for this scenario would be consistent with the "ETIs were here" hypothesis.

In contrast, as noted earlier, discovering the presence of putative ETAs on other worlds is rendered much more difficult, since most of the above distinctive signatures are unlikely to come into play, unless the technological artifacts in question were autonomous and capable of carrying out widespread geo-engineering.

Finding evidence of past extraterrestrial visitors and ETAs on the Moon should be an easier proposition than on the Earth, in theory (Arkhipov 1998; Davies & Wagner 2013), because the former is much less geologically active. NASA's Lunar Reconnaissance Orbiter is endowed with a maximum resolution of 0.5 m/pixel and has already detected signs of human activity at the Apollo landing sites along expected lines. Thus, by judiciously harnessing modern machine-learning techniques, it is feasible to thoroughly comb the lunar surface for ETAs (Lesnikowski et al. 2020), despite the purportedly slim probability of their existence. Of course, it is conceivable that ETAs may be buried underneath the lunar surface akin to the alien monolith described in Arthur Clarke's classic science-fiction novel, *2001: A Space Odyssey* (1968).

ET probes could also be located at the *stable* Earth-Moon and Earth-Sun Lagrangian points (Freitas & Valdes 1980, 1985). The five Lagrangian points are found by calculating the locations where the gravitational pull exerted by the two massive bodies is balanced by the centripetal force for orbiting them. These points are very advantageous since objects located here can maintain a fixed distance from the two massive bodies. Both NASA and ESA have placed many satellites at these locations, and any ET probes interested in surveying our planet might also be situated here. Although no evidence of such probes has been found via either direct imaging or listening for electromagnetic signals, these preliminary searches were far from being comprehensive. Co-orbital objects—namely, objects with an orbit similar to that of Earth but not orbiting our planet—represent another promising candidate in this regard (D. Steel 1995; Benford 2019).

Michael Papagiannis (1978) espoused an interesting idea based on the Hart-Tipler argument: if interstellar travel is indeed easy and the number of such spacefaring species is nonzero, the highest likelihood for finding ET probes is in the asteroid belt. The asteroid belt has the advantage of providing a plentiful supply of raw materials (especially heavy metals), solar radiation (as an energy source), and numerous sites for undertaking mining activities. Up to this moment, we have neither detected any sources of electromagnetic radiation nor come across any objects with anomalous effective temperatures, albedos, and other properties. However, the number of asteroids subjected to careful scrutiny hitherto is a minuscule fraction of the total population.

Lastly, we turn our attention to the outer Solar system. The vast number of Kuiper Belt Objects (KBOs) are obvious hiding places for ET probes, but we will briefly describe a more interesting possibility. In 1936, Albert Einstein published an important paper pointing out the fact that gravity of a massive object bends light and thereby functions effectively like an optical lens in focusing light. The advantages of the *gravitational lens,* as it was christened later, from the viewpoint of interstellar observations and communication were first realized by Eshleman (1979), who pointed out that the Sun can be used as a gravitational lens. The minimum focal distance z_0 for the SGL (solar gravitational lens) is

$$z_0 \approx \frac{R_\odot^2}{2r_g}, \tag{8.33}$$

where $r_g = 2GM_\odot/c^2$ is the Schwarzschild radius; upon substituting the appropriate values, we obtain $z_0 \approx 550$ AU. A quick way of deriving the above formula is by noting that the deflection angle α of the SGL is known to be $\alpha \approx 2r_g/R_\odot$ from the general theory of relativity, and it must also equal the angular diameter of the Sun at the focal distance—that is, $\alpha = R_\odot/z_0$. The maximum SGL gain (magnification) for a point source, denoted by μ_L, is found from

$$\mu_L = \frac{4\pi r_g}{\lambda}, \tag{8.34}$$

where λ is the wavelength at which observations (or communications) are carried out. In addition, the gain resulting from the telescope (antenna) is proportional to $1/\lambda^2$, yielding a total gain that is proportional to $1/\lambda^3$ (Maccone 2009). The theoretical total gain due to the SGL is very high as

per this calculation,[7] and the photon flux can be enhanced by a factor of $\sim 10^9$; for further details, the reader may consult Hippke (2018b).

Serious technical issues associated with the SGL are derived from pointing, signal-to-noise ratio (because of the solar corona), and focal blurring. In principle, none of them appear to be insurmountable, especially given that spacefaring ETIs would probably possess a higher level of technology. The next question that arises is the size and mass of any probes near the SGL for the purpose of interstellar communication. This question was investigated by Gillon (2014), who posited that the probe was held in position by means of a solar sail. The area of the sail can be determined by balancing gravity with radiation pressure as follows,

$$P_{\rm rad} A_s \sim \frac{GM_\odot \left(M_0 + A_s \rho_s\right)}{z_0^2}, \tag{8.35}$$

where $P_{\rm rad}$ is the solar radiation pressure at z_0, M_0 is the mass of the payload, and A_s and ρ_s are the area and surface density of the sail, respectively. For $\rho_s \sim 5 \times 10^{-4}$ kg/m^2 and $M_0 \sim 10^3$ kg, the radius and mass of the sail are $\sim 5.5 \times 10^2$ m and 5×10^2 kg, respectively. The ET probe may have a visual magnitude of ~ 30.5 due to reflected sunlight and is not expected to be detectable by current or upcoming telescopes, either via direct-imaging or occultation methods.

8.2.2.3 Alternative hypotheses for "ETIs are/were here"

Here we will discuss some other suggestions that can be loosely classified under the category of "ETIs were here."

Directed panspermia: We discuss panspermia in Chapter 10 in greater detail, including the prospect of directed panspermia. Directed panspermia was first expounded by Francis Crick and Leslie Orgel (1973). The basic idea was that advanced ETIs deliberately chose to seed planets with life via probes. Crick and Orgel put forward two potential lines of evidence for the directed panspermia hypothesis: (1) life on Earth is based on a universal

7. There are associated subtleties involved with the ray optics approach, which we shall not delve into, since the final results are based on the assumption that the target size is diffraction-limited.

genetic code, and (2) molybdenum plays a central role in biology (despite its low cosmic abundance). This hypothesis has rightfully attracted criticism because neither of the above points necessitates an artificial origin and may be explained through orthodox biochemistry. Nonetheless, directed panspermia does constitute a viable solution to Fermi's paradox *in principle,* wherein the past (or current) presence of technological ETIs is explained through our own existence.

The Zoo hypothesis: In 1973, John Ball proposed one of the first sociological explanations for Fermi's paradox (Ball 1973). The basic idea was that ETIs that attain stability and undertake space exploration are common and much more advanced than our own. ETIs in this category would therefore set aside "wilderness" areas where other species are allowed to develop naturally; the obvious analogy drawn here is with humans setting up national parks and wildlife sanctuaries on Earth to protect, preserve, and observe other species. Ball's scenario was also quite prevalent at that time in science fiction—for example, the Prime Directive in *Star Trek*. A major shortcoming of this scenario is that it cannot be readily tested, since the ETIs considered herein are so advanced as to remain virtually undetectable (Deardorff 1987). Another notable drawback is that it draws on human practices and extrapolates them to other ETIs.

However, the most problematic aspect of the Zoo hypothesis is that it assumes that *every* technologically advanced species, at *every* moment in our history, is willing to remain invisible to humans. As such, it represents a classic example of the "monocultural fallacy" outlined in J. T. Wright, Mullan, et al. (2014), which points out the inherent error in assigning spatially and temporally homogeneous motives and practices to *all* ETIs. In other words, the Zoo hypothesis effectively requires the existence of a *Galactic Club* where all of its members act in unison. Forgan (2017) carried out Monte Carlo simulations to assess the conditions under which the Zoo hypothesis could be valid and arrived at the following conclusions:

1. If ETIs are short-lived (i.e., with lifetimes $< 10^6$ yrs), the number of culturally connected groups is much greater than unity. In other words, the Galaxy would comprise several "Galactic cliques", thereby making it very unlikely that all of these cliques will uniformly abstain from communicating with humans.

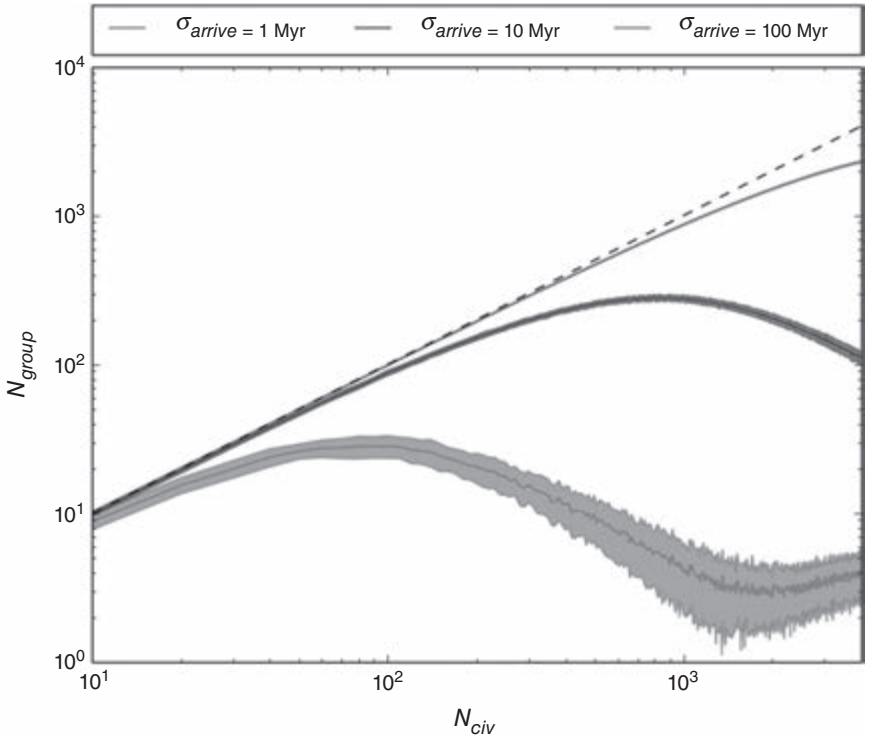


Figure 8.3 The number of Galactic cliques (N_{group}) as a function of the total number of ETIs in the Galaxy (N_{civ}). σ_{arrive} is the standard deviation for the emergence times of ETIs and serves as a measure of their temporal spacing (i.e., it signifies how closely spaced are their origination times). The lower, middle, and upper solid curves depict $\sigma_{arrive} = 1$ Myr, $\sigma_{arrive} = 10$ Myr, and $\sigma_{arrive} = 100$ Myr, respectively. (© Cambridge University Press. *Source:* Duncan H. Forgan [2017], The Galactic Club or Galactic cliques? Exploring the limits of interstellar hegemony and the Zoo hypothesis, *International Journal of Astrobiology* 16[4]: 349–354, fig. 1.)

2. The creation of the Galactic Club is rendered feasible only if all ETIs are relatively long-lived (i.e., with lifetimes $\gtrsim 10^6$ yrs) *and* arise within a fairly narrow time interval. In general, when the spacing between ETIs is increased, the chances of establishing the Galactic Club diminish. These trends can be seen in Figure 8.3.
3. In other words, even the presence of ancient, long-lived ETIs that originated much earlier than the rest of the group will not

suffice to impose hegemony and consequently form a single Galactic Club.

The Interdict hypothesis: This scenario was introduced by Martyn Fogg in 1987 and shares a close relationship with the Zoo hypothesis (Fogg 1987). The Galaxy is assumed to be hospitable for the origin of technological spacefaring species that quickly embark on rapid expansion before entering an enlightened "Steady State" era—both of these stages would occur prior to abiogenesis on Earth. In the latter era, information gathering and sharing is accorded the highest priority (over physical territory); hence these ancient ETIs would voluntarily opt to preserve life-bearing planets without settling on them and wiping out indigenous lifeforms. Even setting aside the potential difficulties and major transitions necessary for the Steady State eon to arise, we can readily see many of the objections raised in connection with the Zoo hypothesis are applicable here. For instance, the Interdict hypothesis also operates under the monocultural fallacy, since it assumes that all advanced ETIs are uniformly willing to prioritize the pursuit of knowledge and abstain from contacting humans.

One other striking candidate in this category is the simulation hypothesis, which envisions the physical world around us as not real but an essential part of a massive simulation being run by ETIs. One might perhaps argue that the existence of God, provided that God is viewed as an extraterrestrial entity, falls under the umbrella of "ETIs are here" (S. Webb 2015).

8.2.3 ETIs exist, but we have not yet encountered them

Here we will discuss different classes of solutions claiming that technological ETIs do exist but we have not found them yet. To be specific, the detection of technosignatures may require sensitivity beyond the capability of current instruments, in which case we would possibly detect many signals once we cross the threshold value (as with the recent detection of gravitational waves by LIGO in 2015–2016). Although we have focused primarily on ETIs undertaking interstellar travel, in this category we also discuss solutions pertaining to ETIs indulging in interstellar communication.

8.2.3.1 Interstellar travel is slow

One simple solution for Fermi's paradox is that interstellar travel by ETIs *or* their probes is impractical. With regard to the former, one might argue that biological life spans are limited, and deploying chemical rockets would

require the use of generation ships (also called world ships) that are probably difficult to actualize. We shall not go into the details here, as the different methods of interstellar propulsion will be discussed in more detail in Chapter 10. It suffices to say that weakly relativistic propulsion systems (most notably light sails powered by laser arrays) are conceivably within the scope of near-future human technology, and other ETIs could likewise have the capacity to develop them. By doing so, they may undertake limited interstellar travel to neighboring stars within relatively short timescales ($\sim 10^2$ yrs at the minimum). Next, if we consider probes, they ought to have a much longer life span than humans and be well suited for large-scale manufacture and dispersal. Thus, invoking the impossibility of interstellar travel as an explanation for Fermi's paradox seems untenable prima facie.

Another candidate for explaining Fermi's paradox is that ETIs or their probes have indeed commenced their expansion but have not had time to reach us. William Newman and Carl Sagan published a detailed mathematical analysis based on diffusion dynamics to estimate the time required by the nearest ETIs to reach us (Newman & Sagan 1981). On the basis of the assumptions involved, the travel time ranged between $\sim 2 \times 10^6$ yr and 1.3×10^{10} yr. If the upper bound were correct, it would indeed provide a nice explanation for Fermi's paradox, but this limit was based on the assumption of zero population growth and crucially ignored the differential (i.e., nonuniform) rotation of the Galaxy. Some of the subsequent numerical analyses have also yielded long timescales (comparable to Gyr) for the exploration of the Galaxy by ET probes.

However, earlier studies did not incorporate certain factors: (1) the prospects of ETIs or their probes settling on other worlds is ignored, (2) the motion of the stars is neglected, and (3) SRPs (self-replicating probes) are not taken into account. In order to understand why (3) matters in particular, let us carry out a simple estimate. We note that average spacing $\langle r \rangle$ between two stars can be computed via

$$\left(\frac{4\pi \langle r \rangle^3}{3} \right) N_{MW} \sim \pi R_g^2 H_g, \tag{8.36}$$

where $N_{MW} \sim 10^{11}$ is the number of stars in the Milky Way; note the similarity between the above equation and (8.4). We find that $\langle r \rangle \sim 0.8$ pc,[8]

8. We know that the nearest star to the Sun, Proxima Centauri, is 1.3 pc away, which is consistent with this mean distance estimate.

and using $v \sim 10^4$ m/s for chemical rockets, the characteristic timescale is $\bar{\tau} \sim 2\langle r \rangle / v \sim 1.6 \times 10^5$ yr. Next, let us suppose that a probe is launched and lands on the destination planet. It proceeds to make a copy of itself (over a timescale smaller than $\bar{\tau}$), and both these probes head off to the next star. The process continues indefinitely until all the stars are explored. In this case, the total number of iterations (n_c) required would be given by $2^{n_c} \sim 10^{11}$. Instead, we will opt for a more conservative value of 1.5 stars explored per iteration to account for the fact that some of the SRPs can fail; for a sophisticated analysis of this issue, the reader should consult Borgue and Hein (2020). Upon solving $(1.5)^{n_c} \sim 10^{11}$, we find $n_c \sim 62.5$. Thus, the time required for completing the exploration of the Galaxy (t_c) is

$$t_c \sim n_c \cdot \bar{\tau} \sim 10^7 \text{ yr}. \tag{8.37}$$

Hence, as per our toy model, the exploration of the Galaxy by SRPs might take only $\sim 10^7$ yrs. A similar result was obtained by Nicholson and Forgan (2013), who concluded that the use of gravitational slingshots by SRPs could result in a Galactic exploration timescale of $\sim 10^7$ yr. The number of stars visited and the associated timescale for optimized slingshot dynamics is presented in Figure 8.4.

The potential existence of SRPs does appear to lower the value of t_c in agreement with the reasoning outlined in Tipler (1980). In turn, this would imply that the oldest ETIs must have arisen within the past $\sim 10^7$ yr in order for its probes to not have reached us yet; interestingly, this timescale is quite close to the origin of the *Australopithecus* genus (which eventually gave rise to genus *Homo*) a few Myr ago. Given that ~ 10 Myr is very short compared to astronomical timescales such as the ages of the Earth and the Galaxy, the premise that ETIs have not yet reached Earth comes across as rather unlikely.

8.2.3.2 We live in a bubble

In Section 8.2.2.3, we saw that the Zoo hypothesis envisions our Solar system as a designated wilderness area, i.e., it suggests that humans are enclosed in a protective bubble. We can now ask the question whether such a phenomenon could occur because of other mechanisms that are not primarily sociological in nature.

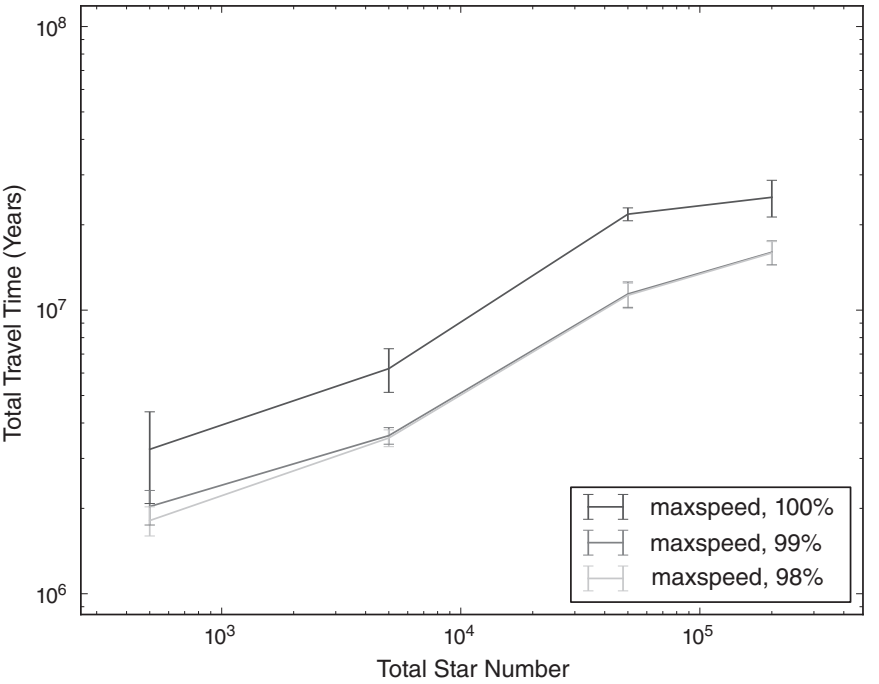


Figure 8.4 The time taken to explore 100%, 99%, and 98%—embodied by the top, middle, and bottom curves, respectively—of the stars (in a total sample of 2×10^5 stars) for the "maxspeed" case. This refers to the scenario in which SRPs select their next target star such that the slingshot maneuver from the present star yields the maximum velocity boost. The observed delay in reaching 100% is because of additional constraints in the model dictating that multiple probes cannot visit the same star. (© Cambridge University Press. *Source:* Arwen Nicholson and Duncan Forgan [2013], Slingshot dynamics for self-replicating probes and the effect on exploration timescales, *International Journal of Astrobiology* 12[4]: 337–344, fig. 5.)

An innovative solution along these lines was proposed by Landis (1998). There are a few assumptions involved in the model. First, it is supposed that there is a maximum interstellar travel distance for the purposes of settling. Second, the settled worlds will eventually diverge culturally from the parent world and become an independent technological species. This is an important point, and one that has not been sufficiently appreciated by some explanations of Fermi's paradox that operate under the premise that both the parent and settled worlds will possess a uniform culture. In reality, Earth is rife with instances where new species are born as a result of geographical

isolation: this process goes by the name of allopatric speciation, and the Galápagos finches famously studied by Charles Darwin during his voyage on the *HMS Beagle* are a classic example. It is further assumed that a world that was already settled will not be resettled by another species.

Lastly, the central rule in this model is that a given species is either expansionist or nonexpansionist with probabilities p_0 and $1 - p_0$, respectively. With this rule, and the previous assumptions, the problem of interstellar travel can be reformulated in the language of percolation theory. The simplest way to envision percolation theory is as follows. Suppose that we have a 2D lattice comprising square tiles, and each site has a probability p_0 that it is occupied and a probability $1 - p_0$ that it is not. Clearly, for very small p_0, most of the lattice will be unoccupied, and vice versa. There exists a percolation threshold p_c, above which we can start from one edge of the lattice and reach the opposite boundary continuously—that is, moving from one occupied tile to another without any gaps in between.

Landis (1998) utilized the percolation model to study how ETIs may expand outward by implementing the above rule and the other assumptions. Three distinct cases emerged:

1. If $p_0 < p_c$, the expansion is terminated at some stage, and the final landscape consists of distinct clusters, with the boundary of each cluster being comprised of nonexpansionist species.
2. If $p_0 > p_c$, these clusters grow indefinitely and fill the available space (forming a spanning cluster). However, even in this scenario, small voids exist that are surrounded on all sides by nonexpansionist species.
3. If $p_0 \approx p_c$, fractal structures are formed with arbitrarily large clusters (i.e., filled regions) but also arbitrarily large voids.

Figure 8.5 shows the case corresponding to $p_0 \approx p_c$, and the existence of large unfilled regions is readily apparent. Hence, a natural explanation for Fermi's paradox could be that we live in one of the voids and have therefore not made contact with any spacefaring ETIs. However, it is worth noting that Landis's model does not incorporate two important dynamical elements: (1) extinction (a species may die out and its home world could be settled by another species) and (2) evolution (a species may shift from an expansionist to a nonexpansionist mode or vice versa at some point in its history). When these two elements are incorporated, the validity of the

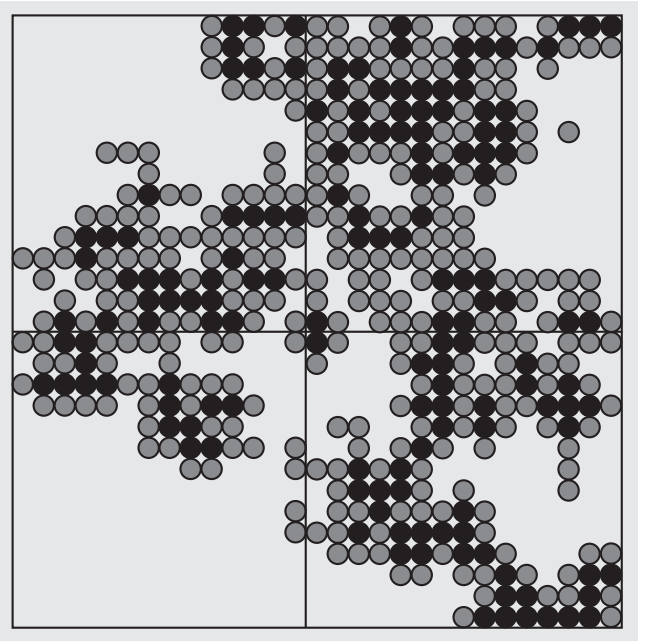


Figure 8.5 A 2D slice of the percolation simulation on a 3D simple cubic lattice. $p_0 = 1/3$; the percolation threshold is given by $p_c = 0.311$. Black circles represent expansionist species, gray circles signify the nonexpansionist species, and the unfilled regions represent voids—to wit, the sites that are not visited. (© Geoffrey A. Landis. *Source:* Geoffrey A. Landis [1998], The Fermi paradox: An approach based on percolation theory, *Journal of the British Interplanetary Society* 51[5]: 163–166, fig. 1.)

three different outcomes discerned in the original model remains indeterminate. However, Landis's basic qualitative conclusions are loosely consistent, mutatis mutandis, with recent numerical simulations based on probabilistic cellular automata for certain parametric choices (Vukotić & Ćirković 2012; Galera et al. 2019); similar analytical results are presented in Lingam (2016b).

We can also approach the question of "Does Earth exist in a void?" from the standpoint of population dynamics, as suggested by Stull (1979) and Ostriker and Turner (1986); the reader is referred to Carroll-Nellenback et al. (2019) for a comprehensive and state-of-the-art quantitative analysis along these lines. To illustrate this point, we begin by appealing to the famous metapopulation model developed by ecologist Richard Levins (1969), wherein the metapopulation was regarded as a "population of populations." Levins propounded a coarse-grained model in which each

population was treated as a single entity, and he wrote a differential equation that accounts for their migration and extinction,

$$\frac{dN}{dt} = \frac{N}{\tau_M}\left(1 - \frac{N}{N_T}\right) - \frac{N}{\tau_E}, \tag{8.38}$$

where N is the number of populations at a given time t, N_T is the total number of available patches, and τ_M and τ_E are the characteristic immigration and extinction timescales, respectively. The first term on the right-hand side quantifies the rate at which new populations are being established, and the factor $1 - N/N_T$ accounts for the fact that the total number of populations cannot exceed the carrying capacity N_T. The second term on the right-hand side encapsulates the rate at which populations are becoming extinct. If we choose to interpret N as the number of ETIs and N_T as the total number of planets that can be inhabited, the metapopulation model is useful for understanding interstellar travel by ETIs. The steady-state value N_0 is given by

$$N_0 = N_T\left(1 - \frac{\tau_M}{\tau_E}\right). \tag{8.39}$$

The key feature worth noting here is that, as long as $\tau_M > 0$ is valid (and the criterion $\tau_M/\tau_E < 1$ is satisfied), $N_0 < N_T$. In other words, the total number of ETIs will never attain its maximum theoretical value—that is, not every habitable planet will actually end up being inhabited. As a concrete example, let us suppose that the typical timescale for extinction is comparable to the total lifetime of an ETI in the detectable phase ($\tau_E \sim L \sim 10^6$ yrs) and we assume that the migration timescale is approximately equal to the travel time between neighboring stars via chemical rockets ($\tau_M \sim \bar{\tau} \sim 1.6 \times 10^5$ yr). Upon plugging these values into (8.39), we obtain $N_0/N_T \approx 0.84$, implying that 16 percent of the total number of inhabitable planets remain unsettled.

The dynamical model proposed by Gros (2005) also falls under this category. In this formulation, all technological species are divided into expansionist and nonexpansionist categories. The dynamical equations are given by

$$\dot{E} = \left(g_E - e_E - m_E\right)E + m_S S + b_E \tag{8.40a}$$

and

$$\dot{S} = m_E E - \left(e_S + m_S\right)S + b_S, \tag{8.40b}$$

where E and S denote the number of expansionist and nonexpansionist (stagnant) species, respectively. $e_{E,S}$ are the extinction rates and $b_{E,S}$ are the effective birth rates for the two groups. m_E denotes the rate at which expansionist species become nonexpansionist, and m_S is the rate for the opposite process. g_E is the growth rate for the expansionist species and distinguishes them from the nonexpansionist ones. It can be readily verified that the solutions for E and S are of the form $\exp(\Lambda t)$ with Λ determined by a quadratic equation.

If m_E is large, it turns out that $\Lambda < 0$ is possible, implying that the number of expansionist species declines over time; consequently, the Earth may not have been visited by any expansionist species. A high value of m_E corresponds to the case where ETIs are characterized by long periods of quiescence and transient periods of expansion. We must ask ourselves whether such behavior has any known analog on Earth, and the answer appears to be tentatively affirmative. Punctuated equilibrium, a prominent theory in evolutionary biology that was delineated by Niles Eldredge and Stephen Jay Gould (Eldredge & Gould 1972), proposes that biological evolution is characterized by long periods of stasis interspersed with intermittent periods of rapid speciation (cladogenesis). Another classic example in this context is the landmark work *The Structure of Scientific Revolutions* by Thomas Kuhn, which envisioned scientific progress as being essentially episodic in nature and characterized by bursts of "revolutionary science" (1962).

One can also solve for the steady-state populations of expansionist and nonexpansionist species. For the former, we find that

$$E_0 \propto m_S b_S + (e_S + m_S)\, b_E, \tag{8.41}$$

where E_0 is the steady-state value of E. From the above equation, we see that $E_0 \to 0$ when $b_{E,S} \to 0$. In other words, if the birth of new ETIs is very rare—a scenario that we shall explore subsequently in more detail—the total number of expansionist species will be greatly reduced, implying that the chances of Earth encountering one might be diminished.

Although the above discussion centered on the use of highly simplified toy models, it serves to highlight the fact that complex pattern formation is an ubiquitous feature in many areas ranging from evolutionary game theory to morphogenesis. For a detailed overview of evolutionary dynamics and the manifold spatial patterns manifested therein, we direct the reader to Nowak (2006). Collectively, these models offer strong evidence in favor

of the claim that patchy Galactic landscapes are common; as a result, the notion that Earth is located within a void is certainly feasible. However, two broad objections can be raised against this line of reasoning. First, accepting that our Solar system is located in a void still renders it anomalous to some degree and therefore may be regarded as a weak violation of the Copernican Principle (which posits that the Earth is not privileged). Second, even if ETIs have not visited the Earth, they could still participate in interstellar communication, and the absence of their detection requires a separate explanation.

8.2.3.3 Sustainability solutions

In discussing ETIs that are inherently expansion driven, most studies, either implicitly or explicitly, operate under the assumption that unlimited exponential growth is possible. Yet, the problematic issue about exponential functions, from a mathematical standpoint, is that they grow rapidly over relatively short timescales. For instance, over the past ~ 400 yr, the United States has witnessed exponential growth in energy consumption at a rate of 2.9 percent per year. In other words, the e-folding timescale is around 35 yr. If we assume that the same growth rate can be indefinitely sustained across the Earth and use the fact that the global current power consumption is $\sim 1.8 \times 10^{13}$ W, the power P_E at a time t is given by

$$P_E(t) = 1.8 \times 10^{13}\,\mathrm{W}\,\exp\left(\frac{t}{35\,\mathrm{yr}}\right). \tag{8.42}$$

If we consider the case where all the power radiated by the Sun is used up by humans, we can estimate the time at which this would come to pass. Using $P_E(t) \approx 3.8 \times 10^{26}$ W in conjunction with the above expression, we find $t \approx 1.1 \times 10^3$ yr. Even a power consumption equal to that of the Milky Way, corresponding to $P_E(t) \approx 1.5 \times 10^{37}$ W, would require a time of only $t \approx 1.9 \times 10^3$ yr.

From this simple calculation, we see that two possibilities immediately open up for ETIs around stars. Either the growth rate of energy consumption must be significantly reduced or this untrammeled growth will eventually lead to the collapse of a given species. The former can be interpreted as the "Sustainability Solution" to Fermi's paradox: either ETIs will be expanding outward very slowly—perhaps explaining why they are not

here yet (see Section 8.2.3.1)—or they underwent rapid growth and collapse (Haqq-Misra & Baum 2009). Another candidate for resolving Fermi's paradox, closely related to sustainability, is that advanced ETIs may find it unnecessary to leave their home world, or, at the most, they stay confined within a local region of space. The reasons advanced in support of this idea are many and varied, ranging from stagnation and apathy to postbiological evolution, unncessary expenditure of energy and resources on interstellar travel, and ethical or political considerations.

While this class of solutions to Fermi's paradox is certainly appealing in light of humanity's own concerns about sustainable development, a potential limitation must be mentioned here. Even if one species possesses the technological wherewithal to create interstellar probes (preferably SRPs) without expending too many resources, it should be able to undertake interstellar exploration of the entire Galaxy quite easily. As a result, we would still be left with the question of why we haven't detected any such probes in the Solar system; of course, as noted in Section 8.2.2.1, they could easily be present in our Solar system and yet avoid being discovered.

8.2.3.4 ETIs are sending signals, but we have not identified them

Hitherto, we have focused on analyzing solutions to Fermi's paradox that seek to explain why ETIs have not traveled to our Solar system. However, the absence of interstellar communications, as much as the absence of ET visitors, still poses the question of "Where is everybody?" In order to explain why we have not detected ET signals, a different class of solutions has been advanced, some of which are described below.

Exotic signals: Conventional SETI has focused almost exclusively on the search for electromagnetic signals, although the last few years have witnessed significant progress in our understanding and search for extraterrestrial artifacts. Yet, at least in principle, other forms of communication are also possible. The best studied among them is communication via neutrinos since they possess numerous advantages: (1) the chances of neutrinos being attenuated is much lower since photons regularly encounter opaque media during transit, and (2) unlike photons, where the signal-to-noise ratio is usually a major issue, neutrino signals are almost noise-free. On the other hand, the present cost of undertaking neutrino communication is enormous, and the most efficient forms of data encoding and decoding are not known. For

further details regarding neutrino communication, the reader may consult Learned et al. (2009).

More exotic possibilities include signals based on gravitational waves. For instance, Abramowicz et al. (2019) proposed that advanced technological entities might opt to construct planet-sized devices near the supermassive black hole at the center of the Milky Way to extract energy from it or conduct observational tests of physical theories. In consequence, a distinctive gravitational-wave signal that is detectable by the Laser Interferometer Space Antenna (LISA) mission within a span of a few years would be manifested.

In either of the above instances—to wit, neutrinos and gravitational waves—we would not detect any signals that ETIs broadcast by means of these methods because we are not actively carrying out such searches.

Searching in the wrong frequencies: We will deal with radio and optical SETI—that is, searching for signals in the radio and visible wavelengths—in Chapter 9. In the case of radio signals, the frequency band 1–10 GHz has the dual advantage of low atmospheric distortion and minimal background radiation. As a result of dispersion by electrons in the interstellar medium (ISM), the bandwidth of any transmissions is presumed to be $\lesssim 0.1$ Hz. Hence, in the radio band mentioned above, there are a total of $\sim 10^{10}$ channels with a bandwidth of 0.1 Hz. This clearly illustrates the difficulty of locating ET signals even if ETIs were incessantly broadcasting them. Optical SETI offers certain advantages compared to radio SETI, but systematic searches have been even fewer in number. Hence, viewed collectively, even if one supposes that ETIs were sending us signals, we may not be cognizant of the right frequencies to identify them.

Searching in the wrong locations: Yet another difficulty surrounding SETI programs is that the choice of targets matters a great deal. Detecting unintentional leakage, as we saw in Section 8.1.2, is a very challenging endeavor; the same holds true for eavesdropping on communications between two ETIs, since they would most likely be using narrow beams. The likelihood of detecting artificial beacons, either deliberately pointed toward us or isotropic emitters, is much higher. As a result, this poses the question of determining the optimal search strategy. Many, but not all, SETI programs have tended to focus on observations of selected target stars, although wide-sky surveys were also undertaken. Hence, another explanation for the putative absence of ET communication is that we have failed to look at the right places.

Insufficient time for detection: Not surprisingly, this point is one of the most common defenses offered by SETI advocates, and certainly there is much truth to this assertion. Not only is the field relatively nascent (about sixty years old), but SETI programs have often been strapped for funding and have had to inefficiently piggyback on other missions. If one subscribes to this worldview, perhaps detecting ET signals is simply a matter of allocating sufficient observational time and resources.

Insufficient detection sensitivity: We noted in Section 8.1.2 that the likelihood of detection is not just a function of the emitting species but also the receiving species. Let us suppose that the number of broadcasting species was very low and that the nearest one was located at a distance $\gtrsim 10$ kpc. If this species had technological capabilities similar to our own and opted to communicate using a transmitter with specifications comparable to the Arecibo radio telescope, the signal received at Earth would be undetectable by present-day technology. In other words, even if there are current broadcasts from ETIs, we may simply be too primitive in the technological sense to detect them.

ETIs are not interested in transmission: ETIs might prefer to be passive listeners instead of active broadcasters. Two obvious reasons favoring this line of reasoning are that (1) the cost of broadcasting (especially in the case of isotropic emission) is much higher than the cost of listening and (2) selecting suitable targets for broadcasting signals is challenging since the chances of success are contingent on identifying the appropriate wavelengths and planets and accurately transmitting signals to those locations. A closely related explanation is that technologically advanced ETIs opt to leave us alone for multiple reasons including apathy, altruism, and derision. However, the central difficulty with both of the above hypotheses is that they assume all ETIs possess universal impulses; in doing so, they operate under the monocultural fallacy.

Many of the above points are broadly classifiable under the category of searching for electromagnetic needles in the cosmic haystack. This haystack is multidimensional in nature, with the exact number of dimensions varying from study to study depending on the survey constraints taken into account. J. T. Wright et al. (2018) identified nine dimensions in total, which included the likes of transmission bandwidth and frequency (two dimensions), spatial location (three dimensions), and polarization associated

with the signal. After analyzing the major radio surveys undertaken until 2017, Wright et al. argued that the fraction of the haystack searched was roughly 6×10^{-18}. It is worth sketching an alternative line of reasoning outlined in Tarter et al. (2010), which yields qualitatively similar results.

First, the number of stars in the Milky Way is $\sim 10^{11}$. Most searches for radio signals focus on the frequency interval of 1–10 GHz. If we assume that each channel has a bandwidth of 1 Hz, there are 9×10^9 channels to be scanned. Thus, the total number of star-channel combinations is $\sim 10^{21}$. This number is further elevated by the fact that different types of modulation schemes are possible (e.g., frequency or amplitude modulation), and dispersion caused by plasma in the interstellar medium constitutes another degree of freedom. Hence, the total parameter space incorporating these additional effects was taken to be $\sim 10^4 \times 10^{21} = 10^{25}$. Of this space, surveys conducted by Project Phoenix and the Allen Telescope Array searched for narrow-band signals and were estimated by Tarter et al. (2010) to have encompassed $\sim 2 \times 10^{12}$ combinations in total. Hence, the resultant fraction obtained is $\sim 2 \times 10^{-13}$.

Next, the parameter space must account for the power of the transmitter. If we require a hypothetical transmitter situated at the center of the Milky Way to be detectable, its effective isotropic power must be $\sim 2 \times 10^4$ times higher than the Arecibo transmitter. For the sake of caution, if we suppose that the transmitter operates at current human technology, the above fraction is further reduced to $\sim 2 \times 10^{-13} / \left(2 \times 10^4\right) = 10^{-17}$. Finally, we must consider the duty cycle of the signal—that is, the fraction of time during which the signal is active and detectable. For convenience, it is specified to be ~ 0.5, with the fraction correspondingly being lowered to $\sim 5 \times 10^{-18}$. The final factor to consider is whether the transmitters are operational in the same epoch when we are carrying out these surveys. This factor is roughly given by L/t_{MW} and potentially lies between $\sim 10^{-8}$ (for $L \sim 100$ yr) and $\sim 1/2$ (for $L \sim 5$ Gyr). Therefore, the fraction of the haystack searched as per this methodology and its attendant assumptions ranges from 5×10^{-26} to 2.5×10^{-18}.

Lastly, before moving to the next class of solutions, we note that there also exists the possibility that we have detected ET signals without realizing it. Although this scenario seems outlandish, it is worth mentioning that we know of symbols from ancient civilizations on Earth whose meaning we have not been able to decipher (e.g., Indus Valley civilization). As these

symbols were produced by humans only a few millenia ago and yet remain mysterious, the paramount difficulties of identifying and deciphering ET signals become apparent.

8.2.3.5 ETIs are far beyond us

Numerical simulations of planetary formation on cosmological timescales have concluded that the average ages of rocky planets orbiting F, G, and K stars are approximately 2.5 ± 1 Gyr older than the Earth, while those around M-dwarfs are 3.5 ± 1 Gyr older (Zackrisson et al. 2016). In other words, there is a high likelihood that the extant ETIs are much older (and potentially more advanced) than our society. Hence, we are confronted by the difficulty of envisioning their current motives, technology, and actions. The hypotheses that we discuss below mostly have one thing in common: they attempt to understand the evolution of incredibly advanced ETIs and thereby explore the ramifications for Fermi's paradox.

The post-biological Universe: By drawing on current developments in the field of artificial intelligence (AI), Dick (2003) presented thought-provoking arguments about why most technological "species" in the Universe would be based on AI—that is, the Universe would be predominantly post-biological. The significance of cultural evolution, which was asserted to be primarily driven by the amplification of knowledge and intelligence, in the Drake equation was highlighted in this work. Even if we restrict ourselves to humans (who are biological), the importance of culture in shaping our evolution has been extensively studied, and there exists sufficiently compelling evidence in this regard. Dick goes on to further contend that species with $L > 10^3$ yr will mostly likely enter a post-biological stage. If this is indeed correct, the ramifications for both SETI and Fermi's paradox are expected to be profound since it would not be apparent how and where one can locate such species.

ETIs are no longer located around stars: A number of hypotheses speculate that advanced ETIs will no longer be located on planets around stars. In particular, several authors have argued that black holes represent natural "attractors for intelligence" (Vidal 2014) for a number of reasons. The Transcension hypothesis by Smart (2012) holds that complex systems are characterized by a rapid increase in spatial, temporal, energetic, and material

(collectively denoted by STEM) density as well as efficiency per computation. Eventually, in order to maintain this STEM compression, advanced ETIs will relocate to the vicinity of black holes. The many advantages offered by black holes from the perspective of intelligence include (1) gravitational time dilation (enabling long-lived species), (2) maximal information transfer rates for a given energy, (3) efficient energy extraction, and (4) gravitational lensing (facilitating interstellar / intergalactic observations and communications). Hence, if technological ETIs were to achieve superintelligence and settle around black holes, conventional SETI focusing on planets would have a low likelihood of success. This conjecture could explain why ETIs exist but have not been detected.

Reducing the cost of computation: In 1961, Rolf Landauer authored a landmark paper, drawing inspiration from the likes of Leo Szilárd, John von Neumann, and Léon Brillouin, in which he argued that the *irreversible* manipulation of one bit of information necessitates the dissipation of a minimum amount of energy (Landauer 1961). This limit, widely known as the Landauer bound, is linearly proportional to the reservoir temperature T.[9] There are several subtleties associated with deploying the Landauer bound, one of them being the fact that logically reversible computations are executable at no energetic cost. If advanced ETIs do need to undertake certain irreversible operations (e.g., erasure), it would seemingly make more sense for them to function in low-temperature environments. This premise led to suggestions that advanced post-biological ETIs would either migrate into the outer regions of the Galaxy or lie dormant until the Universe cools further (Sandberg et al. 2016); the latter notion has, however, been critiqued by Bennett et al. (2019).

A difficulty with some of these explanations is that they presuppose a certain degree of evolutionary convergence and consequently argue that all post-biological ETIs will be driven by similar considerations. Nonetheless, this category is an important and valuable one since it attempts to progress beyond the reliance on overly anthropocentric reasoning.

---

9. As an interesting addendum, we point out that protein synthesis in cells is merely one order of magnitude removed from the Landauer bound, in sharp contrast to supercomputers that function at efficiencies several orders of magnitude smaller than this limit (Kempes et al. 2017).

8.2.4 We are alone

One of the simplest answers to "Where is everybody?" is "We are alone" or "We are extremely rare." However, the reason that this possibility is vehemently excluded by many individuals is because it constitutes a clear violation of the Copernican Principle. If we are truly the only intelligent species in our Galaxy (and perhaps even the Universe), it demarcates us as being privileged observers. Nevertheless, as we shall describe below, there are numerous scenarios in which the emergence of technological intelligence can be highly uncommon. Historically, this stance was championed vigorously by several famous evolutionary biologists (e.g., G. G. Simpson 1964; Mayr 1985), who emphasized the role of pure "chance" in the evolution of humans; by extension, these authors contended that the likelihood of other ETIs in the Galaxy is minimal (Tipler 2003).

If even one of the factors in the Drake equation (8.1) is zero or infinitesimally small, the number of ETIs will also end up being very low. In 1998, the economist Robin Hanson introduced the evocative term *Great Filter* to characterize bottlenecks that might exist between prebiotic organic molecules on the one hand and long-lived expanding technological intelligences on the other.[10] If there does exist a Great Filter (or more than one), the likelihood of ETIs can be greatly reduced, since the corresponding factor in the Drake equation would become very small. With the advancements in detecting signatures of nontechnological and technological entities, it is not inconceivable that the Great Filter paradigm might be testable in the upcoming decade(s) (Haqq-Misra et al. 2020).

8.2.4.1 The anthropic argument

Nearly four decades ago, Brandon Carter presented an influential line of reasoning favoring the rarity of ETIs (Carter 1983). The gist of Carter's argument is as follows.

The stellar lifetime $\tau_\star$ sets an upper bound on the duration that biological organisms survive on their host planet. Let us denote the timescale for intelligent observers to evolve on that planet by τ_i. We can then imagine three different possibilities: (1) $\tau_i \ll \tau_\star$, (2) $\tau_i \sim \tau_\star$, and (3) $\tau_i \gg \tau_\star$. Let us begin by evaluating the likelihood of (1) being correct. On Earth, we

10. George Mason University Information Tech Services (1998), The great filter—are we almost past it?, http://mason.gmu.edu/~rhanson/greatfilter.html

have $\tau_i \sim 4.5$ Gyr and $\tau_\star \sim 10$ Gyr. Hence, if $\tau_i \ll \tau_\star$ were indeed valid, this would not explain why we observe $\tau_i \sim \tau_\star$; to put it differently, if scenario (1) were accurate, we should have typically found ourselves having evolved at a time when the age of the Earth was much smaller than 10 Gyr (e.g., ~ 100 Myr). The second case is less probable since we operate under the assumption that these two timescales are independent; hence there is no a priori reason to suppose that (2) is always true. As a result, we are left with the third explanation. At first glimpse, it may seem as though even (3) cannot be correct because $\tau_i \sim \tau_\star$ for the Earth. However, we have not taken into consideration an important effect: our very existence ensures that $\tau_i \lesssim \tau_\star$ must be valid on Earth as otherwise we (in our capacity as intelligent observers) would not have evolved before the Sun died out. To put it differently, evolution on our planet must have occurred unusually rapidly compared to the norm to enable the emergence of intelligent observers prior to the death of the host star.

This line of reasoning constitutes a classic example of the application of the anthropic principle. As per B. Carter (1974, p. 291), the anthropic principle is encapsulated as follows:

> What we can expect to observe must be restricted by the conditions necessary for our presence as observers. (Although our situation is not necessarily *central,* it is necessarily privileged to some extent.)

In the above discussion, the selection of only those planets with $\tau_i \sim \tau_\star$, despite the general tendency toward $\tau_i \gg \tau_\star$, to ensure the evolution of intelligent observers is a manifestation of the *observation-selection effect* (also referred to as the *self-selection effect*). To conclude, if Carter's argument is correct, our evolution was a fortuitous occurrence and the majority of habitable planets would remain uninhabited.

We can quantify this further in the following fashion. Let us suppose that the emergence of intelligence is but one of n "hard" evolutionary steps. We discussed the critical steps paradigm in Section 3.10, where it was argued that modern evolutionary biology favors $n < 10$, with the most likely values being $n = 5$ or $n = 6$. Assume that the characteristic timescale for the occurrence of each step was τ_0, with $\tau_0 \gg \tau_\star$ as per Carter's reasoning. In this case the probability of technological intelligence evolving on a given planet is $\sim (\tau_\star/\tau_0)^n$. Hence, the total number of planets in the Galaxy N_i on which intelligence arises is estimated to be

$$N_i \sim N_{MW} f_p N_e \left(\frac{\tau_\star}{\tau_0}\right)^n, \tag{8.43}$$

and we choose $\tau_\star/\tau_0 \sim 10^{-2}$ and $n = 5$ as fiducial values. Hence, we end up with $N_i \sim 1$, and this value can fall below unity for lower values of $\tau_\star/\tau_0$ or higher values of n. Thus, Carter's argument might serve as a powerful explanation of why we are alone in the Galaxy.

Yet, several inherent assumptions in Carter's model have garnered due criticism. To begin with, as pointed out by Livio (1999), Carter's model presumes that τ_i and $\tau_\star$ are wholly independent timescales. In contrast, as noted in Chapter 2 on the origin of life, there is a strong possibility that the flux of ultraviolet (UV) radiation played a key role in regulating abiogenesis. Moreover, especially on low-mass stars like M-dwarfs, UV radiation may play a significant role in splitting water (photolysis) and building up oxygen in the atmosphere. On Earth, the rise of oxygen during the Neoproterozoic era is perceived as one of the probable causes for the emergence of animals. Thus, taken collectively, it is conceivable that the two timescales are not wholly unconnected (see Section 4.3 for additional details); on the other hand, the notion that $\tau_\star$ is the only regulator of τ_i is patently incorrect.

A second argument against Carter's reasoning was propounded by Ćirković et al. (2009) on the basis of their formulation of "galactic punctuated equilibrium." As we have seen, Carter's anthropic argument is based on the idea that *well-defined* timescales for both τ_i and $\tau_\star$ do exist, corresponding to biological and astrophysical processes, respectively. In reality, even in the case of the latter, a diverse array of astrophysical processes could disrupt habitability, each with their own timescale. Some examples in this category include (1) catastrophic events like GRBs, supernovae, and superflares (that can trigger mass extinctions), (2) atmospheric erosion by stellar winds and UV radiation from the host star, and (3) the Sun's passage through the Milky Way (which can perturb comets and cause impacts and extinctions). Even at the planetary level, multiple timescales are associated with magnetic field reversals, Snowball Earth episodes, and Milankovitch cycles, to name a few, and these phenomena have potentially important consequences for biological evolution. Thus, the plethora of timescales and the fact that the Earth-Sun system is not closed may collectively diminish the validity of Carter's analysis.

Lastly, even if we accept the basic premises of Carter's argument, its applicability to both subsurface ocean worlds and free-floating worlds is questionable. Even if one does have a concrete biological timescale, the corresponding astrophysical timescale is less clear. It would not be the stellar lifetime since the habitability of both classes of planets and moons is mostly independent of the availability of stellar energy.

8.2.4.2 The value of N_e is very low

In Section 8.1.2, we discussed the importance of N_e, which represents the number of potentially habitable planets. In our subsequent calculations, we used $N_e \sim 0.1$, as that represented the number of rocky planets in the HZ of the host star. However, we also emphasized that the habitable zone does not equate to habitability.

Astrophysical catastrophes: One of the preliminary requirements for habitability is imposed at the Galactic level, i.e., in the inner regions of the Galaxy, the probability of astrophysical catastrophes that can sterilize planets is greater. As a result, Annis (1999a) proposed that the frequency of GRBs explained the absence of ETIs in the Galaxy in the past and that a phase transition is currently underway and will lead to a Galaxy filled with intelligent life. However, the likelihood of GRBs causing large-scale extinctions varies both spatially and temporally, and, moreover, the extent of damage caused by GRBs to the biosphere (for a given fluence) is still not conclusively understood. Similar studies have also been undertaken to assess the threats posed by supernovae, superflares, and quasar activity, but it seems unlikely that any of these phenomena could wholly account for Fermi's paradox.

Availability of liquid water: We have already seen that rocky planets in the HZs of their host stars are relatively common, since the corresponding fraction is $\gtrsim 0.1$. However, not all of the planets in the HZ are guaranteed to have liquid water. Simulations that take into account initial water inventory, stellar evolution, and atmospheric erosion indicate that a reasonably high fraction of planets in the HZ should end up with Earthlike water contents or more (Tian & Ida 2015). Thus, the proposal that liquid water may be rare on these worlds does not appear to be a valid one.

The role of Jupiter: In their famous book, *Rare Earth* (2000), Peter Ward and Donald Brownlee argue that the existence of Jupiter was necessary in order to serve as a "shield" by preventing asteroids and short-period comets from regularly impacting the Earth and causing mass extinctions. In a broader sense, Ward and Brownlee contend that our Solar system had several unique architectural properties that enabled life to flourish on our planet. Later studies do not favor simplistic variants of the "Jupiter as shield" hypothesis, although it is clear that changing Jupiter's properties (e.g., eccentricity of orbit, mass, inclination) does affect the flux of impactors on Earth. Grazier (2016) carried out detailed numerical simulations and concluded that the primary role of Jovian planets was to modulate volatile delivery to the inner planets instead of acting as a shield, as originally assumed. Lastly, we note that ~ 10 percent of stellar systems might possess Jovian planets on wide orbits with low eccentricities. Collectively, the relationship between Jupiter and the evolution of life on Earth is not straightforward, and we cannot therefore invoke the absence or presence of Jovian planets to explain Fermi's paradox.

The Moon: The Moon has been argued to be essential for the origin and evolution of life on Earth for a number of reasons. The most widely cited reason, partly owing to the Rare Earth hypothesis, is that it helps stabilize the Earth's obliquity (axial tilt). However, recent simulations indicate that even in the absence of the Moon, the Earth could have maintained a fairly stable obliquity over $\sim 10^8$ yr timescales (Lissauer et al. 2012; Li & Batygin 2014). The large size of the Moon has also been hypothesized to be important in minimizing climactic fluctuations on Earth because of promoting slow Milankovitch cycles and rotation rate.

Next, we note that the tidal force per unit mass (denoted by a_t) exerted by an object of mass M_t and distance d_t from the Earth is given by

$$a_t \propto \frac{M_t}{d_t^3}. \tag{8.44}$$

Now, suppose that we calculate a_t for the Earth-Moon system (a_m) and for the Earth-Sun system (a_s). It can be shown that $a_s \approx 0.45 a_m$; in other words, we end up with $a_m \sim a_s$. This apparent coincidence is reminiscent of $\tau_i \sim \tau_\star$, discussed in Section 8.2.4.1. In a thought-provoking work, Balbus (2014) argues that this "coincidence" could be a consequence of anthropic bias.

More specifically, the condition $a_m \sim a_s$ leads to strong tidal modulation, which may have formed a network of tidal pools during the Devonian era and provided the impetus for the origin of tetrapods (four-limbed vertebrates). We have also recently proposed that strong tides play a potentially important role in the inception of life, dictating biological rhythms, redistributing nutrients, and facilitating photosynthesis (Lingam & Loeb 2018b). Hence, when all these reasons are considered in tandem, perhaps there is indeed a strong case to be made for the centrality of Earth's large moon in the origin and evolution of life. Even if this were true, it should be recognized that the *absence* of such large moons does not automatically preclude planets from giving rise to complex life.

Plate tectonics: Plate tectonics is often considered to be one of the classic requirements for planetary habitability. The most important facet of plate tectonics on Earth has been its role in regulating the surface temperature, and therefore the availability of water, as part of the carbonate-silicate cycle; in particular, plate tectonics enables subduction to occur and regulate the atmospheric levels of carbon dioxide. In addition, it has been linked with generating the planetary magnetic field and creating and shaping continents. Despite these manifold advantages, plate tectonics cannot be said with certainty to be absolutely essential for the evolution of complex life, as noted in Section 5.2. The question of what planets are most suited for plate tectonics has been the source of much debate, with some studies concluding that planets more massive than the Earth are ideal for plate tectonics, while other studies obtained the opposite result.

In addition, many other factors are traditionally considered to be requirements for habitability; a review of this subject is expounded in Cockell et al. (2016). If all these factors are taken into account, the value of N_e will probably be lower than the estimate of ~ 0.1 that we have used thus far. Having said that, it seems unlikely that N_e would become so minuscule as to serve as the dominant explanation for Fermi's paradox. To further decrease the value of N_s in the Drake equation, we must look to other factors in it.

8.2.4.3 The value of f_l is very low

One of the most favored explanations for those who adopt the stance of "We are alone" is that f_l must be infinitesimally small. As we have already

dealt with the origin of life in detail in Chapter 2, we shall not address how and when it occurred.

In Section 8.1.5, we saw that the probability that the Earth is the only world to host life is very small, as it was on the order of 10^{-24} to 10^{-26}. Yet, an argument often invoked by creationists, as well as a few scientists who contend that life on Earth is unique, is that incredibly small probabilities always arise in the context of abiogenesis. For instance, consider a hypothetical protein composed of 300 amino acids (a real-world example, albumin, contains 584 amino acids). As there are twenty different canonical amino acids, the probability that it assembles through pure chance is $20^{-300} \approx 10^{-390}$. The hidden assumptions present in this derivation are that all outcomes are equally favored and that the processes involved in assembly and synthesis are completely random. This picture is clearly an oversimplification. Laboratory experiments in prebiotic chemistry have demonstrated that the synthesis of amino acids and other biomolecules is not impossible and that subsequent polymerization can occur in multifarious conditions.

Now, let us consider the opposite viewpoint, which posits that life is easy on the basis of arguments related to the origin of life on Earth. On our planet, the timescale for abiogenesis was < 0.8 Gyr and possibly as low as ~ 0.01 Gyr. Since the Earth is expected to be habitable for ~ 6 Gyr, we see that life appears to have arisen quickly on Earth. This has led some scientists to conclude that abiogenesis is common in the Universe by drawing on the Copernican Principle. In reality, the situation is a lot more complicated as the detailed analysis by Spiegel and Turner (2012) demonstrates—the final answer turns out to depend crucially on the prior assumptions made about abiogenesis. To wit, neither of the two scenarios (life is easy; life is hard) can be ruled out at this stage given the available information. Discovering a second independent origin of life elsewhere will enable us to place more stringent constraints on the value of f_l.

As per our current knowledge, all that can (and should) be said is that we possess insufficient knowledge to make any definitive statements pertaining to this subject. Yet, the fact that life did emerge fairly rapidly on Earth offers us tentative grounds to believe, while bearing the aforementioned caveats in mind, that it might not be extremely uncommon in the Universe.

8.2.4.4 The value of f_i is very low

After the origin of life, there were some major evolutionary innovations before intelligence arose (by *intelligence,* we do not refer to technological

intelligence, as we will address the latter when discussing f_c). We tackled the potential critical steps in Section 3.9, so only a brief summary is provided below.

Prokaryotes to eukaryotes: The transition from prokaryotes to eukaryotes remains one of the most famous examples of a major evolutionary transition. Eukaryotes possess many unique traits that are absent in prokaryotes, including phagocytosis (facilitating the acquisition of food via predation) and the presence of mitochondria (essential for energy metabolism) and plastids (which enable photosynthesis in algae and plants). Eukaryotes are believed to have evolved only once in the entire history of the Earth, and they possess features—most notably, the presence of mitochondria—that may have permitted the subsequent increase of biological complexity. Hence, if the origin of eukaryotes was indeed a fluke event, there is a real possibility that many planets are home to microbial life but not to complex life.

Complex multicellularity: Multicellularity apparently evolved independently several dozens of times on Earth, suggesting that its emergence is probable on other planets. On the other hand, the evolution of complex multicellularity (which includes plants, animals, and fungi) has been much rarer (six times in all). *Complex multicellularity* refers to organisms that possess cell and tissue differentiation, enabling them to efficiently transport nutrients, oxygen, and signals. The evolution of animals and plants led to a profound alteration of the biosphere, especially after the Cambrian explosion, which was accompanied by a major diversification of these clades. Some recent hypotheses—outlined in Section 3.6.4—claim that the aftermath of Snowball Earth episodes (characterized by extensive glaciation) led to the delivery of nutrients to oceans and eventually gave rise to more complex ecosystems in the Ediacaran and Cambrian eras. If this hypothesis proves to be correct, it demonstrates the intricate interplay of astronomical, geological, chemical, and biological factors that enabled the radiation of complex multicellular organisms. Hence, if the analogs of the Cambrian explosion (and the myriad factors that led to it) are rare, perhaps many planets have simple multicellular or unicellular organisms but not the likes of plants and animals.

Is intelligence rare? In referring to *intelligence* here, we are concerned only with higher cognitive functions and not technological intelligence specifically. Mounting evidence indicates that complex brains have evolved many times in different species; see Section 3.7. Although primates, elephants, and dolphins are the best known among them, this list also includes crows and octopuses. Thus, at least insofar as the evolution of higher cognitive functions on other worlds is concerned, at first glimpse it does not appear to be particularly difficult.

Gaian bottleneck: The Gaia hypothesis was first proposed by James Lovelock in the 1970s, and it has proven to be the source of much controversy ever since. Lovelock argued that organisms interact with their environment in a synergistic fashion and are responsible for maintaining clement conditions for their continued existence via biotic feedback mechanisms. While it is generally accepted that co-evolutionary processes between organisms and their environment do exist, the degree to which biological regulation occurs is less clear. Drawing on the Gaia conjecture, Chopra and Lineweaver (2016) hypothesize that life does not evolve sufficiently quickly in most instances to modify its environment to maintain planetary habitability thereafter. While this idea comes across as promising, its applicability is manifestly contingent on the Gaia hypothesis being valid on Earth and other habitable worlds.

On the whole, the transition from the origin of life to high (but non-technological) intelligence does appear to have involved a certain number of hard evolutionary steps. As such, it seems plausible that this paradigm could be invoked to explain why complex life is rare in the Universe, even if simple life is not. This idea was precisely the central thesis of the Rare Earth hypothesis by Ward and Brownlee (2000).

8.2.4.5 The value of f_c is very low

In our journey thus far, we are at the point where we have discussed the likelihood of intelligence on other worlds. However, note that the factor f_c requires at least two further transitions: the first is technological intelligence, and the second involves the development of technology to the degree that it becomes detectable. We will briefly tackle these aspects below.

1. Clearly, in order for technology to arise, tool making and tool usage are essential prerequisites. Tool usage has been shown to be present in a wide range of animals, suggesting that it is not particularly unique. On the other hand, tool making is rarer, especially when it increases in complexity; the great apes are the most widely studied animals in this regard.
2. When we turn our attention to genus *Homo* (our ancestors), it seems safe to conclude that none of them evolved the same degree of technology as *H. sapiens* prior to their extinction. Hence, this raises the question of whether some extraterrestrial species may not evolve beyond the usage of rudimentary technology so that they would not be able to undertake either interstellar travel or communication.
3. Another crucial factor which ostensibly differentiates us from all extant animals of Earth is our capacity for language. As this topic was covered in Section 3.8, we will not delve into it further. Given the available evidence at this stage, humans are unique in having a symbolic language with grammar and syntax, which is partly responsible for allowing them to share and acquire information on an unprecedented scale. Without a sophisticated method of communication, it seems unlikely that ETIs would be able to take part in interstellar travel and information transmission.
4. Lastly, especially in discussions of post-biological intelligence, it has been suggested that not all (technologically) intelligent species will be conscious, and vice versa. While this may be true, the more contentious claim in some quarters is that ETIs sans consciousness would not be interested in either communication or travel in order to reach out to other intelligent species.

Thus, on the whole, there are reasons to believe that humans have evolved certain unique features that enabled them to become the planet-spanning technological species that we observe today. Naturally, we cannot say for certain whether technological intelligence is extremely rare or relatively common in the Universe. However, on account of the reasons we have discussed above, at the very least, the prospect that it is rare should not be dismissed.

8.2.4.6 *The value of L is very low*

> The sound of the Gion *shōja* bells echoes the impermanence of all things; the color of the *sōla* flowers reveals the truth that the prosperous must decline. The proud do not endure, they are like a dream on a spring night; the mighty fall at last, they are as dust before the wind.
>
> —*Heike monogatari* (c. 1240)

It is widely believed that the value of L is finite,[11] i.e., even highly advanced technological species will eventually metamorphose into "dust before the wind," in the spirit of the above poetic excerpt from *Heike monogatari*. Hence, it is conceivable that the lifetime over which ETIs remain detectable could be very low. This solution was quite popular in the early days of the Drake equation as humanity was in the midst of the Cold War. Many authors speculated that ETIs would inevitably destroy themselves. For instance, Shermer (2002) estimated that $L \approx 300$ yr for modern civilizations on Earth since the fall of Rome. The fiducial value of $N_s \sim 10^4$ was derived from the assumption that $L \sim 10^6$ yr. Instead, if we choose $L \sim 300$ yr, it can be seen that $N_s \sim 3$, implying that there is a high likelihood we are the only species with a sufficiently advanced level of technology.

Hence, a potential explanation for Fermi's paradox, albeit a depressing one perhaps, is that ETIs face a variety of catastrophes and do not succeed in bypassing them.[12] For a detailed assessment of global catastrophic risks confronting humanity, the reader may consult Bostrom (2002), Bostrom & Ćirković (2008), and Rees (2018). We have listed some of the salient catastrophes below; they are grouped by whether the causes are primarily natural or anthropogenic. Examples often listed in the latter category are as follows:

1. Climate change: The ramifications arising from anthropogenic climate change are well-documented. The most commonly

---

11. However, Dyson (1979) proposed an innovative scheme by which life may persist ad infinitum in an open universe; on the other hand, the validity of his underlying physical assumptions has been called into question by others (Krauss & Starkman 2000).

12. Another possibility, raised by science-fiction writer Karl Schroeder, is that environmental conditions change gradually instead of abruptly. As a consequence, the adaptive significance of technological intelligence can end up being diminished. In turn, it may lead to this particular trait being lost over time; to put it differently, intelligence is viewed as being ephemeral.

cited example is global warming that may eventually lead to a pronounced greenhouse effect, but other stressors connected with climate change include rapid loss of biodiversity, reduced nutritional yields, and ocean acidification.

2. Artificial Intelligence: Owing to the rapid advancements in AI, the advent of humanlike as well as superintelligent AI seems possible even within the next century. Many scientists have suggested that humans would be to superintelligent AI what cockroaches are to us. Hence, the feasibility of mutual coexistence would be governed by the motives of superintelligent AI to a substantial degree; any conflict between them and humans could precipitate the extinction of the latter.
3. Biotechnology: There are two different avenues by which biotechnology can become a global catastrophic risk. The first is a deliberate bioterrorism attack, and the second is the inadvertent release (or production) of genetically engineered organisms. The Black Death may have killed as much as 60 percent of Europe's population, and one could readily envision higher extinction percentages if biotechnology comes into play (Sotos 2019).
4. Nanotechnology: The primary risk from nanotechnology arises from molecular manufacturing that could be used to either augment existing weapons or facilitate the emergence of other catastrophic risks (e.g., malignant AI). Another scenario often invoked in discussions of L is "gray goo," in which self-replicating nanobots consume the entire biosphere, but its likelihood has been lowered in some current estimates.
5. Warfare: The risk posed by nuclear and other forms of warfare is self-evident for the most part and has already been discussed extensively in SETI studies during the Cold War.

At this stage, two important caveats are worth highlighting. First, some of the scenarios will not inescapably lead to the extinction of all humanity. Second, and perhaps more importantly, these are *anthropogenic* catastrophic risks—it would seem rather implausible to contend that one of these fates will befall every single technological species on the basis of crude extrapolation of past and current challenges confronting *Homo sapiens*.

Turning our attention to natural catastrophic risks, some of the potential astrophysical candidates are large asteroid and comet impacts, superflares, GRBs, and supernovae, whereas Earthbound dangers include natural pandemics, megatsunamis, and supervolcanoes. However, in most cases, the probability of these events (per year) appears to be relatively minuscule, and, assuming that humanity survives long enough to navigate the anthropogenic catastrophes, our descendants could formulate effective mitigation strategies (at least in principle) that lower the risks posed by these phenomena.

Whatever the actual reason, if L is indeed very small for virtually all ETIs, it would serve as a viable explanation for Fermi's paradox.

8.3 CONCLUSION

> Thou dost think that thy mind wonderful Nature can grasp.
> Thus the astronomer draws his figures over the heavens,
> So that he may with more ease traverse the infinite space,
> Knitting together e'en suns that by Sirius-distance are parted,
> Making them join in the swan and in the horns of the bull.
> But because the firmament shows him its glorious surface,
> Can he the spheres' mystic dance therefore decipher aright?
>
> —Friedrich von Schiller, *Human Knowledge*

In this chapter, we have sought to address two central foundations of SETI—namely, the Drake equation and Fermi's paradox. Although both of these (especially the latter) may be perceived as philosophical twaddle by some individuals, it is important to recognize that they play a fundamental role in guiding our search strategies and are therefore relevant from a practical standpoint.

For instance, as we shall discuss in Chapter 9, the importance of searching for extraterrestrial artifacts has become increasingly pertinent in recent times. This development has partly been motivated by our improved understanding of the difficulties involved in traditional SETI—that is, the search for deliberately transmitted electromagnetic signals. In adumbrating the proposals for resolving Fermi's paradox, we have come to appreciate that hunting for electromagnetic signals is akin to looking for a needle in an astronomical haystack as well as to recognize that post-biological ETIs would probably look and act very different from us.

We began our discussion by introducing the history of the Drake equation and explaining the various terms and the concomitant underlying assumptions. The Drake equation has been rightly critiqued by many authors for being incomplete, and we delineated some of these issues. Yet, at the same time, it is necessary to reiterate that the original purpose of the Drake equation was to function as a heuristic tool and stimulate further research, and we opine that it has done an excellent job on this front. Subsequently, rough estimates for the number of extant ETIs capable of interstellar communication in our Galaxy were furnished, and we noted that it might be orders of magnitude larger than unity under optimal circumstances; in contrast, pessimistic estimates imply that humanity is the only technological intelligence in the Milky Way or even the entire observable Cosmos. When one further accounts for the expectation that even a single ET species could establish settlements across the entire Galaxy in a span much shorter than the latter's age, the question of "Where is everybody?" becomes all the more puzzling; this question has come to be misleadingly dubbed Fermi's paradox.

There are three broad classes of solutions to Fermi's paradox. The first, and perhaps the most unlikely of the trio, is that ETIs have already visited our neighborhood. Although this notion may seem far-fetched, we drew on earlier analyses to show that the probability of ruling out the absence of ETIs (and their artifacts) is not considerable. Future explorations of the Solar system, which would be carried out at increasingly high resolutions, will enable us to further constrain the probability of ETIs being absent in our neighborhood. The second class of explanations contends that ETIs do exist, but we have not detected them. There are a wide spectrum of solutions ranging from the possibility that we live in an isolated bubble not visited by ETIs to the fundamental axiom that absence of evidence cannot be construed as evidence of absence. The latter is presumably quite reasonable, since we have undertaken explicit searches for signatures of ETIs only during the past sixty years.

The last set of solutions contends that humans (in their capacity as technological species) are alone, or very rare, in our Galaxy. Although this stance violates the Copernican notion that we are not privileged, contingency does play an important role in evolution, and the series of unlikely events that led to the emergence of technological intelligence on Earth might have been incredibly unusual. On the other hand, more examples of evolutionary

convergence are coming to light (McGhee 2011), due to which we cannot say for certain that the analogs of major evolutionary innovations that paved the way for technological intelligence on our planet could not be reproduced elsewhere. If we are indeed truly alone, perhaps we would do well to heed what Nobel Prize winner Jacques Monod had to say at the conclusion of his seminal work, *Chance and Necessity* (1971, p. 180):

> The ancient covenant is in pieces; man knows at last that he is alone in the universe's unfeeling immensity, out of which he emerged only by chance. His destiny is nowhere spelled out, nor is his duty. The kingdom above or the darkness below: it is for him to choose.

Chapter 9

THE QUEST FOR TECHNOSIGNATURES

> There are countless suns and an infinity of planets which circle round their suns as our seven planets circle round our sun. . . . It is not unreasonable to suppose that there are planets which circle round other suns, not perceived by reason of great distance or small mass or not much (reflecting) water on their surface . . . and there must be plants and minerals in the worlds of space like those of our earth or different. We can attribute life to worlds with better reason than we can to our own earth.
>
> —Giordano Bruno, *De l'infinito, universo e mondi*

One of the most profound and ubiquitous characteristics of life on Earth is niche construction. We encountered several remarkable examples in this context at the inception of Chapter 6. The chief advantage of niche construction from the viewpoint of detecting extraterrestrial life is that the ensuing environmental modifications give rise to tangible signatures that are discernible by telescopes (e.g., molecular oxygen). At this stage, it must be recognized that niche construction is not solely restricted to nontechnological life.[1] In fact, by virtue of their reliance on technology, human beings and other intelligent systems have become adept at carrying out energy capture and conversion, computation, and habitat transformation on unprecedented scales. Niche construction by technological entities is accordingly characterized by its unique combination of large spatial scales and fast temporal scales. It is, therefore, the resultant "accelerated" transformation of the biosphere that engenders a veritable panoply of signatures indicative of technology (Vernadsky 1945; Zalasiewicz et al. 2017).

1. The basic datum that both nontechnological and technological organisms participate in niche construction is compatible with the postulate that members of both categories evince a distinctive propensity for biological learning (Watson & Szathmáry 2016).

As we know of just one species on Earth that has attained a high level of technological intelligence (to wit, *H. sapiens*),[2] it is instructive to adumbrate some of the radical transformations wrought by humanity:

- The invention of the Haber-Bosch process facilitated a massive increase in the production of ammonia (used in fertilizers), thus precipitating changes in the global nitrogen cycle that are potentially more dramatic than any prior fluctuations within the past ~ 2.5 Gyr.
- The release of industrial carbon into the atmosphere since the eighteenth century has increased the atmospheric CO_2 inventory to levels that are higher than the last ~ 1 Myr and yielded a rate of ocean acidification not documented in the past ~ 300 Myr.
- Anthropogenic activities such as mining, construction, and waste disposal may have brought about the largest proliferation of new minerals since the Great Oxidation Event. The manufacture and dispersal of plastics and concrete have been so widespread that it is easy to currently identify their signatures in sediments; in view of the longevity of plastics, these signatures might persist over Myr timescales.
- One of the strongest signals of human technology is the sharp increase in the surface abundance of radionuclides, owing to nuclear weapons testing. For example, an excess of ^{14}C and the appearance of ^{239}Pu in the geological record is traceable to the 1950s, with their respective concentrations reaching a peak in 1964. As these two radionuclides are not sufficiently long-lived (i.e., with half-lives of $\sim 10^4$ yr), other candidates such as ^{244}Pu and ^{247}Cm may be detectable over Myr timescales in the event of extensive nuclear warfare because their half-lives are $\gtrsim 10^7$ yr.
- The rates of species extinction are ~ 100 to 1000 times higher than their pre-anthropogenic values. It is becoming starkly evident that our planet is experiencing the sixth large-scale mass extinction in the Phanerozoic.

---

2. Many species on our planet are ingenious tool makers, as described in Section 3.7, but it can safely be averred that none of them displays technological abilities that are sophisticated enough to dramatically alter the biosphere. Needless to say, advanced technological intelligence does not, in any manner whatsoever, automatically imply that human beings represent the zenith of evolution.

In toto, there is broad consensus that human activities have been prominent enough to merit the introduction of a new epoch—namely, the Anthropocene (Lewis & Maslin 2015; Waters et al. 2016). A key point worth reiterating here is that the aforementioned changes have occurred over very short timescales—to wit, at apparently higher rates than previous anomalies in the geological record. Given how many of these feedbacks are ostensibly inimical to Earth's biosphere in the long run, there is growing awareness and acknowledgment of the acute necessity to guide and stabilize the eco-evolutionary trajectory of our planet in the increasingly turbulent Anthropocene (Steffen et al. 2018).

When viewed in this spirit, niche construction constitutes one of the key underlying themes and linkages between nontechnological and technological life as well as biosignatures and technosignatures (i.e., signatures of technology); the latter phrase appears to have been coined for the first time by Tarter (2007). As one transitions from abiotic worlds to those with sparse and rich biospheres and thence to sustainable technospheres, the number, complexity, and magnitude of feedback mechanisms are all expected to increase (Vernadsky 1945), as seen in Figure 9.1. In Section 6.7, we saw how thermodynamic disequilibrium—a measure of the available Gibbs free energy—could presumably distinguish between abiotic worlds and those with biospheres. The same concept was extended by A. Frank et al. (2017) to encompass generic technological agents capable of generating free energy; the resultant classification scheme is illustrated in Figure 9.1.

If lifeforms do exist beyond the Earth, at least some of them might have conceivably attained technological levels comparable to, or surpassing that of, *Homo sapiens*. This notion has a long, diverse, and intricate history (Weintraub 2014; Lingam & Loeb 2020g), with the excerpt at the beginning of this chapter by Giordano Bruno constituting one such notable example. The detection of these technological species might therefore be feasible by searching for signatures of their technology. An alternative is that they establish unambiguous direct contact with Earth—for instance, by alighting near the Great Pyramid of Giza and greeting humanity—but this scenario does seem quite improbable prima facie.

Hence, in this chapter, we will explore the rich array of technosignatures that could result from the actions of extraterrestrial technological intelligences (ETIs) along with the means of identifying them. As opposed to many prior conventional treatments of this subject, we will look beyond

Figure 9.1 The dominant mechanisms driving the generation and dissipation of free energy for different categories of worlds. The dissipation of free energy successively increases with the introduction of novel feedbacks associated with each class. (© Elsevier Ltd. *Source:* Adam Frank, Axel Kleidon, and Marina Alberti [2017], Earth as a hybrid planet: The Anthropocene in an evolutionary astrobiological context, *Anthropocene* 19: 13–21, fig. 2.)

electromagnetic technosignatures (e.g., optical and radio signals) and duly examine the prospects for detecting technological artifacts. One of the major advantages with the latter approach is that it permits us to look for signatures of extinct ETIs and not merely extant ones. The artifact technosignatures that humans can envision are, of course, inextricably a function of our current technological level and will likely evolve in the future.

Before embarking on our analysis, a comment regarding nomenclature is in order. In our parlance, SETI refers to the Search for Extraterrestrial Technological Intelligence (B. W. Jones 2003). We adopt this abbreviation because the conventional definition of SETI as the Search for

ExtraTerrestrial Intelligence is rather vague and misleading, since many species on Earth possess high intelligence, albeit of the nontechnological kind. Hence, the adoption of this alternative meaning for SETI is arguably more accurate and pertinent. Lastly, there has been a recent trend in the literature to avoid referencing SETI altogether because of the "giggle factor" associated with it (J. T. Wright & Oman-Reagan 2018). Although this viewpoint is indubitably pragmatic and apposite in multiple respects, we shall draw inspiration from the sentiments espoused in Shakespeare's famous lines from *Romeo and Juliet,* admittedly in a different context:

> What's in a name? that which we call a rose,
> By any other name would smell as sweet

Hence, at the risk of being labeled cavalier and naive vis-à-vis semantics, both of these terms—to wit, *technosignatures* and SETI—will be employed interchangeably hereafter.

9.1 RADIO TECHNOSIGNATURES

Searching for anomalous radio signals is often perceived as being nearly synonymous with SETI. There are several reasons why the two areas have dovetailed. First and foremost, the advantage with radio signals is that they travel at the speed of light; naturally, this also applies to electromagnetic signals at other frequencies. Second, unlike alternative information carriers such as gravitational waves or neutrinos, it is comparatively cheaper and energy efficient to build detectors and seek out electromagnetic signals (Hippke 2018a). Third, from a historical perspective, modern SETI was initiated in 1959 by the seminal work of Giuseppe Cocconi and Philip Morrison, who advocated searching for radio signals. This study was shortly followed by Project Ozma, the pioneering SETI experiment conducted by Frank Drake in 1960.

It must be appreciated, however, that the feasibility of interstellar communication via electromagnetic signals was articulated by many intellectuals in both technical and popular science outlets several decades prior to Cocconi and Morrison (1959), as meticulously reviewed in Dick (1993). In one of the earliest such treatises published in *Nature,* Ernest William Barnes, the English scientist and theologian (who was the bishop of Birmingham), posited that (1931, p. 722):

> As I have already indicated, I have no doubt that there are many other inhabited worlds, and that on some of them beings exist who are immeasurably beyond our mental level. We would be rash to deny that they can use radiation so penetrating as to convey messages to the Earth. Probably such messages now come. When they are first made intelligible a new era in the history of humanity will begin.

This publication, which was farsighted in many respects, purveyed a curious mixture of optimism and pessimism with regard to humanity's establishing contact with ETIs and the ensuing implications.

Humanity has transmitted radio signals for a period of ~ 100 years. Hence, the radio waves would have traversed a radial distance of ~ 30 pc, which encompasses $\sim 3 \times 10^4$ stars. If there are any ETIs situated within this bubble, we might hear from them in the future. It should be noted that the volume of the bubble exhibits a cubic dependence on the duration of radio signaling (until the scale height of the Milky Way is reached), thus implying that our likelihood of hearing back from any existing ETIs also obeys the same scaling.

For a detailed historical account of radio SETI, the reader may consult Dick (2018). Succinct overviews of radio technosignatures are presented in the likes of Morrison et al. (1979), Tarter (2001), and Gindilis and Gurvits (2019), whereas Margot et al. (2019) provide a synopsis of upcoming projects and telescopes in this arena.

9.1.1 Searching for radio signals

Let us contemplate the scenario where ETIs are broadcasting deliberate signals (beacons) and we seek to identify them. A vast number of variables confront us, but some of the most prominent among them include the choice of signaling frequency (for electromagnetic signals) and when and what direction to search for signals. This question can be naturally expressed in the language of game theory, since it amounts to studying equilibria in cooperative games wherein the participants cannot communicate with one another. These strategies, which are considered mutually obvious to all parties, are known as focal points or Schelling points. The latter nomenclature stems from Nobel laureate Thomas Schelling, who emphasized the importance of such strategies. Remarkably, in his classic work, *The Strategy of Conflict* (1960), Schelling noted that the identification of the "right"

signaling frequencies in SETI constitutes an example of a focal point. For an up-to-date overview of Schelling points in the context of technosignatures, the reader is referred to J. T. Wright (2018a).

9.1.1.1 Potential frequencies for interstellar beacons

As the preceding discussion illustrates, there is a pressing need to discern what frequencies are most suitable for beacons. Ever since the inception of modern SETI by Cocconi and Morrison (1959), the hydrogen 21 cm line (H I line) has always commanded a special place. This line originates from the transition of an electron between two hyperfine levels of the ground state of the hydrogen atom leading to the emission of a photon; these levels correspond to the cases where the magnetic dipole moments of the electron and proton are aligned or anti-aligned. The energy difference associated with this transition is approximately 5.9×10^{-6} eV, which translates to a photon wavelength of 21.1 cm and frequency of 1420.4 MHz (1 MHz is 10^6 Hz). Indeed, this frequency was precisely the one adopted by Frank Drake for Project Ozma.

In the same spirit, the hyperfine transition associated with the hydroxyl radical (OH) is approximately 1.67 GHz (1 GHz is 10^9 Hz). Hence, in the classic SETI literature, the band of frequencies between 1.4 and 1.7 GHz has been dubbed the *water hole*. A key reason behind the attractiveness of the water hole as a potential Schelling point is that the two chemical species involved (H and OH) are formed from the dissociation of water, one of the chief prerequisites for life-as-we-know-it. From a more whimsical perspective, the water hole is reminiscent of pools (watering holes) in arid climates where wild animals congregate to satiate their thirst. Other physical and psychological arguments for choosing the water hole can be found in Oliver (1979). Last, and most important, the water hole falls within the terrestrial microwave window (TMW). The TMW ranges from ~ 1 to 15 GHz and is valuable because this region has a low level of background noise. As one approaches frequencies > 15 GHz, emission and absorption by water and oxygen molecules in Earth's atmosphere raise the background noise. On the other hand, if we consider frequencies < 1 GHz, synchrotron radiation (photons emitted by accelerated charged particles) from the Milky Way increases the ambient noise.

Apart from the TMW, a number of other radio bands have been described in the literature. For instance, instead of beacons, it is also possible

to eavesdrop on leakage emitted by ETIs. Most of the radio power emitted by human technology lies in the range 50–400 MHz, owing to which it was proposed by Loeb and Zaldarriaga (2007) that searches for artificial radio leakage should focus on these frequencies. As electromagnetic leakage is typically isotropic, it would be possible to detect ETIs at the same technological level as humans only up to a distance of ~ 10–100 pc (Loeb & Zaldarriaga 2007). However, the strength of this signal is boosted during gravitational microlensing events, implying that the latter can serve as natural telescopes for unearthing artificial sources. For optimistic choices of the Drake equation, Rahvar (2016) found that ~ 1 beacon per year may be detectable through microlensing.

Drake and Sagan (1973) pointed out another natural frequency that was well suited for space-based telescopes. The current temperature of the cosmic microwave background (CMB) is $T_{\rm CMB} \approx 2.7$ K. Hence, one can compute the associated frequency $\nu_{\rm CMB}$ using

$$\nu_{\rm CMB} = \frac{k_B T_{\rm CMB}}{h} \approx 56.3\ {\rm GHz}. \tag{9.1}$$

However, inasmuch as the above frequency is concerned, we remark that $T_{\rm CMB} \approx 2.7\,(1+z)$ K, where z denotes the redshift. Hence, any ETIs transmitting at redshifts $z \gtrsim 1$ may employ frequencies that are greater than $\nu_{\rm CMB}$. A comparable frequency was advocated by Kardashev (1979), who suggested that the wavelength $\lambda \approx 1.5$ mm ($\nu \approx 200$ GHz) was optimal for sending and receiving reliable beamed signals across interstellar distances.

9.1.1.2 Types of artificial signals

A number of modulation schemes might be employed by ETIs to distinguish artificial radio signals from those produced by astrophysics. The conventional expectation is that such signals are likely to have a bandwidth that is narrower than those originating from astrophysical sources. Astrophysical masers, which emit coherent electromagnetic radiation through stimulated emission, have a bandwidth of $\Delta\nu \sim 500$ Hz. In contrast, human technology is easily capable of achieving $\Delta\nu/\nu \sim 10^{-12}$ (Tarter 2001) at certain wavelengths, which is much narrower than the typical value of $\Delta\nu/\nu \sim 10^{-6}$ for astrophysical sources.

However, an important point must be appreciated as regards narrowband signals, i.e., signals whose bandwidth is smaller than those generated

by astrophysical phenomena. A number of factors can contribute to the broadening or modulation of these signals. First, the observed frequency is shifted due to the orbital motion of the Earth around the Sun and its spin. The rotational velocity of the Earth is expected to introduce a frequency shift ($\Delta\nu_{\rm spin}$) on the order of

$$\Delta\nu_{\rm spin} \sim \nu\left(\frac{v_{\rm spin}}{c}\right) \sim 1.5\times10^{3}\,{\rm Hz}\,\left(\frac{\nu}{1\,{\rm GHz}}\right) \tag{9.2}$$

over the course of one day (rotation period), where $v_{\rm spin} \approx 460$ m/s. Similarly, over the span of one year (orbital period), the Earth's orbital velocity induces a frequency shift ($\Delta\nu_{\rm orb}$) that is given by

$$\Delta\nu_{\rm orb} \sim \nu\left(\frac{v_{\rm orb}}{c}\right) \sim 10^{5}\,{\rm Hz}\,\left(\frac{\nu}{1\,{\rm GHz}}\right), \tag{9.3}$$

where $v_{\rm orb} \approx 30$ km/s for planets in the habitable zones of solar-type stars.

As the signal passes through the interstellar medium (ISM) and the interplanetary medium (IPM), it is subject to multiple scatterings by the ionized component of the respective medium. This phenomenon induces spectral broadening of the signal, thereby yielding a practical lower bound for the minimum frequency resolution. The broadening caused by the ISM ($\Delta\nu_{\rm ISM}$) is expressible as

$$\Delta\nu_{\rm ISM} \approx 0.14\,{\rm Hz}\,\left(\frac{\nu}{1\,{\rm GHz}}\right)^{-6/5}\left(\frac{V_{\perp}}{100\,{\rm km/s}}\right){\rm SM}^{3/5}, \tag{9.4}$$

where $V_{\perp}$ is the relative velocity of the source perpendicular to the observer's line of sight, and SM is known as the scattering measure that embodies the magnitude of electron density fluctuations; for more details, the reader may consult Cordes and Lazio (1991). The broadening arising from the IPM ($\Delta\nu_{\rm IPM}$) is given by

$$\Delta\nu_{\rm IPM} \approx 0.075\,{\rm Hz}\,\left(\frac{\nu}{1\,{\rm GHz}}\right)^{-6/5}\left(\frac{R_{\rm sun}}{100\,R_{\odot}}\right)^{-9/5}, \tag{9.5}$$

where $R_{\rm sun}$ signifies the radial distance from the Sun. The total spectral broadening contributed by these regions was estimated to be $\sim 0.1\,{\rm Hz}\,(\nu/1\,{\rm GHz})^{-6/5}$ in Siemion et al. (2014). Apart from these constraints, a

radially accelerating transmitter with respect to the receiver will emit signals whose frequency is not constant over time. It is thus necessary to correct for this phenomenon, known as the *drift rate* or *chirp,* to maximize the efficacy of future searches for narrow-band electromagnetic signals, while bearing computational limitations in mind. A cumulative drift rate of 200 nHz (e.g., tantamount to drift of 200 Hz/s at 1 GHz) was espoused in Sheikh et al. (2019) on the basis of rigorous calculations.

Although narrow-band signals were considered the norm in radio SETI for decades, many recent studies have argued in favor of broadband signals. An overview of this subject is presented in Siemion et al. (2014), while more technical treatments can be found in Benford et al. (2010) and Messerschmitt (2015). Broadly speaking, several major factors have been advanced in favor of broadband signals. First, from an economic standpoint, the construction and operation costs of beacon transmitters are potentially lowered for broadband systems. Second, wide bandwidth emerges as a corollary of minimizing the energy expended per bit delivered. Third, using techniques akin to spread spectrum (which has wide bandwidth) on Earth is advantageous because (1) they are immune to radio-frequency interference, (2) they mitigate the dispersion and scattering effects of the ISM and IPM, and (3) they are easier to design and implement.

Hitherto, we have implicitly concerned ourselves with continuous wave signals. In contrast, it is possible for short electromagnetic pulses to be emitted. At this stage, we note that the time-bandwidth product is well-known to possess a lower bound imposed by the classical uncertainty principle. The time-bandwidth product is defined as the product of the duration of the signal (τ_s) and its bandwidth ($\Delta\nu$), with its minimum value being 0.25. Now, in the event that $\tau_s \Delta\nu \approx 1$, we see that decreasing $\Delta\nu$ leads to an increase in τ_s, and vice versa. In particular, it is feasible to emit highly transient pulses that are unlikely to originate from astrophysical phenomena; we will return to this topic later.

9.1.1.3 On the detectability of signals

The sensitivity of the receiving telescope sets limits on the maximum distance at which transmitters can be detected for a given power, and vice versa. To quantify this statement, a few concepts must be introduced.

The effective isotropic radiated power (EIRP) is defined as

$$\mathrm{EIRP} = \frac{4W_t}{\theta_A^2}, \tag{9.6}$$

where W_t is the transmitter's power and θ_A denotes the angular width of the beam. For the diffraction-limited case, we have $\theta_A \approx c/(\nu D_t)$, with D_t representing the transmitter's effective diameter. For the sake of comparison, note that the EIRP for the planetary radar at Arecibo is $\sim 2 \times 10^{13}$ W. The signal power (W_{sig}) is given by

$$W_{\mathrm{sig}} = \eta_A \Phi_{sig} A_{\mathrm{eff}}, \tag{9.7}$$

where Φ_{sig} is the flux (from the emitter) received at the location of the detector, A_{eff} is the effective collecting area, and η_A is a constant of order unity that encapsulates the aperture efficiency. The standard deviation (σ_{noise}) associated with photon noise is expressible through the so-called radiometer equation as

$$\sigma_{\mathrm{noise}} = k_B T_{\mathrm{sys}} \sqrt{\frac{\Delta\nu}{\Delta t}}, \tag{9.8}$$

where Δt is the integration time and T_{sys} is known as the system noise—that is, the equivalent temperature of the input noise. The signal-to-noise ratio (SNR) is defined as $W_{\mathrm{sig}}/\sigma_{\mathrm{noise}}$, which allows us to solve for Φ_{sig} using the above equations. Thus, we end up with

$$\Phi_{sig} = \mathrm{SNR} \frac{k_B T_{\mathrm{sys}}}{\eta_A A_{\mathrm{eff}}} \sqrt{\frac{\Delta\nu}{\Delta t}}. \tag{9.9}$$

Hence, if Φ_{sig} is known for a given receiver, it is possible to solve for the EIRP as follows,

$$\mathrm{EIRP} = 4\pi d_\star^2 \Phi_{sig}, \tag{9.10}$$

where $d_\star$ is the distance from the Earth to the location of the putative transmitter. In fact, by combining (9.6) and (9.10) and assuming the signal is diffraction-limited, it is possible to solve for the actual power of the transmitter (W_t). Instead of computing the flux, it is advantageous to calculate the flux density $S_{\mathrm{sig}} = \Phi_{sig}/\Delta\nu$, which yields

$$S_{\rm sig} \approx 2.2 \times 10^{-2}\,{\rm Jy}\,\eta_A^{-1} \left(\frac{\rm SNR}{5}\right) \left(\frac{\Delta\nu}{1\,{\rm Hz}}\right)^{-1/2} \times \left(\frac{\Delta t}{10^3\,{\rm s}}\right)^{-1/2} \left(\frac{A_{\rm eff}/T_{\rm sys}}{10^4\,{\rm m^2/K}}\right)^{-1}. \quad (9.11)$$

In closing, we note that the above formulae are applicable to artificial transmitters. Instead, it is conceivable that ETIs might piggyback onto astrophysical sources and repurpose them as powerful transmitters. Radio pulsars are a natural example in this regard, as it may be feasible for ETIs to modulate pulsar emission and thereby encode information. The presence of such pulsar modulators is potentially discernible by searching for excess thermal emission in the UV during the phases when the pulses are nullified (Chennamangalam et al. 2015).

9.1.1.4 Temporal markers for signals

It is important to not only identify the right frequency but also look in the right direction at the right *time*. The latter becomes particularly important when the signal, from the perspective of an observer on Earth, is not *on* at all times. The obvious strategy that comes to mind is using rare astrophysical phenomena for synchronized signal transmissions (Makovetskii 1978). A number of astrophysical events have been postulated as temporal Schelling points, some of which are described below.

1. One of the earliest Schelling points for transmission was proposed by G. W. Pace and Walker (1975). They argued that ETIs around single stars would broadcast their signals to binary star systems at apastron and periastron—that is, the greatest and least distance between the stars, respectively.
2. A number of transient and explosive astrophysical phenomena have been propounded as temporal signposts that can be utilized by the transmitting ETI to initiate their transmission. Four of the most notable candidates in this regard are supernovae (Tang 1976; Lemarchand 1994), stellar flares (Siebrand 1982), Gamma Ray Bursts (Corbet 1999), and binary neutron star mergers (Nishino & Seto 2018; Seto 2019).

3. A closely related approach was outlined in Edmondson and Stevens (2003) with millisecond pulsars serving as the beacons. The basic idea is that transmitting ETIs temporally synchronize their transmissions with a pulsar and broadcast their signals toward habitable stars situated within a given angular separation and distance from the pulsar. Receiving ETIs would need to therefore look for signals broadly arriving in the direction of the pulsar with temporal modulations characteristic of the pulsar.
4. Corbet (2003) suggested a simple search strategy from the standpoint of the receiver: observations of the targets are to be conducted when their angular separation from the Sun is highest as seen from the Earth—namely, when they are at *opposition,* in astronomical parlance. However, it must be noted that this approach requires the transmitter to have accurate knowledge of the Earth's orbit, the Sun's location, and relative velocity.
5. The importance of transiting planets in SETI has been appreciated since the early study by Filippova and Strelnitskij (1988). The Earth's transit zone has attracted much attention in this context, since any ETIs in this region could have discerned the presence of life on Earth via transit spectroscopy (see Section 6.1). Inverting this argument, ETIs on transiting exoplanets (as seen from Earth) may utilize the duration of transit to broadcast electromagnetic signals or distort the transit light curves into artificial shapes (Kipping & Teachey 2016). Over time, ETIs might establish contact with each other through this avenue, thereby forming nodes and edges in a Galactic network.

9.1.1.5 All-sky surveys and targeted searches

There are two different modes of searching for radio signals. One can either scan the entire sky by looking in all directions or choose individual targets that have potentially higher prospects for hosting ETIs. Unfortunately, rigorous decision-making is next to impossible as we know neither the spatial distribution of ETIs nor the EIRP of their putative transmitters. Despite this uncertainty, a couple of qualitative conclusions are derivable. If the majority

of transmitters have low values of EIRP, it is more likely that targeted observations of nearby planetary systems will prove to be successful. The converse is true in the event that there are a sufficiently large number of transmitters continuously broadcasting at high power.

Since it is not yet apparent which of the above two strategies represents the more optimal one, several approaches have sought to combine elements of both. One possible approach is to select an isotropic volume of random stars, corresponding to a uniform prior, because this entails a minimum of anthropocentric assumptions about what stars and planets are more likely to develop ETIs. On the other hand, an equally good case can be made for handpicking target stars and planets that are Earthlike in superficial terms, given we know for certain that technological intelligence emerged at least once on Earthlike planets (namely, our own). Extending this notion of searching for radio signals in the vicinity of potentially habitable planets, one may also search within regions of the Galaxy that are believed to be most conducive to hosting life—that is, the so-called Galactic Habitable Zone (GHZ).

9.1.2 SETI surveys to date

In Section 8.2.3.4, we pointed out the remarkable (albeit not surprising) fact that we have searched only a minuscule fraction ($\sim 10^{-18}$) of the cosmic haystack for radio signals. Hence, it seems reasonable to surmise that the absence of evidence to date cannot be construed as robust evidence of absence; much more of the search volume needs to be sampled by future surveys before the existence of ETIs can be ruled out with a high degree of certainty. A thorough historical summary of the radio searches undertaken through the end of the twentieth century is presented in Tarter (2001). We will restrict ourselves only to describing a few of the major SETI projects initiated since 2010. For more details, the reader is referred to the excellent web resource aptly named Technosearch and hosted by the SETI Institute.[3]

1. The Wow! signal was detected in 1977 by Ohio State University. It was a strong narrow-band ($< 10^4$ Hz) emission near the 21 cm line of hydrogen with an estimated flux density

3. See https://technosearch.seti.org/

of ~ 60 Jy. A number of subsequent searches were undertaken for this putative signal, but they proved to be unsuccessful. Gray and Ellingsen (2002) searched the location of the Wow! signal but did not find any evidence after scanning a bandwidth of 2.5 MHz across fourteen hours (spread over multiple days) with a flux density threshold of ~ 14 Jy. An even more sensitive and exhaustive survey also yielded null results (Harp et al. 2020).

2. Siemion et al. (2013) searched for narrow-band radio emissions between 1.1 and 1.9 GHz with spectral resolution of $\delta\nu = 0.75$, 1.49, and 2.98 Hz using the Green Bank Telescope. The total number of targets searched was eighty-six stars confirmed to host transiting exoplanets, with the flux sensitivity being $\Phi_{sig} \sim 2 \times 10^{-26}$ W/m^2. The survey concluded that $\lesssim 1$ percent of transiting exoplanetary systems were likely to host radio transmitters with an EIRP of $\sim 1.5 \times 10^{14}$ W.
3. Harp, Richards, Tarter, et al. (2016) used the Allen Telescope Array to search for narrow-band radio signals between 1 and 9 GHz in multiple bands with a spectral resolution of $\delta\nu \geq 0.7$ Hz. The total number of stars surveyed was 9293, and the telescope had a frequency-dependent flux sensitivity of $\Phi_{sig} \sim 1.8$–3.1×10^{-24} W/m^2. No persistent radio signals were detected in the sample at this threshold.
4. Tingay et al. (2016) carried out the first search for radio signals using the Murchison Widefield Array (MWA) at low radio frequencies ranging from 103 to 133 MHz, with a spectral resolution of $\delta\nu = 10^4$ Hz. The field of view encompassed by this survey was 400 sq deg (note that 1 sq deg $\approx 3 \times 10^{-4}$ steradians) and comprised forty-five known exoplanets. The sensitivity limit, expressed in terms of the flux density defined in (9.11), was $S_{\rm sig} \sim 0.4$–0.6 Jy. The upper bound on the EIRP was $\gtrsim 10^{14}$ W, which is much higher than equivalent transmitters in this frequency range on Earth.
5. A study similar to the one above was conducted by Tingay et al. (2018) using the MWA. The frequency range chosen was 99–122 MHz and the spectral resolution employed was $\delta\nu = 10^4$ Hz. The survey's field of view was 625 sq deg and consisted of twenty-two known exoplanets. The sensitivity

limit associated with this survey was $S_{\rm sig} \sim 0.25$–0.6 Jy, and the upper bound on the EIRP was computed to be $\gtrsim 10^{13}$ W; note that the EIRP varies with the distance to the target, as seen from (9.10).

6. R. H. Gray and Mooley (2017) employed the Jansky Very Large Array (VLA) to search for radio signals from the Andromeda (M31) and Triangulum (M33) galaxies. The frequency chosen was the 21 cm line with a spectral window of 0.125–1 MHz and a spectral resolution of $\delta\nu = 15$–122 Hz. No continuous narrow-band signals were found to exceed the flux density threshold of $S_{\rm sig} \sim 0.24$–1.33 Jy.

Last, the Breakthrough Listen project merits special attention because it has allotted $100 million (US) in funding for the next ten years and is the largest SETI undertaking to date.[4] The reader may consult Worden et al. (2017) and Isaacson et al. (2017) for a description of the scope and aims of Breakthrough Listen, while Lebofsky et al. (2019) elaborate on the process of collecting, reducing, and archiving data. As massive quantities of data are being continually generated and analyzed by Breakthrough Listen, our description cannot be wholly up-to-date and has little pretensions to this effect.

Bearing this caveat in mind, we will briefly mention some of the salient achievements by Breakthrough Listen. For starters, Breakthrough Listen has conducted several studies of Fast Radio Bursts (FRBs), which are powerful radio pulses with temporal durations of $\lesssim 10^{-3}$ s that originate from distant galaxies at gigaparsec distances; we will return to FRBs later in this chapter. Second, Breakthrough Listen has investigated unusual objects in the vicinity of our Solar system such as the retrograde asteroid (514107) 2015 BZ_{509} (of possibly interstellar origin) and the first confirmed interstellar object ‘Oumuamua, which was detected in 2017 while passing through our Solar system. Third, a transient radio signal allegedly from the nearby M-dwarf Ross 128 situated at a distance of 3.4 pc was scrutinized by Breakthrough Listen, and subsequent analysis suggested that the emission originated from Earth-based satellites.

4. Breakthrough Initiatives (n.d.), Listen, https://breakthroughinitiatives.org/initiative/1

On a related note, ten planetary systems, including the well-known TRAPPIST-1 and LHS 1140 systems, were surveyed by the Green Bank Telescope at 1.15–1.73 GHz, with the result being that no candidates of nonterrestrial origin were discovered (Pinchuk et al. 2019). Likewise, twenty nearby stars in the Earth Transit Zone (ETZ)—to wit, the region of the sky where hypothetical observers would see the Earth transit the Sun—were scanned by the Green Bank Telescope at frequencies of 3.95–8.00 GHz (Sheikh et al. 2020). The null evidence for radio transmitters was used to rule out the presence of emitters with power capabilities ~ 0.1 to 88 percent that of the Arecibo radar transmitter. It was also concluded that $\gtrsim 0.6$ percent of all systems in the ETZ at distances of ≤ 150 pc are unlikely to host these emitters. Lastly, Breakthrough Listen intends to cover the comprehensive *Exotica Catalog* encompassing "one of everything" in astrophysics that is of interest from a SETI perspective (Lacki et al. 2020).

Apart from these individual objects, Breakthrough Listen also carried out a systematic survey of 692 nearby stars using the Green Bank Telescope (Enriquez et al. 2017). The observations were made in the standard 1.1–1.9 GHz at a spectral resolution of $\delta\nu = 3$ Hz. At this resolution, the minimum detectable flux density was estimated to be $S_{\text{sig}} \sim 17$ Jy. After making use of (9.10), it was determined that none of the analyzed systems appeared to host radio transmitters with an EIRP of $\sim 10^{13}$ W, which is achievable by present-day human technology. Hence, this result indicates that the fraction of stellar systems at distances < 50 pc having radio transmitters with the above EIRP may be $< 10^{-3}$. A more exhaustive survey along similar lines was carried out by D. C. Price et al. (2020) for 1327 nearby stars in the frequency range of 1.1–3.45 GHz, and no concrete evidence for putative transmitters with EIRP of $\sim 10^{12}$–10^{13} W was identified.

Given the different types of surveys that have been undertaken as well as the infinitesimal fraction of the search volume covered, several metrics have been proposed to quantitatively gauge the reach of SETI surveys. Metrics of this kind are known as "figures-of-merit" and vary from study to study; a comprehensive list can be found in Wright, Kanodia, and Lubar (2018). One of the most widely employed figures-of-merit is the one proposed by Frank Drake in 1984. This figure-of-merit (denoted by DFM) is given by

$$\text{DFM} \propto \frac{\text{BW}_i \Omega_{\text{sur}}}{\Phi_{sig}^{3/2}}, \tag{9.12}$$

where BW_i is the total bandwidth of the instrument and Ω_{sur} denotes the solid angle of the sky encompassed by the survey. The DFM is proportional to the product of the total instrumental bandwidth and the volume covered by the survey. To see why, note that the latter is approximately expressible as $\Omega_{\mathrm{sur}} d_\star^3$. After using (9.10) while holding EIRP fixed, we obtain $d_\star \propto \Phi_{sig}^{-1/2}$. Substituting this expression into the survey volume, we end up with $\Omega_{\mathrm{sur}}/\Phi_{sig}^{3/2}$. It must be recognized that the DFM assumes that all regions of the sky are equally likely to host detectable transmitters. This premise may represent an idealization, except in the interesting scenario wherein most ETIs are technologically very advanced (perhaps post-biological in nature) and therefore capable of living anywhere in the vast expanse of space. This runs counter to the conventional intuition that ETIs would be found in the vicinity of stars, galaxies, and the like.

Another figure-of-merit outlined in Enriquez et al. (2017) is termed the survey speed figure-of-merit (SSFM). The basic idea is that the SSFM is roughly proportional to the instrumental bandwidth and the integration time (Δt). The inclusion of the latter implies that a survey with SSFM twice that of another survey will take only half the time to complete the same search, provided the bandwidth is held constant. From (9.9), it is apparent that $\Phi_{sig} \propto \Delta t^{-1/2}$, implying that $\Delta t \propto \Phi_{sig}^{-2}$. Hence, the SSFM can be simplified to

$$\mathrm{SSFM} \propto \mathrm{BW}_i \Phi_{sig}^{-2}. \tag{9.13}$$

The SSFM does not incorporate the factor of Ω_{sur} because it is applicable to targeted surveys in which the fraction of the sky covered does not play a major role. Figure 9.2 depicts the two figures-of-merit outlined here (DFM and SSFM). It is apparent from inspecting the figure that Breakthrough Listen scores highly on both metrics.

9.2 OPTICAL AND INFRARED TECHNOSIGNATURES

One of the chief reasons why radio technosignatures were historically preferred was because the radio power output of the host star is minimal compared to that emitted by artificial transmitters. In contrast, the overall background contribution from the host star is substantial, which imposes comparatively stringent requirements on the transmitter. However, on the basis of the experimental and theoretical strides made in the development of masers during the 1950s, Robert Schwartz and Nobel

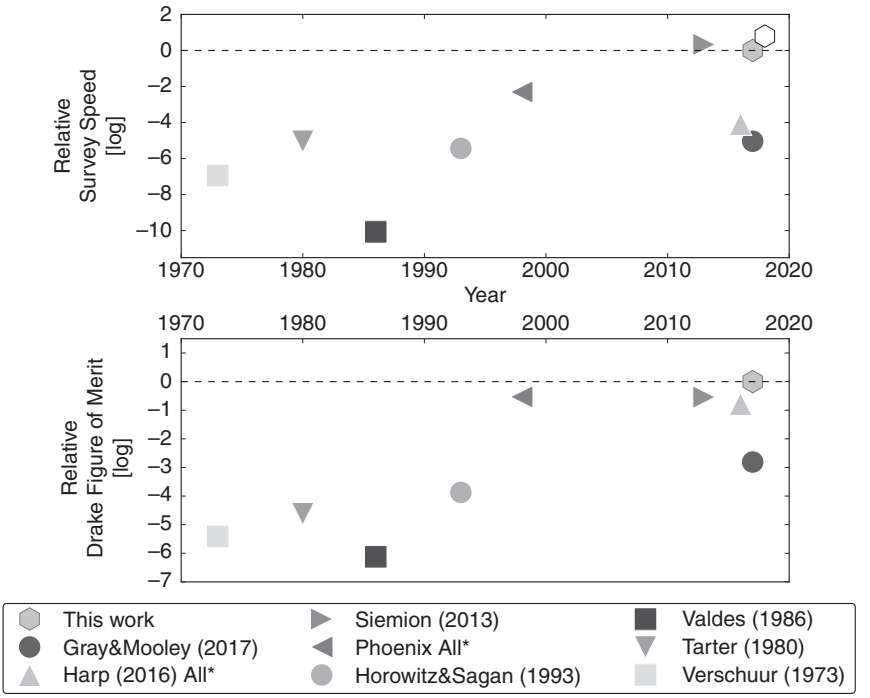


Figure 9.2 *Top:* The survey speed figure-of-merit (SSFM) defined in (9.13). *Bottom:* The Drake figure-of-merit (DFM) defined in (9.12). Both figures-of-merit are normalized such that their values for the Breakthrough Listen survey by Enriquez et al. (2017) equal unity. The white hexagon in the upper panel represents the magnitude of the SSFM if the entire bandwidth accessible to Breakthrough Listen was utilized (around 5 GHz).

*As some SETI projects were conducted in multiple stages or components, the use of the word "All" signifies that the cumulative value was calculated. For information on the references cited above, please consult the original paper by Enriquez et al. (© The American Astronomical Society. *Source:* J. Emilio Enriquez, Andrew Siemion, Griffin Foster, Vishal Gajjar, Greg Hellbourg, Jack Hickish, Howard Isaacson, Danny C. Price, Steve Croft, David DeBoer, Matt Lebofsky, David H. E. MacMahon, and Dan Werthimer [2017], The Breakthrough Listen search for intelligent life: 1.1–1.9 GHz observations of 692 nearby stars, *Astrophysical Journal* 849[2]: 104, fig. 6.)

laureate Charles Townes (who developed the maser) proposed that interstellar signals could be transmitted at optical and infrared (OIR) frequencies (Schwartz & Townes 1961). A vital point worth appreciating about OIR SETI is that the signals are likely to be highly collimated. Hence, the feasibility of eavesdropping on such signals through happenstance is very low.

Detecting OIR technosignatures has a higher probability of success when the beams are either weakly collimated or deliberately targeted toward the Earth (Forgan 2014). Note that there is an intrinsic trade-off: increasing the beamwidth leads to a decrease in photon flux and consequently lowers the prospects for detection, whereas decreasing the beamwidth enhances the likelihood of the signal standing out at the cost of requiring a very precise knowledge of the planet's position at all times (J. R. Clark & Cahoy 2018).

9.2.1 Principles of OIR technosignatures

Although surveys or theoretical analyses in the realm of OIR technosignatures are not as numerous compared to their radio counterparts, much progress has nonetheless been accomplished since the pioneering work of Schwartz and Townes (1961). To understand why OIR frequencies might be advantageous in some respects, we will adopt a similar line of reasoning to the one delineated in Townes (1983). After combining (9.6), (9.9), and (9.10) and solving for the SNR in the diffraction-limited regime, we find

$$\mathrm{SNR} \propto \nu^2 \Delta\nu^{-1/2}. \tag{9.14}$$

Now, suppose that the major contributor to $\Delta\nu$ is Doppler broadening. In this event, as $\Delta\nu \propto \nu$, we arrive at $\mathrm{SNR} \propto \nu^{3/2}$, which is consistent with the scaling relations presented in Townes after taking the appropriate limits. Hence, in this instance, we see that the SNR is higher at shorter wavelengths, implying that OIR might be favored over radio in terms of signal detection. Instead of the Doppler effect, if we specify broadening from the ISM or IPM that obeys $\Delta\nu \propto \nu^{-6/5}$, we obtain $\mathrm{SNR} \propto \nu^{13/5}$; as before, we see that shorter wavelengths yield higher SNR. We caution that our analysis is not definitive, as a number of physical factors were held fixed (e.g., system temperature) or neglected.

One of the first issues to address is whether any magic frequencies exist in OIR SETI, analogous to those noted in radio SETI (e.g., the 21 cm line). This issue has not attracted much attention, perhaps on account of the fact that the possible range of wavelengths (or frequencies) is narrower. Narusawa et al. (2018) proposed two favorable wavelengths from the perspective of the transmitting ETIs: the 1064 nm line associated with the Nd:YAG laser (involving a neodynium atomic transition) and its second harmonic generation, the 532 nm line. Other lines that were identified are close to

standard absorption lines known from astrophysics, such as 393.8 nm (near Ca K line), 589.1 nm (Na D_2 line), and 656.5 nm (near the Hα line).

In the search for OIR technosignatures, most surveys have been devoted to looking for short pulses as opposed to continuous wave (CW) signals. Let us define the pulse duration as τ_p and the transmitting laser's power by W_t. In other words, the number of photons emitted by the laser is given by $N_{\rm trans} = \tau_p W_t / (hc/\lambda)$, where the denominator is the energy of a single photon with wavelength λ. The photon flux received at a distance $d_\star$ is $N_{\rm trans}/\pi \Omega_t d_\star^2$, where $\Omega_t = (2\lambda/\pi D_t)^2$ and D_t is the aperture of the transmitting telescope; the factor of $2/\pi$ is present inside the brackets because lasers have a narrower waist that decreases the effective beamwidth. This flux must be multiplied by the aperture of the receiving telescope, whose diameter and area are D_r and $\pi D_r^2/4$, respectively, to calculate the number of photons received ($N_{\rm rec}$). Thus, after simplification, we obtain

$$N_{\rm rec} = \frac{\pi^2 \tau_p W_t D_t^2 D_r^2 \Gamma_{\rm ext}}{16 hc \lambda d_\star^2}, \tag{9.15}$$

where $\Gamma_{\rm ext}$ is the wavelength-dependent extinction factor that is generally close to unity for distances smaller than 1 kpc. For $\tau_p = 3 \times 10^{-9}$ s, $W_t \approx 10^{15}$ W, $D_r = D_t = 10$ m, $d_\star \approx 300$ pc, $\lambda = 349$ nm, and $\Gamma_{\rm ext} = \exp\left[-4.6 \times 10^{-4} (d_\star/1\,{\rm pc})\right] \approx 0.87$, we find that $N_{\rm rec} \sim 1500$ photons. In contrast, the number of photons emitted by a Sunlike star in the interval τ_p is ~ 0.1 (A. W. Howard et al. 2004). Hence, even at a technological level identical to that of humans in the year 2000, laser pulses outshine their host star by four orders of magnitude. Under the assumption of Moore's Law, human-made lasers in the upcoming decades could outshine Sunlike stars by ~ 6 orders of magnitude.

In Section 9.1.1.3, we saw how the minimum detectable flux could be calculated as well as the EIRP (and actual power) of the transmitter. A similar treatment is feasible for laser SETI by adopting the simplified analysis presented in Reines and Marcy (2002); we shall focus on the case where $D_r = D_t$. First, we begin by noting that the fraction of starlight contained within the beamwidth of the laser is given by

$$f_{\rm star} \approx \frac{1}{4\pi} \left(\frac{2\lambda}{\pi D_t} \right)^2. \tag{9.16}$$

Next, the spectral energy distribution of a Solar-type star has a total width of ~ 400 nm. We will assume that the linewidth of the putative transmitters is closely similar to the spectral resolution of the Keck High Resolution Echelle Spectrometer (HIRES)—namely, we utilize $\sim 7.5 \times 10^{-3}$ nm. Hence, the number of elements of this width across the spectral energy distribution of the star is $400/(7.5 \times 10^{-2}) \approx 5.3 \times 10^4$; this value may vary by a factor of $\gtrsim 10$ depending on the star's spectral type and the choice of wavelength. In order to ensure an $n\sigma$ detection, the minimum laser power required is

$$W_t \sim \frac{1}{4\pi} \left(\frac{2\lambda}{\pi D_t} \right)^2 \left(\frac{1}{5.3 \times 10^4} \right) \left(\frac{n\sqrt{2}}{\mathrm{SNR}} \right) L_\star, \tag{9.17}$$

where $L_\star$ is the stellar luminosity. Here, we have used the fact that the ratio of the photon signal to that of the noise constitutes the SNR; the signal in turn is determined from the fraction of stellar output within the beamwidth at the specified spectral resolution of the laser. The factor of $\sqrt{2}$ is included because the laser must compete against two sources of photon noise: the star and itself. After simplification, the above equation becomes

$$W_t \sim 2 \times 10^4 \ \mathrm{W} \left(\frac{n}{6} \right) \left(\frac{\lambda}{0.45\,\mu\mathrm{m}} \right)^2 \left(\frac{D_t}{10\,\mathrm{m}} \right)^{-2} \left(\frac{\mathrm{SNR}}{200} \right)^{-1} \left(\frac{L_\star}{L_\odot} \right). \tag{9.18}$$

For a pulsed laser, the above value of W_t must be multiplied by $\Delta t/(N_{\mathrm{pulse}} \tau_p)$, with N_{pulse} denoting the number of pulses received within the exposure time of Δt. The minimum detectable photon flux can be computed via

$$\Phi_{\mathrm{laser}} = \frac{W_t \lambda}{hcA_{\mathrm{laser}}}, \tag{9.19}$$

where A_{laser} is the area of the laser beam at the Earth that is given by

$$A_{\mathrm{laser}} = \pi \left(\frac{2\lambda}{\pi D_t} \right)^2 d_\star^2, \tag{9.20}$$

which is obtained by calculating the spot size (product of the beam opening angle and the distance) and, consequently, the corresponding area. After simplification, we end up with

$$\Phi_{\rm laser} \sim 1.8 \text{ photons m}^{-2}\,\text{s}^{-1} \left(\frac{n}{6}\right)\left(\frac{\lambda}{0.45\,\mu\text{m}}\right) \times \left(\frac{\text{SNR}}{200}\right)^{-1}\left(\frac{L_\star}{L_\odot}\right)\left(\frac{d_\star}{100\,\text{pc}}\right)^{-2}. \quad (9.21)$$

Instead of expressing the threshold in terms of the photon flux, it is possible to rewrite the expression in terms of the flux density. The latter is obtained by making use of $S_{\rm sig} \sim hc\Phi_{\rm laser}/(\lambda\delta\lambda)$, where $\delta\lambda \sim 7.5 \times 10^{-3}$ nm in our case. For the characteristic values specified above, we find that $S_{\rm sig}$ is on the order of 10^{-16} W m^{-2} nm^{-1}, although this merely represents a fiducial estimate.

9.2.2 OIR searches

Even though the number of searches for OIR signals have been comparatively fewer, a fair number of them have been carried out since 2000. We shall describe some of the most salient ones below.

1. Reines and Marcy (2002) searched for non-astrophysical emission lines in the wavelength range of 400–500 nm from 577 nearby F-, G-, K-, and M-type stars using Keck HIRES. It was concluded that none of these stellar systems appeared to host CW lasers with $W_t > 5 \times 10^4$ W that were oriented toward our planet.
2. A. W. Howard et al. (2004) carried out a detailed survey of $\sim 1.3 \times 10^4$ stars with a cumulative observation time of $\sim 2.4 \times 10^3$ h. The goal was searching for nanosecond pulses that could outshine Solar-type stars by ~ 4 orders of magnitude, but no valid events were detected.
3. Stone et al. (2005) carried out observations of 4605 F-, G-, K-, and M-type stars situated within 60 pc using the facilities at the Lick Observatory. No laser pulses were detected in this sample across 4.5 years of observations.
4. A. Howard et al. (2007) carried out a survey of the sky spanning an extent of 0.3 sq deg for nanosecond optical pulses. After adopting simplified models for the stellar density in the Milky Way and photon losses in the ISM, a tentative upper

bound of $\sim 3 \times 10^5$ for the number of currently transmitting ETIs was obtained.

5. Hanna et al. (2009) employed the Solar Tower Atmospheric Cherenkov Effect Experiment (STACEE) detector to search for nanosecond laser pulses at 420 nm around 187 stars with a detection threshold of 10 photons m^{-2}. No evidence for such laser pulses was found in the sample.
6. Tellis and Marcy (2015) made use of Keck HIRES to search for narrow emission lines in the vicinity of 2796 stars over the range 364–789 nm. It was concluded that there were no CW lasers with $W_t > 10^2$ W at $d_\star \sim 30$ pc and $W_t > 10^3$ W at $d_\star \sim 300$ pc directed toward the Earth in the observed sample.
7. The above study was extended by Tellis and Marcy (2017) using Keck HIRES to scrutinize 5600 stars across the wavelength range of 364–789 nm. The detection threshold for the power of the CW lasers ranged between 3×10^3 W and 1.3×10^7 W. The observed null result suggests that $< 10^{-3}$ temperate Earth-sized planets appear to have transmitters above this threshold oriented in the Earth's direction.
8. The Near-InfraRed and Optical SETI (NIROSETI) instrument installed at the Lick Observatory has the capacity to search for nanosecond pulses emitted at near-IR wavelengths of 0.9–1.75 μm. The instrument sensitivity is approximately 60–380 photons m^{-2} per nanosecond pulse, which enables it to detect current human-level laser emissions from $\lesssim 50$ pc away (Maire et al. 2016, 2019). The instrument has surveyed 1280 astronomical objects as of 2019, but no laser signals above this threshold have been identified (Maire et al. 2019).

The F-type star KIC 8462852—colloquially known as Tabby's Star or Boyajian's Star after Tabetha S. Boyajian, the lead author of the paper that elucidated the star's properties—has attracted much attention owing to its unusual and irregular changes in brightness. One of the candidates advanced to explain this phenomenon was ETI activity. Hence, both radio and optical SETI searches of this object were undertaken. Radio observations were made across the frequency range of 1–10 GHz, but no narrow-band signals were detected in a 1 Hz channel above the flux density threshold of

$S_{sig} \sim 180$–300 Jy; in contrast, if the channel was assumed to be 10^5 Hz, the threshold was lowered to 10 Jy (Harp, Richards, Shostak, et al. 2016).

The search for pulsed optical beacons discovered no evidence of signals above the sensitivity threshold of 1 photon m^{-2} at the Earth within an integration window of 12 ns (Abeysekara et al. 2016); a concomitant study by the Allen Telescope Array arrived at similar conclusions, albeit at a lower sensitivity of 67 photon m^{-2}. A more recent study of this star by the Breakthrough Listen initiative (using the Lick Observatory's Automated Planet Finder) effectively ruled out the prolonged existence of lasers with power of > 24 MW on the basis of the analysis of high-resolution spectra in the wavelength range of 374–970 nm (Lipman et al. 2019). Besides Tabby's Star, the irregular transiting star HD 139139 (EPIC 249706694) was examined by the Breakthrough Listen initiative via the Green Bank Telescope, but no radio signals with an EIRP of $> 10^{13}$ W were found (Brzycki et al. 2019).

To date, we have scanned $> 10^4$ stars for pulsed and CW laser signals with null detection, but it must be noted that the average dwell time per star has been low ($\sim 10^3$ s). The Pulsed All-sky Near-infrared Optical SETI (PANOSETI) observatory is an upcoming survey that is expected to address many of the limitations inherent in previous searches. For starters, it has a field of view of 10^4 sq deg, implying that the solid angle covered by this survey will be higher than current targeted searches by a factor of $\sim 10^6$. Second, the dwell time per source will be increased by $\sim 10^4$ times compared to previous OIR surveys. Lastly, the PANOSETI instrument will probe a much wider range of wavelengths (350–1650 nm) than present-day searches. The extension of the search into the near-IR is particularly advantageous since photon losses in the ISM are minimized and the background noise from the Milky Way is also comparatively lower. The reader is referred to S. A. Wright et al. (2018) for the technical details underlying PANOSETI.

Lastly, we note that optical signals transmitted by putative ETIs need not always occur in the form of beamed energy. Instead, they may opt to embed their signals in starlight, as the latter is a plentiful source of optical photons. The basic idea is that the signals would be manifested in the form of unusual temporal variations that are discernible in the stellar background (Kipping & Teachey 2016). Signals that are several orders of magnitude fainter than the light emitted by the host star are ostensibly detectable by a 1 m optical telescope (Stanton 2019).

9.2.3 Electromagnetic signals at other frequencies

Most of the surveys undertaken to date have been at radio and optical wavelengths, with a few in the IR regime. In principle, however, there is no reason why photons of higher energies could not be employed for the purposes of interstellar signaling. One important caveat is that the Earth's atmosphere typically absorbs photons at $\lambda < 200$ nm, owing to which such hypothetical transmissions would be detectable by space-based observatories.

Let us turn our attention to (9.15). If all other factors are held fixed, we see that the number of received photons is proportional to $\nu\,\Gamma_{\rm ext}$. Setting aside the issue of $\Gamma_{\rm ext}$, as the transmittance is expected to be nearly unity, it is apparent that photons with higher energies are more advantageous in terms of detectability. If we assume that each photon can encode a fixed number of bits, this also implies that the data rates are enhanced for high-energy photons. It is, however, important to appreciate that producing photons with arbitrarily higher frequencies may prove to be impractical even for technologically advanced ETIs due to (1) economic constraints on building emitters and (2) physical constraints on collimating beams (e.g., mirror roughness).

Bearing these caveats in mind, we can consider a handful of theoretical studies that posit the usage of X-rays for transmitting signals. Fabian (1977) proposed that dropping a kilometer-sized rock onto a neutron star may result in the emission of a very powerful X-ray pulse ($\sim 10^{29}$ W) that is potentially detectable throughout the Milky Way given Earth-level telescopes. Alternatively, emission from strong X-rays sources (e.g., X-ray binaries) could be modulated by sufficiently advanced ETIs, and such signals are possibly identifiable by piggybacking onto X-ray astronomy facilities (Corbet 1997).

9.3 MODALITY OF INTERSTELLAR SIGNALING

We have taken it for granted that any ETIs intent on deliberately broadcasting their presence would opt to employ electromagnetic signals. While many advantages stem from the use of photons to transmit information, they are not the only viable information carriers. ETIs may reveal their presence, either deliberately or unintentionally, through the artifacts that they produce. As this constitutes the subject of Section 9.5, we will not tackle this theme further. It suffices to say that artifact SETI has its own advantages and

encompasses a more diverse spectrum of technosignatures. In our subsequent analysis of artifact SETI in Section 9.5, we implicitly operate under the premise that the artifacts are typically large-scale and situated at the location of the ETIs.

However, it must be recognized that small-scale artifacts (either functional or defunct) could be "mailed" to other worlds. We can therefore merge the two broad themes encountered thus far: Is it feasible for ETIs to transmit artifacts, in lieu of photons, as repositories of information? In theory, the answer is yes and these artifacts will be referred to as *inscribed matter* hereafter. The inscribed matter "packages" would, in theory, reach the destination system via probes. The idea of deploying interstellar probes as a means of communication was first espoused by Bracewell (1960) in a pioneering study. Later overviews of the benefits accruing from interstellar probes are presented in Gertz (2016) and Hippke (2020). We will not spend much time on this topic, because we encountered it while discussing Fermi's paradox in Section 8.2. The seeming absence of probes in our Solar system has been interpreted as indicative of the absence of ETIs altogether by some authors, but a number of potential alternatives have been advanced, as elucidated in Section 8.2.

Here, we will exclusively focus on comparing the relative merits of inscribed matter (i.e., artifacts) and photons in their capacity as information carriers. Our approach will closely mirror that of Rose and Wright (2004) and Hippke et al. (2018). We do not deal with more exotic possibilities such as neutrinos and gravitational waves, as our understanding of their information encoding schemes is comparatively limited; we briefly commented on signaling using exotic particles in Section 8.2.3.4.

9.3.1 Energetic costs of information carriers

Let us commence our analysis with inscribed matter. The energy that needs to be expended (E_{mat}) to transmit $\mathcal{B}$ bits of information is merely the kinetic energy,

$$E_{\mathrm{mat}} = \frac{1}{2} m_{\mathrm{tot}} v_{\mathrm{mat}}^2, \tag{9.22}$$

where m_{tot} is the total mass and v_{mat} is the velocity at which the probe is launched. A couple of subtle points are worth highlighting with regard to m_{tot}. It is evident that the mass of the payload carrying the actual information

must be smaller than the total mass on account of two reasons: (1) onboard fuel and (2) ample shielding.

There are different routes by which the contribution from (1) can be minimized. First, it is possible to avoid carrying the fuel: one of the most optimal solutions in this respect is to rely on radiation pressure exerted by either lasers or starlight. Spacecraft reliant on this mode of propulsion are known as light sails, and they are further explicated in Section 9.5.2.1. In this scenario, it is feasible for the payload mass to become comparable to, albeit smaller than, the total mass. Second, the probe could be subjected to an initial impetus, preferably from space, in the general direction of a chosen target. The overall function of the probe in this context is to serve as a passive message in a bottle sans any propulsion system that is catapulted toward a given destination.

Next, let us turn our attention to (2). The requirements for shielding relativistic probes from gas and dust are explored in detail by Hoang et al. (2017) and Lingam and Loeb (2020f). In addition, the effects of bombardment by cosmic rays and high-energy photons are not always negligible. Moreover, the degree of shielding required will vary depending on whether the inscribed matter carrying the information is organic or nonorganic in nature, i.e., the shielding is dependent on the chemical composition of the information carrier. It is thus apparent that a general treatment of this issue is rendered difficult, albeit not impossible. We will, however, adopt the simplified prescription in Bialy and Loeb (2018) as it suffices for our purposes; a prior analysis of this subject was carried out by S. A. Stern (1986). If collisions with dust grains are the primary cause of evaporation, the minimum desired thickness $h_{\min}$ is determined using

$$h_{\min} \sim 4.6 \times 10^{-5}\,\mathrm{m} \left(\frac{\rho_{\mathrm{mat}}}{1000\,\mathrm{kg\,m^{-3}}}\right)^{-1} \left(\frac{n_{\mathrm{ISM}}}{10^{6}\,\mathrm{m^{-3}}}\right) \left(\frac{d_{\star}}{1\,\mathrm{kpc}}\right) \times \left(\frac{v_{\mathrm{mat}}}{100\,\mathrm{km/s}}\right)^{2} \left(\frac{\phi_{dg}}{0.01}\right) \left(\frac{\bar{m}_{\mathrm{mat}}}{12\,m_p}\right) \left(\frac{U_v}{4\,\mathrm{eV}}\right)^{-1}, \tag{9.23}$$

where n_{ISM} is the number density of protons in the interstellar medium, ϕ_{dg} is the dust-to-gas mass ratio in the ISM, and ρ_{mat}, $\bar{m}_{\mathrm{mat}}$, and U_v are the mass density, the average atomic or molecular mass, and the vaporization energy of the shield's material, respectively. Instead of dust grains, if collisions with

gas particles are the dominant driver of evaporation, the minimum thickness becomes

$$h_{\rm min} \sim 6.2 \times 10^{-6}\,{\rm m} \left(\frac{\rho_{\rm mat}}{1000\,{\rm kg\,m^{-3}}}\right)^{-1} \left(\frac{n_{\rm ISM}}{10^6\,{\rm m^{-3}}}\right) \left(\frac{d_\star}{1\,{\rm kpc}}\right) \times \left(\frac{Y_{\rm tot}}{0.01}\right) \left(\frac{\bar{m}_{\rm mat}}{12\,m_p}\right), \tag{9.24}$$

with $Y_{\rm tot}$ denoting the total sputtering yield that measures the number of particles that are ejected from the shield during the collision of the probe with a single particle.

Now, let us further suppose that the probe transporting the material has a cylindrical geometry, with the thickness of the shield being $h_{\rm min}$ while its inner radius is $r_{\rm mat}$. If the mass density of the shield and the information carrier within are similar, it is easy to show that the ratio of the shield mass to the total mass is $\sim 2h_{\rm min}/r_{\rm mat}$. By inspecting the above equations, we can see that this ratio is always much smaller than unity even for a small probe of $r_{\rm mat} \sim 1$ cm, provided that the probe velocity is not relativistic and $d_\star \lesssim 1$ kpc. Hence, to leading order, it is fairly reasonable to contend that the shielding mass required is quite minimal compared to that of the actual payload.

In light of the preceding discussion, we will operate under the assumption that a significant fraction of $m_{\rm tot}$ is devoted to the purpose of conveying information. Apart from the issues of shielding and propulsion, the costs associated with deceleration at the target also need to be taken into account. It was, however, shown by Rose and Wright (2004) that this factor is of order unity if the speed of the probe is comparable to that of chemical rockets. If the fraction of the total mass embodied in the information-carrying inscribed matter is $\eta_{\rm mat}$, then we have $m_{\rm tot} = \eta_{\rm mat}^{-1}\mathcal{B}/\bar{\rho}_B$, where $\bar{\rho}_B$ is the information density (units of bits kg^{-1}). Therefore, (9.22) is expressible as

$$E_{\rm mat} = \frac{1}{2}\frac{\mathcal{B}}{\bar{\rho}_B}\frac{v_{\rm mat}^2}{\eta_{\rm mat}}. \tag{9.25}$$

The information density for DNA is capable of reaching $\sim 5 \times 10^{23}$ bits kg^{-1} (Church et al. 2012). It is also possible to reach higher values of $\bar{\rho}_B$ in principle using perfluorocarbon chemistry (Hippke et al. 2018). Let us consider an arbitrary carbon atom in a chain. Two of its bonds are unavailable

because they link to the neighboring carbon atoms. In the remaining two sites, imagine that a total of two functional groups are allowed. Hence, a total of four possibilities are admitted (-CAA-, -CBB-, -CBA-, and -CAB-), but the latter duo exhibit identical chemical and structural properties in most cases. Hence, each carbon atom could store $\log_2(3) \approx 1.6$ bits. Next, we have to determine the two functional groups, A and B. It is advantageous to choose perfluorocarbons owing to their higher stability. It can be shown that this choice translates to $\sim 100\, m_p$ per bit. After further simplification, it is found that $\bar{\rho}_B \sim 6 \times 10^{26}$ bits kg^{-1}.

Jacob Bekenstein identified a fundamental upper bound on the entropy-to-energy ratio in a series of seminal publications (Bekenstein 1981), which can be recast in terms of the information density as follows,

$$\bar{\rho}_B < 2.6 \times 10^{43}\,\text{bits}\,\text{kg}^{-1} \left(\frac{r_0}{1\,\text{m}}\right), \tag{9.26}$$

where r_0 denotes the effective radius of the system. Even if we consider systems on the scale of nanometers, it is apparent that the Bekenstein bound on $\bar{\rho}_B$ is many orders of magnitude higher than the information density of real-world materials. Note that a higher value of $\bar{\rho}_B$ is beneficial since it leads to a reduction in E_{mat} if all other factors are held constant.

Next, let us turn our attention to the transmission of information by means of photons. The Shannon capacity, encountered in Section 7.7, which constitutes an upper bound on the net data rate ($\dot{\mathcal{B}}$), is given by

$$\dot{\mathcal{B}} \leq \frac{\Delta\nu}{\ln 2} \ln\left(1 + \text{SNR}\right) \leq \frac{\Delta\nu}{\ln 2} \text{SNR}, \tag{9.27}$$

where the last expression follows from $\ln(1+x) < x$. The noise is given by $N_0 \Delta\nu$, where $N_0 = k_B T_{\text{sys}}$ is the noise spectral density for thermal noise. The signal corresponds to the power intercepted by the receiving telescope and is determined by following the same line of reasoning outlined in Section 9.2.1. The power is defined as $P = dE/dt$ and $\dot{\mathcal{B}} = d\mathcal{B}/dt$, which allows us to integrate both the left- and right-hand sides of the above equation and invert it to solve for the minimum energy required for transmitting $\mathcal{B}$ bits of information. Denoting this quantity by E_{rad}, we obtain

$$E_{\text{rad}} \approx \mathcal{B} N_0 \frac{8\lambda^2 d_\star^2 \ln 2}{\pi^2 D_r^2 D_t^2}. \tag{9.28}$$

Let us assume that $D_r = D_t = D_0$ and compute the ratio of (9.28) and (9.25). This ratio, denoted by Δ_E, measures the relative energy efficiency of transmitting information via radiation as opposed to inscribed matter. Thus, we end up with

$$\Delta_E \sim 5.1 \times 10^8 \left(\frac{\eta_{\rm mat}}{0.1}\right)\left(\frac{\bar{\rho}_B}{10^{22}\,{\rm bits\,kg^{-1}}}\right)\left(\frac{T_{\rm sys}}{300\,{\rm K}}\right)\left(\frac{v_{\rm mat}}{0.01\,c}\right)^{-2} \times \left(\frac{\lambda}{1\,\mu{\rm m}}\right)^2\left(\frac{d_\star}{100\,{\rm pc}}\right)^2\left(\frac{D_0}{10\,{\rm m}}\right)^{-4}. \quad (9.29)$$

A number of interesting general inferences are feasible from the above formula. First and foremost, we expect $\Delta_E \gg 1$ in most realistic scenarios, thus implying that inscribed matter is a much more energy-efficient information carrier with respect to photons. More precisely, we observe that Δ_E increases when (1) the probe velocity is decreased, (2) the distance between source and destination is increased, (3) the telescope aperture is reduced, and (4) longer photon wavelengths are utilized.

Instead of computing the energy required to transmit a certain number of bits, it is easy to invert the argument and calculate the number of bits that can be transmitted per unit energy expended. In the case of inscribed matter, from (9.25) we obtain

$$\Upsilon_{\rm mat} \approx 2.2 \times 10^8\,{\rm bits\,J^{-1}} \left(\frac{\eta_{\rm mat}}{0.1}\right)\left(\frac{\bar{\rho}_B}{10^{22}\,{\rm bits\,kg^{-1}}}\right)\left(\frac{v_{\rm mat}}{0.01\,c}\right)^{-2}, \quad (9.30)$$

where $\Upsilon_{\rm mat}$ is the information efficiency (i.e., number of bits encoded per unit energy expenditure) of inscribed matter.

In order to determine the analog of (9.30), we begin by observing that the rate of photons $\dot{N}_{ph}$ collected by the receiver under the assumption of diffraction-limited beaming is

$$\dot{N}_{ph} \approx \frac{W_t D_t^2 D_r^2 \Gamma_{\rm ext}}{4hc\lambda d_\star^2}, \quad (9.31)$$

which is nearly identical to (9.15); the slight differences arise due to minor variations in the minimal beamwidth of lasers and telescopes. Next, we note that encoding $\gtrsim 1$ bit per photon is permitted by theory (Giovannetti et al. 2014) and has also been achieved in practice (Tentrup et al. 2017), although

the latter implicitly presupposes that the sender and receiver have complete knowledge of the encoding scheme utilized. Hence, by multiplying (9.31) with the fiducial conversion factor of 1 bit per photon and dividing by the transmitter's power, we arrive at

$$\Upsilon_{\rm rad} \approx 1.3 \times 10^{-3}\,{\rm bits\,J^{-1}}\,\Gamma_{\rm ext}\left(\frac{\lambda}{1\,\mu{\rm m}}\right)^{-1}\left(\frac{d_\star}{100\,{\rm pc}}\right)^{-2}\left(\frac{D_0}{10\,{\rm m}}\right)^{4}, \tag{9.32}$$

where $\Upsilon_{\rm rad}$ denotes the information efficiency for photon communications; to derive this formula, we have used $D_r = D_t = D_0$.

9.3.2 Date rate of information carriers

For the sake of completeness, we will also quantify the data rates of information transfer from the sender to the receiver.

For photons, we determined the photon rate in (9.31) and argued that an encoding of 1 photon per bit is feasible. Hence, the data rate at the receiver's end ($\dot{\mathcal{B}}_{\rm rad}$) is given by

$$\dot{\mathcal{B}}_{\rm rad} \approx 1.3 \times 10^{3}\,{\rm bps}\,\Gamma_{\rm ext}\left(\frac{W_t}{1\,{\rm MW}}\right)\left(\frac{\lambda}{1\,\mu{\rm m}}\right)^{-1} \times \left(\frac{d_\star}{100\,{\rm pc}}\right)^{-2}\left(\frac{D_0}{10\,{\rm m}}\right)^{4}, \tag{9.33}$$

where bps is bits per second. As seen from the above formula, $\dot{\mathcal{B}}_{\rm rad}$ is higher when the aperture size or transmitter power are increased and when the wavelength or distance to the receiver are decreased.

Next, we turn our attention to inscribed matter probes. The amount of information contained in the probe is $\mathcal{B} = \eta_{\rm mat}\bar{\rho}_B m_{\rm tot}$. The time taken by the probe to reach the recipient is approximately $d_\star/v_{\rm mat}$. Hence, the net data rate for inscribed matter ($\dot{\mathcal{B}}_{\rm mat}$) is estimated to be

$$\dot{\mathcal{B}}_{\rm mat} \sim 9.7 \times 10^{8}\,{\rm bps}\left(\frac{\eta_{\rm mat}}{0.1}\right)\left(\frac{\bar{\rho}_B}{10^{22}\,{\rm bits\,kg^{-1}}}\right) \times \left(\frac{m_{\rm tot}}{1\,{\rm kg}}\right)\left(\frac{v_{\rm mat}}{0.01\,c}\right)\left(\frac{d_\star}{100\,{\rm pc}}\right)^{-1}, \tag{9.34}$$

implying that $\dot{\mathcal{B}}_{\rm mat}$ is typically higher than $\dot{\mathcal{B}}_{\rm rad}$ unless the mass of the inscribed matter is very small or the probes are expelled at low velocities. One of the striking features that is evident from comparing (9.33) and (9.34) is that the former falls off as the inverse square of the distance, whereas the latter is inversely proportional to it. As a result, ceteris paribus, inscribed matter is favored over longer distances.

9.3.3 Other aspects of inscribed matter

Although inscribed matter is generally very advantageous from the dual perspectives of energetics and data rates, it also has its share of drawbacks. First, it is difficult to accurately target the probe, especially if there is no onboard propulsion system and the distance to the destination is very far. To bypass both these shortcomings, a large number of inscribed matter parcels would need to be "posted" by the putative ETIs. Second, the travel time is much longer when it comes to inscribed matter. When compared to electromagnetic signals, inscribed matter will take a factor of $\sim c/v_{\rm mat}$ to traverse the same distance. The long travel times, in turn, impose additional constraints on two-way communication. In our scheme, the first ETI sends a probe to the second ETI with information, and this gesture is symmetrically reciprocated. In order for the first ETI to still exist when the return probe arrives, we require $L > 2d_\star/v_{\rm mat}$, where L denotes the longevity of the ETI.

In spite of these caveats, the preceding analysis illustrates that inscribed matter represents a genuine alternative to photons as information carriers. Hence, it is instructive to search for probes and other artificial objects (e.g., space debris) in our Solar neighborhood. It is more likely that such hypothesized artifacts would have been deliberately sent toward our Solar system instead of being random isotropic emissions because the latter strategy necessitates far more probes to achieve the same number density. Suppose that an ETI situated at a distance of $d_\star$ opts to target Earthlike planets in the habitable zone. The angular resolution required is estimated as $10\,{\rm mas}\ (a/1\,{\rm AU})\left(d_\star/100\,{\rm pc}\right)^{-1}$, where $1\,{\rm mas} = 4.85\times10^{-9}$ rad. If the probes are targeted at this spherical sector instead of being emitted in an isotropic fashion, the resultant gain ($\mathcal{G}_{\rm mat}$) will be

$$\mathcal{G}_{\rm mat} \sim 1.7\times10^{15}\left(\frac{a}{1\,{\rm AU}}\right)^{-2}\left(\frac{d_\star}{100\,{\rm pc}}\right)^{2}. \tag{9.35}$$

Not surprisingly, this expression is mathematically analogous to the antenna gain when electromagnetic radiation is emitted as a collimated beam instead of isotropically.

How are artificial objects, perhaps bearing information from ETIs, located within our neighborhood detectable? If they display signatures of artificial illumination or specular reflection, they are potentially identifiable using the strategies explained in Section 9.5.6. On the other hand, if hypothetical probes are conducting flybys of our planet and Solar system, it is much more difficult to detect them. We will need to look for anomalies in their aspect ratios, albedos, orbital trajectories, and chemical composition, to name a few. Bialy and Loeb (2018) conjectured that the interstellar object 'Oumuamua might be artificial in origin. This unconventional proposal runs counter to the mainstream viewpoint, which advocates a "purely natural origin for 'Oumuamua" (Bannister et al. 2019). Owing to the relative paucity of observational data, conclusively invalidating or proving this notion is probably next to impossible.

9.4 ON THE CLASSIFICATION OF TECHNOLOGICAL AGENTS

Thus far, we have deliberately opted not to comment on the characteristics of putative ETIs. Yet, it is apparent that some kind of a classification scheme is in order. Consider, for instance, the variables such as transmitter power and telescope aperture that appeared in many of the equations we encountered while discussing electromagnetic (radio and optical) technosignatures. Both of these quantities are likely to implicitly depend on the resources available to a given ETI, which is itself strongly dependent on the technological level attained by the ETI.

9.4.1 The Kardashev Scale and its extensions

Any classification scheme of ETIs must ideally possess the virtues of simplicity, physical basis, and consequences for detectability. The most famous among them is the scale that was formulated by Nikolai Kardashev in 1964 and now bears his name. The Kardashev Scale measures the energy consumption (i.e., power) of an ETI (W_{ETI}), which may serve as an indirect proxy for its technological level. This classification scheme is easy to quantify, employs power—an intrinsic physical quantity—as the metric, and

can produce a wide range of technosignatures depending on how the ETI utilizes the captured energy.

Kardashev's (1964) original formulation defined three major classes:

- Type I: ETIs whose technological levels are close to humans' in the 1960s, with a corresponding energy consumption of $\sim 4 \times 10^{12}$ W.
- Type II: ETIs capable of harnessing the energy of a Sunlike star, with an energy consumption of $\sim 4 \times 10^{26}$ W.
- Type III: ETIs that possess energy resources roughly on the scale of the Milky Way—that is, whose energy consumption is $\sim 4 \times 10^{37}$ W.

Subsequent studies have tended to redefine Type I ETIs as those with an energy expenditure equal to the total starlight incident on the planet. In other words, Type I ETIs would evince an energy consumption of

$$W_{\mathrm{T1}} \approx 1.7 \times 10^{17}\ \mathrm{W} \left(\frac{L_\star}{L_\odot}\right) \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a}{1\ \mathrm{AU}}\right)^{-2}, \tag{9.36}$$

where $L_\star$ is the stellar luminosity, while R and a denote the radius and orbital distance of the planet, respectively. It is important to avoid confusing Kardashev's version of Type I ETIs and the above definition, as the difference between the two is more than four orders of magnitude for Earth-sized planets around Sunlike stars at 1 AU. Observing that the three classes are separated by approximately ten orders of magnitude, Carl Sagan is often credited with the following classification, although this equation does not seem to appear explicitly anywhere in Sagan (1973):

$$\mathcal{K}_{\mathrm{ETI}} = 1 + \frac{1}{10} \log_{10} \left(\frac{W_{\mathrm{ETI}}}{10^{16}\ \mathrm{W}}\right), \tag{9.37}$$

with $\mathcal{K}_{\mathrm{ETI}}$ representing the non-integer Kardashev type. As per this scheme, humanity falls under the category of Type 0.7 because $W_{\mathrm{ETI}} \sim 10^{13}$ W.

Other classification schemes unrelated to the Kardashev Scale also exist. For example, John Barrow proposed that ETIs should be categorized based on their ability to manipulate and control matter at different length scales. In

this schema, increasing levels of technological sophistication are interpreted in terms of precise matter manipulation at microscopic scales. However, one of the potential issues with this otherwise intriguing approach is that it yields little direct information about the nature of technosignatures manifested by ETIs of a given rank. Ivanov et al. (2020) raised the pertinent point that the complexity of a given ETI is not inescapably correlated with its energy expenditure and propounded a categorization whereby the class of an ETI is determined by the degree to which it modifies the ambient environment and blends in with the latter. The reader may consult Ćirković (2015) and R. H. Gray (2020) for in-depth overviews of the exact nature and benefits of the Kardashev Scale and its extensions.

9.4.2 Stapledon-Dyson spheres

In 1937, Olaf Stapledon authored *Star Maker,* an acclaimed science-fiction novel that drew on the philosophy of Baruch Spinoza and anticipated the likes of genetic engineering, the Zoo hypothesis pertaining to Fermi's paradox (Section 8.2.2.3), interstellar travel and directed panspermia (Chapter 10), and megastructures, i.e., artificial objects comparable to the sizes of planets and moons at the minimum. In particular, the following prescient lines from Stapledon (1968, p. 365, 380) are worth highlighting:

> As the aeons advanced, hundreds of thousands of worldlets were constructed, all of this type, but gradually increasing in size and complexity. Many a star without natural planets came to be surrounded by concentric rings of artificial worlds. In some cases the inner rings contained scores, the outer rings thousands of globes adapted to life at some particular distance from the sun … it began to avail itself of the energies of its stars upon a scale hitherto unimagined. Not only was every solar system now surrounded by a gauze of light traps, which focused the escaping solar energy for intelligent use, so that the whole galaxy was dimmed, but many stars that were not suited to be suns were disintegrated, and rifled of their prodigious stores of sub-atomic energy.

Stapledon may have, in turn, been influenced by J. D. Bernal, who wrote about spherical space colonies in his visionary, but sadly forgotten, treatise *The World, the Flesh and the Devil* (1929). Stapledon's qualitative analysis was subsequently formalized by Freeman Dyson in his seminal paper (1960), owing to which we shall dub these spherical habitats

Stapledon-Dyson spheres. We will briefly describe some salient characteristics of Stapledon-Dyson spheres, and we defer to J. T. Wright (2020) for a meticulous exposition.

For the sake of simplicity, we suppose that the star is a perfect blackbody with an effective temperature of $T_\star$ and a radius of $R_\star$. Now, we consider a Type II ETI on the Kardashev Scale that harnesses all of the energy output from this star. We will assume that this energy is utilized to maintain a spherical habitat at a distance a from the star with an effective temperature of T_w. If the habitat also functions as a blackbody, we have

$$L_\star = 4\pi\sigma R_\star^2 T_\star^4 = 4\pi\sigma a^2 T_w^4. \tag{9.38}$$

This equation is easily solvable for T_w, thus yielding

$$T_w \approx 395\ \mathrm{K} \left(\frac{T_\star}{T_\odot}\right)\left(\frac{R_\star}{R_\odot}\right)^{1/2}\left(\frac{a}{1\ \mathrm{AU}}\right)^{-1/2}, \tag{9.39}$$

wherein $R_\odot$ is the solar radius and $T_\odot \approx 5780$ K denotes the effective temperature of the Sun. Hence, for a Sun-like star, choosing $a \approx 2$ AU would yield a habitable temperature of $T_w \approx 280$ K.

The reader may rightly accuse the preceding estimate of being overly anthropocentric. In order to appreciate why the obtained value of T_w could be reasonable, it is worth recalling that the Carnot efficiency ($\mathcal{W}_C$), which often serves as the maximum thermodynamic efficiency attainable, is expressible as

$$\mathcal{W}_C = 1 - \frac{T_w}{T_\star} \approx 1 - 6.8 \times 10^{-2}\left(\frac{R_\star}{R_\odot}\right)^{1/2}\left(\frac{a}{1\ \mathrm{AU}}\right)^{-1/2}. \tag{9.40}$$

Strictly speaking, a more realistic estimate is given by (4.76), but the gist of our ensuing discussion is still preserved. As seen from the above expression, increasing a will improve the Carnot efficiency, but this result needs further unpacking. For the two cases with $a = 1$ AU and $a = 10^2$ AU, we find that $\mathcal{W}_C \approx 0.932$ and $\mathcal{W}_C \approx 0.993$, respectively. Hence, increasing the size of the Stapledon-Dyson sphere by a factor of ~ 100 contributes to a fairly marginal increase in the Carnot efficiency. For a spherical shell, if we assume that the cost of construction scales with the area, we see that this modest enhancement in the Carnot efficiency entails an economic cost that is $\sim 10^4$ times higher.

A couple of other interesting properties can be worked out in connection with Stapledon-Dyson spheres. First, it must be recognized that rigid Stapledon-Dyson spheres (viz., continuous spherical shells) are not stable; at best they are marginally stable. To understand why, let us recall a well-known result from electrostatics: the electric field is zero in the interior of a spherical shell. Hence, any point mass situated within the shell would not experience any net attraction. The same line of reasoning is applicable to gravity because of its inverse-square law behavior for a point mass (i.e., the star) situated within a spherical shell. Second, the material strength of rigid Stapledon-Dyson spheres must be unrealistically high in order to prevent deformation. A lower bound on the tensile strength (Π) is obtained by equating the pressures attributable to elastic stress and gravity, thereby yielding

$$\Pi = \frac{GM_\star \rho_{\rm shell}}{2a} \approx 2.5 \times 10^{12}\,{\rm N\,m^{-2}} \left(\frac{M_\star}{M_\odot}\right)\left(\frac{\rho_{\rm shell}}{\rho_\oplus}\right)\left(\frac{a}{1\,{\rm AU}}\right)^{-1}, \tag{9.41}$$

where $\rho_{\rm shell}$ and $\rho_\oplus$ stand for the mass density of the spherical shell and the Earth, respectively. The derived tensile strength is orders of magnitude higher than those of most known materials (e.g., graphene has a tensile strength of $\sim 1.3 \times 10^{11}$ N m^{-2}). In addition, there is a second material constraint imposed by the necessity of avoiding buckling that leads to even more unrealistic requirements for the elastic modulus, which we shall not outline here; the reader is referred to J. T. Wright (2020) instead.

Hence, realistic proposals for Stapledon-Dyson spheres are based on the premise that the "sphere" actually comprises smaller and closely packed structures, analogous to how floors or roofs are made up of tiles. These structures may be placed independently around the host star in Keplerian orbits. Alternatively, an interesting proposal advanced by Robert Forward is that each unit can function as a solar sail and maintain a fixed position without orbiting the star. These "statites" (to use Forward's term) are in equilibrium due to the balance between the radiation pressure and the gravitational force,[5]

$$\frac{L_\star A_{\rm sat}}{2\pi a^2 c} = \frac{GM_\star m_{\rm sat}}{a^2}, \tag{9.42}$$

5. Strictly speaking, the general relativistic corrections to Newtonian gravity could grow over time and make a difference since the former does not precisely scale as $1/a^2$.

where $m_{\rm sat}$ and $A_{\rm sat}$ are the mass and area of the statite, respectively; we have also assumed that the statite functions as a perfect reflector. Note that the factor of a^2 cancels out on both sides, implying that this equality holds true at all distances from the star. Defining $\sigma_{\rm sat} = m_{\rm sat}/A_{\rm sat}$ and solving for it, we obtain

$$\sigma_{\rm sat} = \frac{L_\star}{2\pi c G M_\star} \approx 1.5 \times 10^{-3}\,{\rm kg\,m^{-2}} \left(\frac{L_\star}{L_\odot}\right)\left(\frac{M_\star}{M_\odot}\right), \tag{9.43}$$

which is further simplified after invoking the mass-luminosity relationship $L_\star \propto M_\star^3$. If we suppose that the mass density of the object is $\sim 2 \times 10^3$ kg m^{-3}, we see that the thickness of this sail will need to be $\sim 1\ \mu$m. It is also possible to design and deploy dynamic versions of statites, referred to as "quasites" by Kipping (2019), in an analogous fashion. The non-Keplerian motion of these quasites would result in anomalously long transits, thereby rendering them potentially detectable, provided that their cumulative effective area is large enough (Kipping 2019).

The blackbody emission peak of the Stapledon-Dyson sphere is easy to compute by combining Wien's law and (9.39). The corresponding wavelength ($\lambda_{\rm max}$) will be given by

$$\lambda_{\rm max} \approx 7.3\,\mu{\rm m} \left(\frac{T_\star}{T_\odot}\right)^{-1} \left(\frac{R_\star}{R_\odot}\right)^{-1/2} \left(\frac{a}{1\,{\rm AU}}\right)^{1/2}, \tag{9.44}$$

implying that most of the radiation occurs in the infrared along expected lines (Dyson 1960; Sagan & Walker 1966). The associated flux density at the emission peak is determined to be

$$S_{\rm max} = \frac{L_\star \lambda_{\rm max}}{4\pi c d_\star^2} \approx 7.8 \times 10^{-2}\,{\rm Jy} \left(\frac{L_\star}{L_\odot}\right)^{3/4} \left(\frac{a}{1\,{\rm AU}}\right)^{1/2} \left(\frac{d_\star}{1\,{\rm kpc}}\right)^{-2}. \tag{9.45}$$

Thus, for fixed $d_\star$, it is apparent that $S_{\rm max}$ depends only on the stellar luminosity and the distance between the Stapledon-Dyson sphere and its host star.

9.4.3 Other megastructures

Apart from Stapledon-Dyson spheres, a number of other megastructures have been investigated. One of the most famous among them is the ring world that formed the eponymous subject of the science-fiction novel

Ringworld (1970) by Larry Niven. This megastructure comprises a giant rotating ring with a finite thickness Δr at a distance a from the star; the purpose of the rotation is to endow the habitat with sufficient centrifugal acceleration to mimic the Earth's gravity. The requisite angular velocity $\Omega_{\rm ring}$ is found from $\Omega_{\rm ring}^2 a \approx g_\oplus$, which yields

$$\Omega_{\rm ring} \approx 8.1 \times 10^{-6}\,{\rm rad/s}\,\left(\frac{a}{1\,{\rm AU}}\right)^{-1/2}. \tag{9.46}$$

Ring worlds suffer from certain issues that also beset rigid Stapledon-Dyson spheres, such as the necessity for very high tensile strength. Moreover, ring worlds are unstable to radial perturbations, as seen from the following heuristic argument. Suppose that the ring world has been slightly perturbed from the center, with the distance to the closer edge being denoted by r'. Gravity falls off as $1/r'^2$ whereas the "area" of the mass element (akin to a circular sector) scales as r'. Hence, these two effects do not cancel each other out, as they would for a spherical shell, thereby leading to further drift.

Another well-known class of megastructures are stellar engines, which harness the energy of the star to generate thrust. The most famous among them is the Shkadov thruster, which was ostensibly first proposed by Leonid Shkadov and consequently bears his name (see Shkadov 1987). It is worth appreciating, however, that the history of stellar engines goes back to the mid-twentieth century at the minimum (Lingam & Loeb 2020i). Shkadov thrusters are also known as Class A stellar engines in the classification scheme outlined by Badescu and Cathcart (2000), and their calculations are briefly described henceforth. In qualitative terms, the Shkadov thruster can be envisioned as an incomplete Dyson sphere (i.e., a spherical sector) that essentially functions as a mirror and reflects the starlight toward the star. Due to a certain fraction of the star's radiation pressure being reflected back to it, a net thrust is exerted on the star because of the induced force asymmetry. The star is therefore incapable of radiating away energy from the region covered by the Shkadov thruster (i.e., the mirror) as the emitted energy is returned back to the star. Hence, the same power must be dissipated across a smaller area on the surface of the star. We assume that the thruster is situated at a distance a from the star, and the angle subtended by the mirror at the center of the star is denoted by Ψ.

Note that the total surface area of the star is $4\pi R_\star^2$, whereas the area of the star covered by the thruster is $2\pi R_\star^2 \int_0^\Psi \sin\theta'\, d\theta' = 2\pi R_\star^2 (1 - \cos\Psi)$. Hence, the effective area of the star that is not covered by the thruster is given

by $\tilde{A}_\star = 2\pi R_\star^2 (1 + \cos\Psi)$. As noted above, the stellar luminosity must be conserved, which leads us to

$$L_\star = 4\pi\sigma R_\star^2 T_\star^4 = 2\pi\sigma R_\star^2 (1 + \cos\Psi)\ \tilde{T}_\star^4, \tag{9.47}$$

where $\tilde{T}_\star$ is the new temperature of the star after the introduction of the Shkadov thruster. The force (impulse per unit time) is given by dp/dt, and for photons, we have $p = E/c$ from the energy-momentum law. Using these relations, we find that the impulse per unit time of the radiation exiting the complete surface of the star is

$$\frac{dp}{dt} = \frac{4\pi\sigma R_\star^2 \tilde{T}_\star^4}{c}, \tag{9.48}$$

and after simplifying this expression using the above equations, we arrive at

$$\frac{dp}{dt} = \frac{L_\star}{c}\frac{2}{1 + \cos\Psi}. \tag{9.49}$$

The radiation pressure $P_{\rm rad}$ in the direction normal to the base of the spherical cone is found by dividing the force by the area $4\pi a^2$ to obtain

$$P_{\rm rad} = \frac{L_\star}{4\pi c a^2}\frac{2}{1 + \cos\Psi}. \tag{9.50}$$

The thrust $\mathcal{F}_{\rm net}$ that is generated by the star-mirror system due to the asymmetric radiation field is given by

$$\mathcal{F}_{\rm net} = \int P_{\rm rad}\, dA_{sc} = \pi a^2 P_{\rm rad} \sin^2\Psi = \frac{L_\star}{2c}(1 - \cos\Psi), \tag{9.51}$$

where A_{sc} represents the base area of the spherical cone. The acceleration experienced by the star ($\bar{a}_{\rm net}$) is expressible as

$$\bar{a}_{\rm net} \approx 3.2 \times 10^{-13}\ {\rm m\,s^{-2}} \left(\frac{L_\star}{L_\odot}\right)\left(\frac{M_\star}{M_\odot}\right)^{-1}(1 - \cos\Psi). \tag{9.52}$$

Although the thrust provided by Shkadov thrusters is apparently very small, it may nevertheless be useful for causing deviations of host stars from their intrinsic orbits in the event that astrophysical or artificial phenomena in their

vicinity pose existential threats to the survival of ETIs.[6] Aside from Shkadov thrusters, more sophisticated configurations for stellar engines have been expounded in the literature, some of which might be capable of attaining sub-relativistic speeds of ~ 0.01–$0.1\ c$ (Lingam & Loeb 2020i).

Apart from Stapledon-Dyson spheres, Shkadov thrusters, and ring worlds along with their variants, a number of other designs have been propounded for large space-based habitats sensu lato (Birch 1991; Badescu et al. 2006), such as shell worlds (Roy et al. 2013). Enthusiastic, albeit somewhat dated, treatments of this subject can be found in O'Neill (1977) and Rood and Trefil (1981), the latter of which is primarily devoted to an examination of Fermi's paradox and the prevalence of ETIs. Lastly, it is possible to envision a series of Stapledon-Dyson spheres situated in concentric orbits, such that the waste heat of the i-th sphere serves as the power source for the $(i+1)$-th sphere to carry out computations, and so on. This hypothetical megastructure was christened the *Matrioshka brain* by futurist Robert Bradbury in 1999,[7] although some aspects were elucidated by Stapledon in 1937—as evidenced by the excerpt in Section 9.4.2—and by Suffern (1977).

9.4.4 The AGENT framework

Wright, Griffith, et al. (2014) present a formalism dubbed AGENT, after the names of its parameters, to discuss the energy budget of ETIs. For the sake of simplicity, we will concern ourselves with an ETI that is restricted to a single host star, but this is easily generalizable to encompass Type III ETIs that span an entire galaxy; the chief difference is that the stellar parameters must be replaced by their galactic counterparts.

The four parameters of relevance are as follows:

- α_{ETI} is the amount of starlight that is harnessed by the ETI.
- $\varepsilon_{\mathrm{ETI}}$ is the amount of energy that is generated through alternative channels. This may include exotic options such as using rotating

6. The thrust could be significantly enhanced if one engineered a rocket effect via expulsion of material from the surface of the megastructure (Caplan 2019). The momentum gain might increase in select cases by a factor of $\sim c/v_{\mathrm{ex}}$, where v_{ex} is the exhaust speed of the evaporated gas.

7. See https://www.gwern.net/docs/ai/1999-bradbury-matrioshkabrains.pdf

black holes to extract energy or relatively mundane routes like fossil fuels and geothermal power.
- γ_{ETI} is the amount of radiation emitted in the form of waste heat at high entropy, with an associated temperature of T_w.
- ν_{ETI} is the amount of nonthermal energy that is emitted; an obvious example is the transmission of coherent electromagnetic signals.

All of these variables are dimensionless since they are expressed in units of stellar luminosity ($L_\star$). By the conservation of energy, we obtain

$$\alpha_{\mathrm{ETI}} + \varepsilon_{\mathrm{ETI}} = \gamma_{\mathrm{ETI}} + \nu_{\mathrm{ETI}}. \tag{9.53}$$

For ETIs that are primarily dependent on the host star for energy, it follows that $\varepsilon_{\mathrm{ETI}} \ll \alpha_{\mathrm{ETI}}$. Similarly, if most of the output from ETIs is in the form of waste heat—as opposed to transmitting electromagnetic signals, for instance—we see that $\nu_{\mathrm{ETI}} \ll \gamma_{\mathrm{ETI}}$. For a Stapledon-Dyson sphere, both of these orderings are valid; in fact, one has $\gamma_{\mathrm{ETI}} = \alpha_{\mathrm{ETI}} = 1$ and $\varepsilon_{\mathrm{ETI}} = \nu_{\mathrm{ETI}} = 0$. Sagan's version of the Kardashev Scale (9.37) can be reexpressed as

$$\mathcal{K}_{\mathrm{ETI}} = 2.06 + 0.1 \log_{10} \left(\alpha_{\mathrm{ETI}} + \varepsilon_{\mathrm{ETI}} \right). \tag{9.54}$$

In addition to the conservation of energy, if the thermodynamic efficiency is specified by the Carnot efficiency (9.40), a constraint on ν_{ETI}, the fraction of emitted energy that is not waste heat, is obtained as follows:

$$\frac{\nu_{\mathrm{ETI}}}{\alpha_{\mathrm{ETI}} + \varepsilon_{\mathrm{ETI}}} \lesssim \mathcal{W}_C. \tag{9.55}$$

Lastly, it is possible to employ the AGENT formalism to examine the flux density S_{sig} (in wavelength units) received at Earth. In the absence of any ETI, we obtain

$$\lambda S_{\mathrm{sig}} = \bar{f}_{\mathrm{star}} \frac{L_\star}{4\pi d_\star^2}, \tag{9.56}$$

where $\bar{f}_{\mathrm{star}}$ signifies the normalized spectral energy distribution. For an idealized blackbody at the stellar temperature $T_\star$, it is known that $\bar{f}_{\mathrm{star}} = \pi \lambda B_\lambda \left(T_\star \right) / \left(\sigma T_\star^4 \right)$, with B_λ representing the Planck function defined in

(4.59). In the presence of an ETI, only a fraction $(1-\alpha_{\rm ETI})$ of the starlight reaches the Earth. In addition, a fraction $\gamma_{\rm ETI}$ is emitted as waste heat that is modeled as a blackbody at temperature T_w, and there also exists the nonthermal contribution $\nu_{\rm ETI}$, which we neglect without much loss of generality. Thus, the revised version of (9.56) has the form

$$\lambda S_{\rm sig} = \left[(1-\alpha_{\rm ETI})\,\bar{f}_{\rm star} + \gamma_{\rm ETI}\frac{\pi\lambda B_\lambda\,(T_w)}{\sigma\,T_w^4} \right] \frac{L_\star}{4\pi\,d_\star^2}. \tag{9.57}$$

We will now use the blackbody assumption for $\bar{f}_{\rm star}$ delineated earlier, which yields

$$S_{\rm sig} \approx \left[(1-\alpha_{\rm ETI})\,B_\lambda\,(T_\star) + \gamma_{\rm ETI} B_\lambda\,(T_w) \left(\frac{T_\star}{T_w}\right)^4 \right] \frac{L_\star}{4\sigma\,d_\star^2\,T_\star^4}. \tag{9.58}$$

As the resultant expression is still rather complex, we will make use of the Rayleigh-Jeans law. In this regime, $B_\lambda(T) \propto T/\lambda^4$, and we obtain

$$S_{\rm sig} \sim \left[1 - \alpha_{\rm ETI} + \gamma_{\rm ETI} \left(\frac{T_\star}{T_w}\right)^3 \right] \frac{k_B c L_\star}{2\sigma\lambda^4 d_\star^2 T_\star^3}. \tag{9.59}$$

Let us introduce the notation $S_{\rm null} = S_{\rm sig}$, evaluated at $\alpha_{\rm ETI} = \gamma_{\rm ETI} = 0$ (in the absence of an ETI), and $\Delta S_{\rm sig} = S_{\rm sig} - S_{\rm null}$. Using the above expression, we see that

$$\frac{\Delta S_{\rm sig}}{S_{\rm null}} \sim \gamma_{\rm ETI} \left(\frac{T_\star}{T_w}\right)^3 - \alpha_{\rm ETI}. \tag{9.60}$$

Due to the presence of the cubic term involving the ratio $T_\star/T_w$, it is apparent that the left-hand side of (9.60) will be subjected to considerable amplification. To illustrate our point, let us select $T_\star \approx 5800$ K, $T_w \approx 300$ K, and a modest value of 0.15 for $\gamma_{\rm ETI}$. For this choice, we end up with $\Delta S_{\rm sig}/S_{\rm null} \sim 1.1 \times 10^3$, which amounts to a substantial enhancement in the flux density.

As a result, we have seen that there is a clear advantage gained from searching for waste heat from ETIs in the infrared. On the other hand, we caution that natural astrophysical sources, most notably dust, also give rise to significant infrared excesses. Hence, due care should be taken to distinguish

between such false positives and waste heat emitted by putative ETIs. We will return to the issue of detecting ETIs via their waste heat later in this chapter.

9.5 ARTIFACT TECHNOSIGNATURES

Electromagnetic signals present themselves as obvious candidates for technosignatures because of their speed of transmission, their comparatively straightforward generation, and the purportedly facile recognition of their artificial nature if they are detected. Yet, it should be simultaneously recognized that electromagnetic technosignatures have inherent disadvantages. Due to the fact that the parameter space for electromagnetic signals is nine-dimensional in nature, it is quite conceivable that we may spend decades, even centuries, searching for them and end up with null results even if ETIs do exist. Furthermore, although humans find the use of electromagnetic signals for communication purposes natural, it does not automatically follow that all ETIs will think along the same lines.

Hence, to improve our prospects of detecting signatures of ETIs, other approaches must be pursued concomitantly. One of the central themes that unifies both nontechnological and technological life is niche construction, as noted at the beginning of the chapter. Technology modifies the natural environment in numerous ways, thereby producing a diverse array of potential technosignatures. Therefore, from a theoretical standpoint, the identification of viable technosignatures is mainly a function of our capacity to rigorously envision futuristic technologies in parallel with the *physical* traces that they induce in their environments (Kardashev 1985).

Owing to the emphasis on technological constructs and their physical markers, this burgeoning field has come to be known as artifact SETI or Dysonian SETI. The resultant artifacts are not necessarily correlated with *extant* ETIs; in fact, such markers could persist long after their builders have become extinct. As a consequence, the study of artifacts from extinct ETIs has been dubbed "interstellar archaeology" (Freeman & Lampton 1975; Carrigan 2012) or "astro-palaeontology" (J. Armitage 1977) in some quarters. Summaries and arguments in favor of artifact SETI are presented in Bradbury et al. (2011), J. T. Wright, Mullan, et al. (2014), and Lingam and Loeb (2019e).

Two chief advantages are inherently attributable to artifact SETI. First, as we shall see, the range of potential technosignatures is very diverse and

spans physical markers from planetary surfaces, atmospheres, and space. This diversity may increase the likelihood of stumbling on an artifact, as opposed to conventional SETI, which places all of its eggs in a single basket (electromagnetic signals). Second, in most instances involving planetary technosignatures, searches for artifacts can be conducted in tandem with searches for biosignatures, thereby obviating the necessity for separate telescopes and projects.

Owing to the multitude of prospective artifacts that have sprung up in recent times, we will restrict ourselves to only a few select examples below.

9.5.1 Detection of atmospheric pollutants

It is self-evident that the concentrations of certain gases in Earth's atmosphere have evolved rapidly as a result of anthropogenic activity. Setting aside the ongoing extensive debates concerning the optimal levels of carbon dioxide and the unsettling implications for our biosphere, it is worth recognizing that some molecules in our atmosphere are generated solely due to human technology. Hence, it is natural to posit that finding evidence of these gases could signify the existence of either active or extinct ETIs (J. B. Campbell 2006; Schneider et al. 2010).

Perhaps the most notable among the anthropogenic gases are the chlorofluorocarbons (CFCs). Due to their nontoxic and nonflammable nature allied to low reactivity, they have been used in a wide range of industrial applications ranging from aerosol sprays and refrigerants to the manufacture of foams and packing materials. CFCs are known to cause ozone depletion via the following schematic reaction network,

$$\begin{aligned} &\mathrm{CFCl_3} + h\nu \rightarrow \mathrm{CFCl_2} + \mathrm{Cl}, \\ &\mathrm{Cl} + \mathrm{O_3} \rightarrow \mathrm{ClO} + \mathrm{O_2}, \end{aligned} \tag{9.61}$$

where $h\nu$ refers to a UV photon with the requisite energy, and the CFC is represented by $\mathrm{CFCl_3}$. The lifetimes of CFCs range from ~ 10 to $\sim 10^5$ yr. Hence, if a short-lived CFC was detected, it may be indicative of an active ETI, while the detection of a long-lived CFC would imply that an active ETI existed at least within the past $\sim 10^5$ yr.

A number of methods have already been explored extensively for detecting gaseous biosignatures, of which one of the most prominent is

transmission spectroscopy; see Section 6.1 for additional details. Lin et al. (2014) analyzed the prospects of detecting CFCs at concentrations that are ~ 10 times higher than the modern levels in Earth's atmosphere. Assuming the current rate of CFC production can be extrapolated, this concentration would occur on Earth in $\sim 10^3$ yr. The two CFCs considered in the study were CFC-14 and CFC-11, with strong absorption features at 7.8 μm and 11.7 μm, respectively. The detection of these features requires a spectral resolution of $\mathbb{R} \sim 100$, which corresponds to achievable medium-resolution spectroscopy.

Lin et al. (2014) determined that detecting CFCs at the above concentration in the atmosphere of an exoplanet transiting a white dwarf at a distance of ~ 40 pc required a total integration time of ~ 1 day for JWST to achieve a SNR of ~ 5. It is possible to undertake a heuristic calculation for an M-dwarf exoplanet situated at a given distance by using the ensuing scaling relation after applying the Rayleigh-Jeans limit to (6.16):

$$\mathrm{SNR} \propto d_\star^{-1} R_\star^{-1} \Delta t^{1/2} T_\star^{1/2} \tag{9.62}$$

As the SNR is held fixed, we end up with $\Delta t \propto R_\star^2 d_\star^2 / T_\star$. For illustration, we adopt $\Delta t \sim 1$ day, $T_\star \sim 6 \times 10^3$ K, $d_\star \sim 40$ pc, and $R_\star \sim 0.01\, R_\odot$ for the white dwarf, whereas the corresponding parameters for the M-dwarf are $T_\star \sim 2.5 \times 10^3$ K, $d_\star \sim 10$ pc, and $R_\star \sim 0.1\, R_\odot$. Thus, we find that the integration time required for detecting CFCs on an M-dwarf exoplanet is ~ 15 days; if we lower the desired SNR to 3, the integration time drops to ~ 5 days.

Another notable candidate for altering atmospheric chemistry is nuclear warfare; some of the prominent effects are delineated in Stevens et al. (2016). First, the emission of beta particles into the atmosphere would cause significant ionization of the chemical species. In particular, if one supposes that the primary molecules are nitrogen and oxygen, a blue-green airglow akin to an aurora is produced. The most distinctive airglow feature is the O I line at 557.7 nm. It has been predicted that global nuclear war on Earth would yield a photon flux of $\sim 10^{13}$ photons m^{-2}. In Section 6.3.1, we noted that the impact of the stellar wind on the nearby exoplanet Proxima b also yields strong auroral emission; theoretical models indicate that the corresponding photon flux would be $\sim 10^{14}$ photons m^{-2}. Hence, in the event that the inventory of nuclear weapons utilized by ETIs in warfare is a few orders of magnitude higher than what is currently available to human beings, it is

conceivable that the resultant O I emission may be detectable after several nights of integration time for nearby exoplanets.

Second, new species such as nitrogen oxides (NO_x) and hydrogen oxides (HO_x) are produced as a consequence of nuclear warfare. If the entire stockpile of human weapons is deployed, the total quantity of NO_x thus generated is $\sim 3 \times 10^{13}$ kg. As a result, the mixing ratio of NO_x might reach a maximum of $\sim 10^{-5}$ in the immediate aftermath of a nuclear war. The high concentration of nitrogen oxides is expected to have detectable consequences. First, time-varying emission of infrared photons occurs at 5.3 μm for nitric oxide (NO); this band is created due to inelastic collisions between NO and oxygen that excite a number of vibrational-rotational transitions. During the impact of extreme stellar proton events (SPEs) arising from superflares (see Section 4.4.3) with Earthlike planets around G- and K-type stars, it is estimated that the mixing ratio of NO reaches a maximum of $\sim 2 \times 10^{-6}$, and the corresponding power emitted at 5.3 μm is $\sim 10^{15}$ W (Airapetian et al. 2017). The transient enhancement in NO concentration is ~ 2 to 3 orders of magnitude during these events. It is reasonable to contend that the dissipated power at 5.3 μm can attain $\gtrsim 10^{15}$ W at the peak of nuclear war because the associated maximal mixing ratio of NO may be on the order of $\sim 10^{-6}$, as noted earlier.

The formation of nitrogen and hydrogen oxides has two other crucial consequences. First, the production of the OH radical leads to the depletion of NO_x and the formation of nitric acid (HNO_3). A simplified representation of this reaction is

$$NO_2 + OH + \mathcal{X}_s \rightarrow HNO_3 + \mathcal{X}_s, \tag{9.63}$$

where $\mathcal{X}_s$ is a chemical species that does not undergo any net alteration in chemical composition but plays a catalytic role of sorts. Detailed numerical models suggest that the production of HNO_3 gives rise to a strong absorption feature at 11.2 μm that may be potentially detectable by future telescopes. In the case of exoplanets around M-dwarfs, the planet-to-star contrast ratio might change by a factor of $\sim 10^{-6}$ at this wavelength due to HNO_3 generation (Tabataba-Vakili et al. 2016). Although these calculations were undertaken for SPEs, similar principles are expected to come into play after global nuclear catastrophes. Lastly, the formation of nitrogen oxides is predicted to stimulate ozone depletion, as explained in Section 4.4.3. In the case of large SPEs and global nuclear war, the ozone depletion could

exceed ~ 50 percent. The transient depletion of the ozone layer would ostensibly reduce the amplitude of ozone absorption features, most notably the sensitive Hartley-Huggins bands at $0.25\ \mu$m.

Thus, viewed in totality, there are a number of interesting avenues for discerning the effects of nuclear catastrophes. Yet, two important caveats should be borne in mind. First, not only SPEs but also high volcanic activity and impact events by large (kilometer-sized or greater) objects are capable of mimicking most of the above features. Distinguishing SPEs is comparatively easier since they are either directly or indirectly linked with stellar flares, and continually monitoring changes in the light curve of the host star would reveal the existence of flares. Identifying bolide impacts or frequent volcanism is much harder as both global nuclear war and these two natural phenomena generate substantial amounts of dust and aerosols, thereby increasing the planet's opacity by at least one order of magnitude. This brings us to the second caveat: pinpointing the aforementioned spectral features may prove to be very difficult because of the increased atmospheric opacity. What is required, instead, is a series of serendipitous observations (e.g., afterglow, NO and HNO_3 spectral signatures) in specific temporal order to demarcate nuclear catastrophes from natural processes.

It is not merely the planet that might host industrial pollutants or other signatures of ETI. In their classic monograph, Shklovskii and Sagan (1966) discussed the possibility that advanced ETIs may opt to deliberately seed the photospheres of their host stars with rare elements and isotopes to serve as markers of technological activity. Alternatively, megaengineering projects that rely on chemical homogenization and regulation of stellar mass loss were posited as avenues for prolonging the main-sequence lifetime by Martin Beech. In this analysis, Beech (1990) also suggested that stars engineered in such a fashion would exhibit characteristics akin to blue stragglers, which are rare main-sequence stars found in clusters that exhibit main-sequence lifetimes longer than typical stars of their spectral type.

Whitmire and Wright (1980) conjectured that ETIs indulging in large-scale nuclear fission could opt to use stars as repositories of their radioactive waste products. If the host stars have shallow convection zones (valid for $M_\star > 1.5\,M_\odot$) and ~ 1 percent of Earth's uranium reserves have been consumed and dumped into the host stars, such waste might be detectable. In particular, it was suggested that the elements praseodymium and neodymium would occur at anomalously high concentrations. The star HD 101065, christened Przybylski's Star after Antoni Przybylski, who discovered it in 1961, comprises unusually high abundances of not only

praseodymium and neodymium but also other lanthanides as well as short-lived actinides (e.g., einsteinium). Yet, HD 101065 has not attracted any serious theoretical or observational scrutiny hitherto as an intriguing target for SETI (J. T. Wright 2018a).

Lastly, apart from stars and planets, technologically advanced ETIs could decide to mine asteroids for raw materials (Kecskes 1998; Forgan & Elvis 2011) or perhaps even deliberately pollute them. Large-scale mining will give rise to chemical disequilibrium due to the extraction and depletion of certain species, mechanical disequilibrium from disrupting and destroying asteroids, and thermal disequilibrium as a result of dust generated at anomalous temperatures during mining. However, while the detectability of signatures originating from asteroid mining remains very uncertain, searching for them does not incur any additional cost because these searches can piggyback onto existing studies of debris disks.

9.5.2 Leakage of energy from propulsion systems

If the drive to settle other worlds is pronounced in a subset of ETIs, the necessity for developing interplanetary and interstellar propulsion systems is a must. We shall not tackle this topic in detail here, deferring it instead to Chapter 10. For now, it suffices to note that two major disadvantages with conventional chemical rockets are (1) they are relatively slow and (2) the fuel source must be transported concurrently with the actual payload, thereby increasing the total mass of the system by orders of magnitude. Among alternatives for spacecraft propulsion at high (even weakly relativistic) speeds, light sails have emerged as attractive candidates because they are capable of surmounting the aforementioned issues. Antimatter propulsion is another technology that can attain relativistic speeds in theory, although its efficacy is yet to be proven. The chief challenges in employing antimatter as fuel have to do with its large-scale production and long-term storage.

9.5.2.1 Light sails

The basic premise underlying light sails is simple: they are spacecraft propelled by radiation pressure. The source of this radiation pressure could be either starlight, as seen from our brief analysis of the Shkadov thruster in Section 9.4.3, or beamed energy by ETIs. If ETIs were to rely on power beaming to accelerate spacecraft, this activity ought to result in the leakage of electromagnetic radiation. In some respects, the leakage would

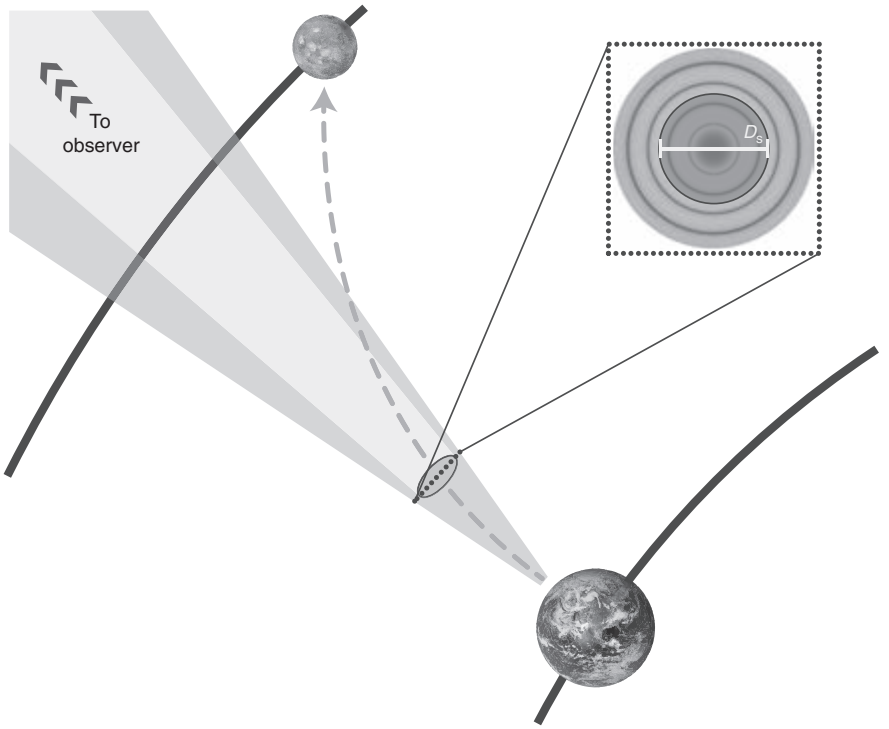


Figure 9.3 Light sail propulsion for interplanetary travel between Earth and Mars. The trajectory of the light sail is illustrated by the dashed curve and the shaded area represents the beam profile. The logarithm of the beam profile intensity is shown within the inset. (© The American Astronomical Society. *Source:* James Guillochon and Abraham Loeb [2015], "SETI via leakage from light sails in exoplanetary systems, *Astrophysical Journal* 811[2]: L20, fig. 1, left panel.)

resemble technosignatures arising from deliberate electromagnetic signals (Sections 9.1 and 9.2), but the underlying cause is very different. Leakage from power beaming is discussed as a technosignature in Guillochon and Loeb (2015a) and Benford and Benford (2016). We will henceforth adopt the notation and formalism in the former reference, which explores the scenario of using light sails for interplanetary travel. An illustration of this propulsion scheme for Earth-Mars transits along with the associated intensity profiles is presented in Figure 9.3.

When it comes to observing beamed electromagnetic radiation, it is taken for granted that the beamwidth (θ_A) is approximated by $\theta_A \approx \lambda/D_t$.

It is, however, important to appreciate that this relationship is only valid in the Fraunhofer (far-field) regime at distances $d_\star > d_F$, where $d_F = D_t^2/\lambda$ is the Fresnel length. At distances $d_\star < d_F$, the angular resolution is much larger because it is given by $\theta_F \approx (d_F/d_\star)\theta_A$. In this limit, the transmitter power (W_t) required is much higher for the same value of EIRP, as seen from (9.6). To mitigate the leakage, it seems advantageous to ensure that the acceleration phase comes to an end when the distance from the original location is equal to the Fresnel distance. For a constant acceleration $\bar{a}_{\max}$, note that the final velocity $v_{\max}$ is found from $v_{\max}^2 = 2\bar{a}_{\max} d_F$. It is therefore straightforward to invert this relationship and solve for the optimal frequency ν. By doing so, we arrive at

$$\nu \approx 68\,\mathrm{GHz} \left(\frac{v_{\max}}{100\,\mathrm{km/s}}\right)^2 \left(\frac{\bar{a}_{\max}}{g_\oplus}\right)^{-1} \left(\frac{D_t}{1.5\,\mathrm{km}}\right)^{-2}, \tag{9.64}$$

implying that the optimal frequency for light sails designed for interplanetary travel may be situated in the microwave range. The power of the beam driving the light sail is connected to the latter's mass and acceleration via $W_t = m_{\mathrm{tot}}\bar{a}_{\max}c/2$, where m_{tot} is the mass of the light sail. This relationship is derivable from (1) the rate of change of momentum equals the force, which is expressible as $2\Delta p/\Delta t = m_{\mathrm{tot}}\bar{a}_{\max}$, and (2) the momentum-energy (Δp and ΔE) relationship for photons $\Delta p = \Delta E/c$. After normalizing this expression, we end up with

$$W_t \approx 1.5 \times 10^{12}\,\mathrm{W} \left(\frac{m_{\mathrm{tot}}}{10^3\,\mathrm{kg}}\right)\left(\frac{\bar{a}_{\max}}{g_\oplus}\right). \tag{9.65}$$

The peak flux density detected by an observer is estimated by applying the formula $S_{\mathrm{sig}} \approx \epsilon_{\mathrm{leak}} W_t/\left(\pi\theta_A^2 d_\star^2 \Delta\nu\right)$ and (9.64); note that ϵ_{leak} denotes the leakage fraction of beam power. The final expression is given by

$$S_{\mathrm{sig}} \approx 5.6\,\mathrm{Jy} \left(\frac{\epsilon_{\mathrm{leak}}}{0.1}\right)\left(\frac{\Delta\nu/\nu}{0.1}\right)^{-1}\left(\frac{W_t}{10^{12}\,\mathrm{W}}\right) \times \left(\frac{v_{\max}}{100\,\mathrm{km/s}}\right)^2 \left(\frac{\bar{a}_{\max}}{g_\oplus}\right)^{-1}\left(\frac{d_\star}{100\,\mathrm{pc}}\right)^{-2}. \tag{9.66}$$

Hence, even up to distances of ~ 100 pc from Earth, it is conceivable that electromagnetic leakage associated with light sail propulsion may be

detectable. An interesting point worth mentioning with regard to the above expression is that it is independent of the aperture of the transmitter.

In theory, it is feasible to distinguish signals generated by the leakage of beamed radiation from those produced by either natural phenomena or ETIs exclusively for signaling purposes. A moving shadow is cast by the light sail on the laser beam, consequently giving rise to an intricate diffraction pattern comprising multiple peaks in the light curve as seen in Figure 9.4. As the path of the beam intersects with the observer's line of sight, a series of transient bursts, whose characteristic duration is ~ 0.1–1 s, ought to be recorded (Guillochon & Loeb 2015a). The chances of observing the leakage of electromagnetic radiation are highest during the accelerating and decelerating phases because the beam must track the spacecraft across the sky, thus resulting in the beam sweeping an angle of a few radians over a span of hours.

In our discussion heretofore, we have seen that short bursts of electromagnetic radiation at radio frequencies are potential outcomes of putative light sail propulsion systems (or signaling beacons) from the standpoint of distant observers when the laser beams sweep across their skies. In addition, photons reflected from light sails moving at relativistic speeds are presumably distinguishable from those emitted by astrophysical sources, which are often nonrelativistic, on account of their conspicuous Doppler shifts and relativistic beaming (Viewing et al. 1977; Garcia-Escartin & Chamorro-Posada 2013; Yurtsever & Wilkinson 2018).[8] It is natural to inquire whether there exist any unexplained astrophysical phenomena that are possibly compatible with any of the aforementioned properties.

Fast Radio Bursts (FRBs), discovered in 2007, seem promising in this respect; a recent review can be found in Petroff et al. (2019). The duration of FRBs is on the order of milliseconds or smaller, whereas their EIRP appears to be $\sim 10^{35}$ W, which is more than eight orders of magnitude higher than the Sun's luminosity. Two other aspects of FRBs that merit mentioning are (1) they originate from cosmological (Gpc) distances, and (2) a few of them (e.g., FRB 121102) are confirmed to repeat, but the precise status of the outwardly non-repeating majority remains indeterminate (Ravi 2019). A broad spectrum of astrophysical progenitors have been proposed for FRBs, ranging from neutron stars to white dwarfs and black holes, but none of the hypotheses are yet definitive, although active magnetars have recently gained

8. In theory, sub-relativistic spacecraft in our Solar system are also identifiable via thermal emission emitted in the course of heating due to gas and radiation (Hoang & Loeb 2020).

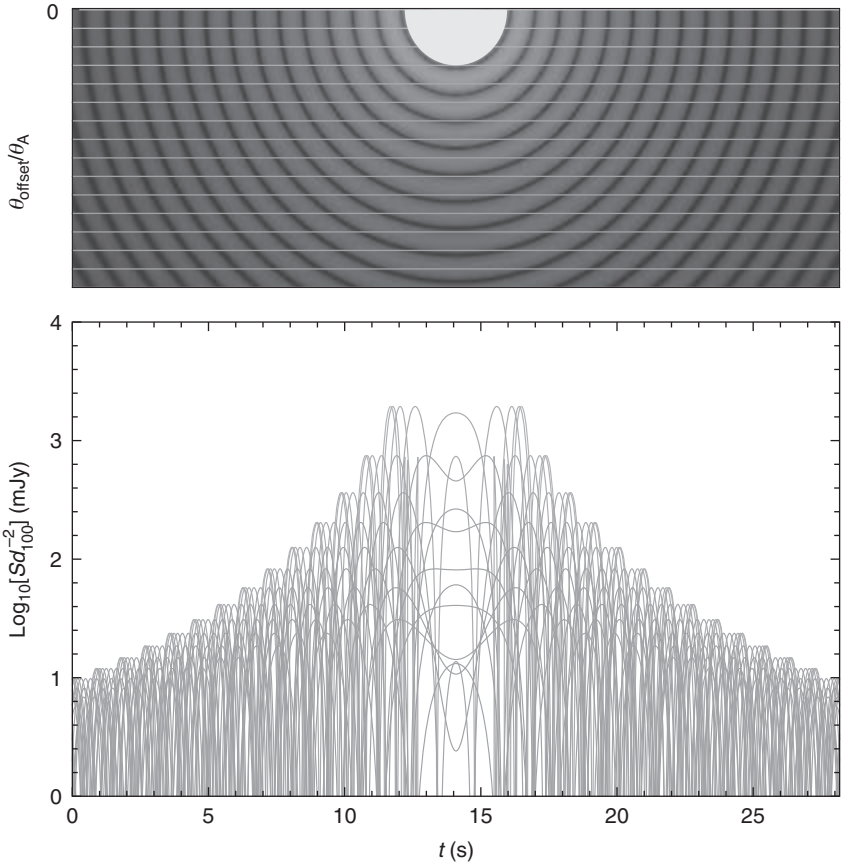


Figure 9.4 *Top:* The grayscale inherent to the diffraction pattern constitutes a proxy for the logarithm of the beam profile intensity. The horizontal chords trace the path of the beam for different choices of the offset angle θ_{offset}. Moving from top ($\theta_{\text{offset}} = 0$) to bottom ($\theta_{\text{offset}} = 10\theta_A$), the chords embody perfectly centered and grazing events, respectively. The light-shaded disk at the top represents the region that is blocked out by the light sail. *Bottom:* Flux density S received by an observer at a distance of ~ 100 pc for chords shown in the top graph as a function of time. The flux density is calculated toward the end of the acceleration phase of the light sail. (© The American Astronomical Society. *Source:* James Guillochon and Abraham Loeb [2015], "SETI via leakage from light sails in exoplanetary systems, *Astrophysical Journal* 811[2]: L20, fig. 4.)

more credibility owing to the observation of fast radio bursts in the Milky Way associated with the magnetar SGR 1935+2154.

Lingam and Loeb (2017b) examined the possibility that FRBs were a consequence of extragalactic light sails employed by ETIs. This work estimated the power of the emitter that would be required and determined

that the requisite energy was comfortably within the limits of a Type II ETI on the Kardashev Scale; in fact, it corresponds to $\mathcal{K}_{\rm ETI} \gtrsim 1.2$ after utilizing (9.37). Moreover, the optimal frequency calculated by means of (9.64) was comparable to the frequencies at which FRBs were detected. Although the study indicated that FRBs are consistent with the laws of physics and may fall within the engineering abilities of advanced ETIs (should they exist), this conjecture is indubitably very speculative; future data garnered from telescopes will undoubtedly shed more light on the precise nature of these intriguing phenomena. On a related note, B. Zhang (2020) suggested that brief radio bursts akin to FRBs, but not exactly identical to them, could be indicative of ETIs.

9.5.2.2 Antimatter propulsion

The reaction of matter with antimatter releases a massive amount of energy even if the amount of mass consumed is small. As per Einstein's famous mass-energy equivalence relation, the reaction of 1 kg of matter with 1 kg of antimatter would yield 1.8×10^{17} J if the entire rest mass is converted into energy. Hence, ever since the 1960s, a number of proposals for antimatter rockets have been mooted and investigated. Two of the best known candidates entail the annihilation of (1) protons and antiprotons and (2) electrons and positrons. The latter yield only γ-ray photons as the final product, whereas the former also produce charged and neutral pions as well as neutrinos.

In both cases, γ-ray emission is observed. Hence, by searching for the leakage of γ-rays, it may be possible to find signatures of antimatter propulsion. In order to understand how this works, let us specialize to the case of electron-positron reactions that yield photons with energies of approximately 0.5 MeV. Suppose that the mass of positrons consumed is $M_{\rm am}$, which happens over the temporal interval τ_j, which represents the burn time. In this event, the total number of photons emitted is $\sim M_{\rm am}/m_e$, implying that the average rate of photon production is $M_{\rm am}/(m_e\tau_j)$. Hence, the photon flux Φ_{phot} detected by a distant observer is

$$\Phi_{phot} \sim \frac{M_{\rm am}}{4\pi\, m_e \tau_j d_\star^2}. \tag{9.67}$$

Hence, for a given detector sensitivity, we can invert the above expression to determine an upper bound on $M_{\rm am}/\tau_j$. For $d_\star \sim 300$ pc and the threshold

sensitivity of $\Phi_{phot} \sim 0.1\ \mathrm{m}^{-2}\ \mathrm{s}^{-1}$ for the Burst and Transient Source Experiment, Harris (1986) calculated that $M_{\mathrm{am}}/\tau_j > 10^8$ kg/s was necessary in order for the γ-ray emission to be detectable. Distinguishing between γ-ray photons produced by natural astrophysical phenomena and those emitted as leakage during antimatter propulsion will require precise calibration of the proper motion (relative velocity) of the source through Doppler measurements. This diagnostic is fairly robust since very few natural objects in the Milky Way—with possible exceptions being hypervelocity objects ejected through gravitational interactions with black holes—are predicted to exhibit relativistic proper motions.

Harris (2002) generalized the above scenario to argue that (1) the annihilation of protons and antiprotons gives rise to distinctive signatures and (2) the accompanying emission of γ-ray photons from this reaction can be differentiated from natural sources by analyzing the spectrum and temporal variability. An all-sky search for antiproton annihilation signatures indicative of artificial sources was conducted using the Energetic Gamma Ray Telescope. The null result was used to obtain an upper bound of $\Phi_{phot} \sim 2.3 \times 10^{-4}\ \mathrm{m}^{-2}\ \mathrm{s}^{-1}$ on the γ-ray photon flux. Along the lines of (9.67), an upper bound of $\sim 3 \times 10^3 \left(d_\star/1\,\mathrm{pc}\right)^2$ kg/s was derived for M_{am}/τ_j. This consumption rate is about fifteen orders of magnitude higher than the current production rate of antimatter on Earth.

9.5.3 Global warming and its mitigation

Waste heat is an inevitable consequence of energy consumption as per the laws of thermodynamics, although its magnitude will vary from one technological species to another. Observations of biological activity on Earth have revealed that waste heat is not distributed uniformly; instead, it is spatially localized in the form of heat islands. This result prompted J. R. Kuhn and Berdyugina (2015) to model ETIs with an energy consumption that is lower than Type I on the Kardashev Scale. More precisely, the authors defined the quantity $\Omega_{\mathrm{ETI}} = W_{\mathrm{ETI}}/W_{\mathrm{T1}}$, where W_{T1} is given by (9.36), and considered ETIs that satisfy $4 \times 10^{-4} < \Omega_{\mathrm{ETI}} < 1$; note that the lower bound signifies the power consumed by humanity in 2013.

Satellite observations of North American cities near Michigan reveal that they have temperature excesses of ~ 10 K compared to their surroundings during summertime. J. R. Kuhn and Berdyugina (2015) demonstrated that studying temporal variations in the IR flux (contributed by reflected

light and thermal emission) at ~ 5–10 μm of exoplanets may be inverted to discern geographically clustered thermal excesses indicative of ETIs with $\Omega_{\rm ETI} \sim 0.01$. This clumping gives rise to distinctive patterns because the longitudinal distribution of waste heat is different from that of natural variations in the planet's albedo; in particular, the normalized residual variability contributed by waste heat might be on the order of ~ 1 percent. Achieving the requisite contrast ratio and inner working angle (see Section 6.2.1 for details) for detecting heat signatures of sub–Type I ETIs will, however, probably require the next generation of telescopes, such as the proposed Exo-Life Finder (ELF) Telescope (Berdyugina et al. 2018).

Although we shall return to the identification of technosignatures arising from megastructures later, one candidate is worth highlighting in this context. At some point in the evolutionary history of an exoplanet situated in the habitable zone (HZ), it will experience a runaway climate. This catastrophic change could occur either via the inexorable increase in stellar luminosity over time or due to the injection of greenhouse gases into the atmosphere by ETIs. There have been several proposals for mitigating such a greenhouse effect engendered by global warming. One strategy relies on reducing the flux of starlight incident on the planet by placing a starshade in the vicinity of the first Lagrange point ($\mathcal{L}_1$) of the planet. The engineering specifics of the starshade differ from proposal to proposal, but it suffices to note that its size is roughly assumed to be comparable to the planet.

Gaidos (2017) studied the effect of a hypothetical starshade on the transit light curve and determined that a local maximum is potentially produced at midtransit due to the occultation of the starshade by the planet. A schematic illustration of the resulting geometry and transit light curve are provided in Figure 9.5. Gaidos determined that the following pair of conditions must be satisfied in order to optimize the likelihood of detection:

$$\rho_p \gtrsim 3\rho_\star \tag{9.68}$$

and

$$b < \left(\frac{3\rho_\star}{\rho_p}\right)^{1/3}, \tag{9.69}$$

where ρ_p and $\rho_\star$ are the mean densities of the planet and star, respectively, whereas b denotes the transit impact parameter defined in Section 6.1.1. It is possible to reduce (9.68) further by making use of $\rho_\oplus \approx 3.9\rho_\odot$, $\rho_p \propto M^{0.19}$, and $\rho_\star \propto M_\star^{-1.4}$, with M and $M_\star$ representing the planetary and stellar masses. Thus, after simplification, we end up with

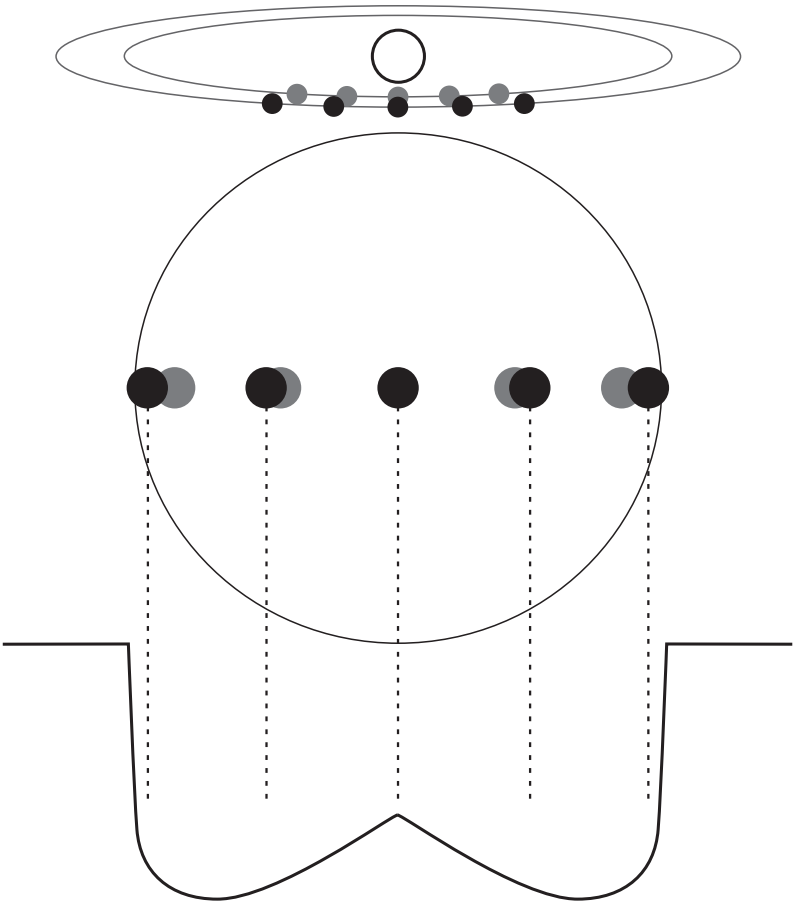


Figure 9.5 The geometry corresponding to a transiting exoplanet and starshade, respectively represented by the black and gray discs. As seen by an observer, the projected separation between the two objects is maximum at the ingress and egress of the planet, owing to which the deepest transits are manifested (minima in the light curve) at these locations. At midtransit, the planet occults the starshade, thereby giving rise to a local maximum in the light curve. (© The Author. Published by Oxford University Press on behalf of the Royal Astronomical Society. *Source:* E. Gaidos [2017], Transit detection of a "starshade" at the inner Lagrange point of an exoplanet, *Monthly Notices of the Royal Astronomical Society* 469[4]: 4455–4464, fig. 3.)

$$\left(\frac{M}{M_\oplus}\right)^{0.19} \gtrsim 0.77 \left(\frac{M_\star}{M_\odot}\right)^{-1.4}. \tag{9.70}$$

If we consider a planet whose mass is close to the Earth's, it is apparent that the above inequality is satisfied only when $M_\star \gtrsim 0.83\, M_\odot$. Hence,

Earth-sized planets around M- and K-type stars are less optimal in terms of identifying hypothetical starshades relative to G-type stars. On the other hand, stars more massive than the Sun pose their own set of difficulties because the contrast ratio becomes smaller, thus making their planets harder to detect. Alternatively, if we hold the stellar mass fixed, we see that the planet's mass must be at least 25 percent that of the Earth.

Potential false positives for starshades include starspots and dust clouds, but the former are predicted to be identifiable through their distinctive shape and features in the transmission spectrum. This is expected because the optimal design of starshades requires the preferential deflection of radiation emitted at wavelengths close to the blackbody peak of the host star. Finally, we note that starshades are not the only means of reducing the effects of global warming. Another option is to increase the planet's albedo by introducing aerosols into the atmosphere. However, discerning such artificially engineered hazes from naturally occurring ones would be difficult and requires in-depth modeling of transmission and reflectance spectra to identify proper diagnostics.

9.5.4 Redistribution of stellar energy

In Section 5.3, we noted that many planetary systems could possess tidally locked planets with the nightside permanently facing away from the host star; this is because of the fact that tidal forces are higher for planets around M-dwarfs, the latter of which are the most common type of stars in the Universe. If any ETIs were to exist on such worlds, they may wish to illuminate the nightside for a number of reasons; an obvious advantage is that the incident radiation can be harnessed to power photosynthesis. We will therefore discuss two strategies that are well suited to achieve this goal.

9.5.4.1 Photovoltaic arrays

The Kardashev Scale and its extensions classify ETIs on the basis of their energy consumption. It is evident that harnessing stellar energy is advantageous because it is both plentiful and renewable up to the star's death. Hence, Lingam and Loeb (2017c) explored the possibility that ETIs may opt to cover a fraction of the planetary surface with photovoltaic cells to convert stellar energy into electricity via the photovoltaic effect.

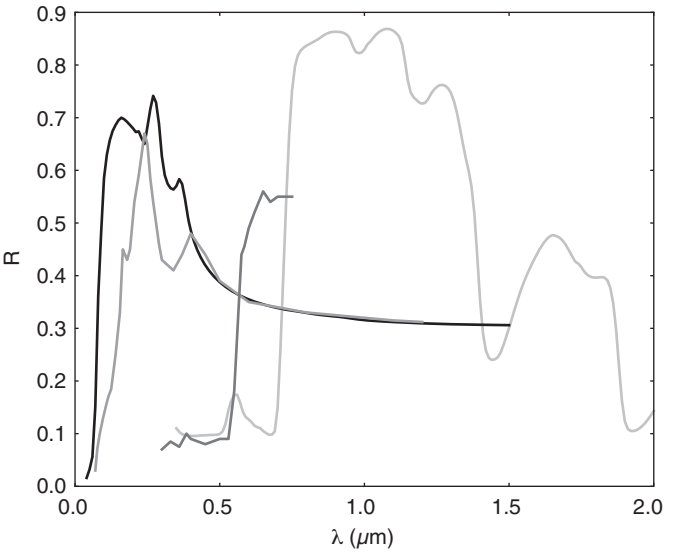


Figure 9.6 Reflectance (R) as a function of wavelength (λ) for several photovoltaic materials. The first three curves (*left to right*) correspond to silicon, gallium arsenide, and perovskite, respectively. The fourth curve, in contrast, depicts the reflectance for healthy oak leaves (vegetation) with the red edge clearly visible. (© The Authors. Published by Oxford University Press on behalf of the Royal Astronomical Society. *Source:* Manasvi Lingam and Abraham Loeb [2017], Natural and artificial spectral edges in exoplanets, *Monthly Notices of the Royal Astronomical Society,* 470[1]: L82–L86, fig. 1.)

If ETIs do opt to harness stellar energy in this manner, we must ask ourselves how large-scale photovoltaic arrays are detectable. The answer is to be found by examining the reflectance of commonly employed photovoltaic materials as a function of wavelength. In all instances, they give rise to distinctive spectral edges in the ultraviolet or visible wavelengths, analogous to the red edge of vegetation encountered in Section 6.5.1, as observed in Figure 9.6. Hence, if a marked increase in the intensity of reflected spectra is observed at the corresponding wavelength, it could indicate the existence of planetary-scale photovoltaic arrays. If we express the change in the geometric albedo (A_g) in the immediate vicinity of the spectral edge as $f_{PV}A_g$, the change in the reflected-light contrast (δC_S) is

$$\delta C_S \sim 4 \times 10^{-10} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a}{0.05\,\mathrm{AU}}\right)^{-2} \left(\frac{A_g}{0.3}\right) \left(\frac{f_{PV}}{0.01}\right), \qquad (9.71)$$

after making use of (6.28); the choice of normalization for a is based on the semimajor axis of Proxima b. In Section 6.2.1, we noted that contrast ratios of $\sim 10^{-10}$ are feasible in the near future via coronographs, although there are stringent constraints regarding angular resolution in need of addressing. Hence, for Earth-sized planets orbiting late M-dwarfs (with $M_\star \lesssim 0.15\, M_\odot$), it may be possible to detect photovoltaic arrays that cover $\gtrsim 1$ percent of the planet's surface in the upcoming decades. A lower bound on the power consumption of putative ETIs is also easy to estimate once f_{PV} is known for a particular photovoltaic material, by using

$$W_{\rm ETI} \sim 1.2 \times 10^{14}\,{\rm W} \left(\frac{\eta_{PV}}{0.1}\right)\left(\frac{f_{PV}}{0.01}\right)\left(\frac{1-A_p}{0.7}\right) \times \left(\frac{L_\star}{L_\odot}\right)\left(\frac{R}{R_\oplus}\right)^2 \left(\frac{a}{1\,{\rm AU}}\right)^{-2}, \tag{9.72}$$

where η_{PV} represents the energy conversion efficiency of the photovoltaic cell and A_p is the planet's Bond albedo.

Photovoltaic cells are not the only means of harnessing stellar energy. In fact, by taking inspiration from oxygenic photosynthesis in the natural world, humanity invented artificial photosynthesis in the twentieth century; one of the best-known pathways entails photocatalytic water splitting by means of titanium dioxide. Hence, when it comes to technosignatures, it behooves us to bear in mind the maxim advanced by science-fiction author and futurist Karl Schroeder (2011):

> Any sufficiently advanced technology is indistinguishable from nature.

If this postulate is indeed correct, we must ask ourselves whether there are other avenues by which we can distinguish *artificial* from *natural* photosynthesis. The answer may lie in observing anomalous levels of heat and light on the nightside of tidally locked exoplanets. For instance, in the absence of an atmosphere or oceans, if we were to discover efficient heat redistribution to the nightside, this could be labeled *anomalous* and potentially indicative of ETI. Detecting anomalous levels of light is feasible via temporal photometric observations, whereas the redistribution of heat is discernible from thermal phase curves, as explained in Section 6.2.2. Finding signatures along

these lines will also assist in ruling out other false positives such as certain minerals that give rise to spectral edges (e.g., enstatite).

9.5.4.2 Mirror fleets

The second possibility we wish to highlight concerns the deployment of large-scale mirror fleets to redirect starlight toward the nightside of the tidally locked planet. An important caveat worth mentioning here is that this approach is not feasible for humanity in the near future on economic grounds. In order to create a fleet of mirrors with a disk area equal to Earth, the total economic cost would exceed $\sim \$10^{16}$ (US), which is orders of magnitude higher than the world's gross domestic product. Nevertheless, this demanding engineering project is probably well within the purview of a hypothetical ETI with $\mathcal{K}_{\rm ETI} \approx 1.5$.

Korpela et al. (2015) examined the transit light curves that would result from mirror fleets and concluded that they are potentially detectable by JWST for nearby stars. The absorptance ($\alpha_{\rm mirror}$) of the mirror ring was parameterized as

$$\alpha_{\rm mirror} = \left[\left(\frac{R_{\rm mirror}}{R} \right)^2 - 1 \right]^{-1}, \tag{9.73}$$

where $R_{\rm mirror}$ denotes the radial distance of the mirror fleet from the planet; the range of values adopted for $R_{\rm mirror}$ was $2R$ to $10R$. Two different scenarios and their accompanying transit light curves were investigated: (1) the planet encompassed by the mirror fleet and (2) a larger planet in the absence of mirrors that gives rise to the same transit depth as (1). It is theoretically possible to distinguish (1) from (2) by studying both the timing and shape of the transit light curve at ingress and egress. Furthermore, the transit for (1) is initiated earlier and the minimum of the light curve is attained more gradually.

Korpela et al. (2015) found that the residual differences in the light curves—obtained by subtracting the light curve of (2) from that of (1)—normalized by the transit depth did not vary much across their simulations and were typically on the order of a few percentage points. To express it differently, the change in the transit light curve due to the mirror fleet was on the order of $\sim 10^{-4}$ times the stellar intensity for a star with $M_\star \approx 0.5\, M_\odot$; this fraction decreases by a factor of 2 when it comes to $M_\star \approx M_\odot$. In order

to identify the mirror fleet at 4σ confidence using the JWST for nearby stars, it was determined that ~ 10 transits may be necessary.

9.5.5 Satellites and space debris

Artificial satellites have quickly become indispensable to modern human society as they are used in telecommunications, navigation, and weather. The geostationary orbit is considered especially useful because satellites at this location will appear stationary as viewed from the planet's surface. The geostationary orbit is a special case of geosynchronous orbits, with the latter characterized by possessing an orbital period equal to the rotation period of the planet. Hence, it is plausible that some ETIs would also opt to employ satellites in geostationary or geosynchronous orbits. In this event, it is natural to ask whether such a belt of satellites is detectable. It is worth noting, however, that the geostationary orbit is not always stable, implying that it may not represent the optimal choice for certain worlds.[9]

This issue was explored by Socas-Navarro (2018) for a wide range of parameters. Two variables taken into consideration were the belt's opacity (χ_o) and width. The belt opacity represents the fraction of light in the observer's line of sight that is blocked by a surface element; it serves as a proxy for the fiducial density and size of the satellites. In the case of present-day humanity, $\chi_o \sim 3 \times 10^{-13}$, but Socas-Navarro adopted much higher values, on the order of 10^{-4}. Another important parameter is the radius of the geosynchronous orbit ($R_{\rm geo}$) as measured from the center of the planet, which is obtained from Kepler's Third Law by specifying the mass to be equal to that of the planet and the orbital period equal to the planet's rotation rate. Thus, we end up with

$$R_{\rm geo} = \left(\frac{GM\tau_{\rm rot}^2}{4\pi^2}\right)^{1/3} \approx 4.2 \times 10^7 \text{ m} \left(\frac{M}{M_\oplus}\right)^{1/3} \left(\frac{\tau_{\rm rot}}{1 \text{ day}}\right)^{2/3}, \quad (9.74)$$

where $\tau_{\rm rot}$ is the rotation period of the planet. It is difficult to determine $\tau_{\rm rot}$, but some potential methods are described in Chapter 6. Alternatively, if a planet is tidally locked, it is straightforward to calculate its rotation period since it will equal the orbital period, which can be measured.

---

9. Tidally locked exoplanets, in particular, could have unstable geostationary orbits. The authors are grateful to Hector Socas-Navarro for raising this point.

Before discussing whether an artificial satellite belt is discernible through transit light curves, a potential false positive must be mentioned: ring systems. In certain respects, ring systems resemble a belt of artificial satellites, but the underlying geometry might be qualitatively different. To understand why, it is necessary to appreciate the point that the geosynchronous satellite belt is concentrated toward a single radial location ($R_{\rm geo}$), whereas the ring system is spread out radially. Likewise, the satellite belt is extended in terms of its inclination, while the ring system is thin. Of course, it must be recognized that both these geometries are simplified descriptions, and real-world systems may not exhibit these properties.

In Figure 9.7, the light curve for a planet with a ring system is contrasted against one possessing a geosynchronous satellite belt and one lacking both rings and artificial satellites. First, it is apparent that the presence of either rings or satellites induces a much deeper transit as well as different characteristics at ingress and egress. In order to distinguish rings from satellites, note that the light curve for the satellite belt at ingress ($\rm CEB_1$) is purely concave, whereas it is always convex at egress ($\rm CEB_2$). In contrast, the light curve of the ring system switches from convex to concave at both locations. More visibly, the minimum of the light curve is attained in a smooth fashion for the ring system, while the slope of the curve is discontinuous for the satellite belt. Hence, in principle, it might be feasible to differentiate between artificial satellites and ring systems for transiting exoplanets.

A major issue worth highlighting here is that a belt opacity of $\chi_0 \sim 5 \times 10^{-4}$ with individual satellites of radius ~ 1 m and ~ 100 kg employed in the simulations translates to a total mass of $\sim 10^{12}$ to 10^{14} kg. If we assume an optimistic launch cost of $\sim \$1000$/kg, which is presumably achievable at present, the total cost amounts to $\sim \$10^{15}$ to $\$10^{17}$ (US), which is far beyond the scope of humanity from an economic standpoint. Second, the dynamical stability of the geosynchronous satellite belt is questionable at high densities. This drawback was comprehensively explored by Sallmen et al. (2019), but we shall develop a simplified version that captures the salient details.

Let us denote the number density of satellites (chosen to have a homogeneous size) by $n_{\rm sat}$ and their cross-sectional area by $\mathcal{A}_{\rm sat}$. We model their average velocity as $v_{\rm sat} = 2\pi R_{\rm geo}/\tau_{rot}$. The collision rate between these objects ($\dot{\mathcal{C}}$) is explicated in Kessler and Cour-Palais (1978) and is given by

$$\dot{\mathcal{C}} = \frac{1}{2} \int n_{\rm sat}^2 v_{\rm rel} \mathcal{A}_{\rm sat} \, d\mathcal{V}, \tag{9.75}$$

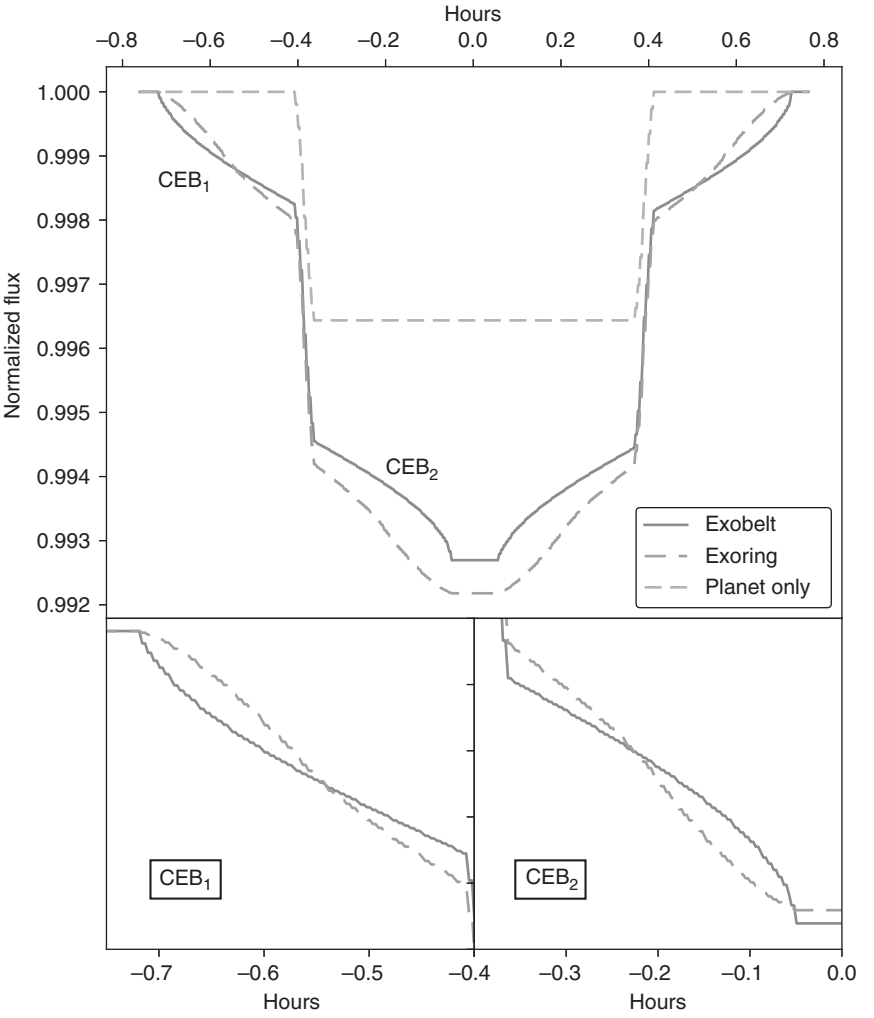


Figure 9.7 Synthetic transit light curves for an M-dwarf exoplanet with either a ring system, geosynchronous satellite belt, or neither of them. *Top:* Entire transit; the shallower transit (smaller dashed curve) corresponds to that of the planet alone. *Bottom:* Two zoomed-in sections. The location of CEB_2 is not wholly congruous in the top and bottom panels because it is shifted vertically in the former to permit comparison. The ring system is assumed to span the region 0.5 R_{geo} to R_{geo}. (© The American Astronomical Society. *Source:* Hector Socas-Navarro [2018], Possible photometric signatures of moderately advanced civilizations: The Clarke exobelt, *Astrophysical Journal* 855[2]: 110, fig. 8.)

where $d\mathcal{V} = r'^2 \sin\theta'\, dr' d\theta' d\phi'$ is the volume element in spherical coordinates and $v_{\rm rel}$ is the ensemble-averaged relative velocity. In the case of a satellite belt centered on geosynchronous orbit, we choose $r' \in \left[R_{\rm geo} - 0.5\Delta r_{\rm sat}, R_{\rm geo} + 0.5\Delta r_{\rm sat}\right]$, $\theta' \in [0, \theta_{\rm max}]$, and $\phi' \in [0, 2\pi]$, where $\Delta r_{\rm sat}$ represents the radial width of the satellite belt and $\theta_{\rm max}$ is the maximum inclination. For this ansatz, the total volume is expressible as

$$\int d\mathcal{V} = 2\pi R_{\rm geo}^2 \Delta r_{\rm sat} \left(1 - \cos\theta_{\rm max}\right). \tag{9.76}$$

We shall select the fiducial values of $\Delta r_{\rm sat} = 0.01\, R_{\rm geo}$ and $\theta_{\rm max} = 20°$, but other values can be easily utilized. Hence, after simplifying (9.75), we obtain

$$\dot{\mathcal{C}} \sim 4 \times 10^{-3}\,{\rm yr}^{-1} \left(\frac{n_{\rm sat}}{3 \times 10^{-18}\,{\rm m}^{-3}}\right)^2 \left(\frac{v_{\rm rel}}{10\,{\rm km/s}}\right) \left(\frac{\mathcal{A}_{\rm sat}}{10\,{\rm m}^2}\right) \times \left(\frac{M}{M_\oplus}\right) \left(\frac{\tau_{\rm rot}}{1\,{\rm day}}\right)^2, \tag{9.77}$$

where the choice of normalization for $n_{\rm sat}$ is highly conservative, corresponding to the number density of satellites in near-Earth orbit during the late twentieth century. In contrast, if we specify $n_{\rm sat} \sim 3 \times 10^{-9}\ {\rm m}^{-3}$, which is potentially consistent with the very high belt opacity introduced previously, we arrive at $\dot{\mathcal{C}} \sim 4 \times 10^{15}\,{\rm yr}^{-1}$. As a result, the satellites would be quickly reduced to debris via collisional cascades. This phenomenon is known as the Kessler syndrome, named after Donald Kessler, who analyzed it during the 1970s. In order to address this issue, ETIs would need to reduce either $v_{\rm rel}$ or $\mathcal{A}_{\rm sat}$ by many orders of magnitude. Adjusting $\mathcal{A}_{\rm sat}$ is constrained by satellite functionality, owing to which the condition $v_{\rm rel} \to 0$ must be fulfilled. It is worth noting that the same constraints are applicable to nonrigid Stapledon-Dyson spheres (Section 9.4.2), which are best envisioned as swarms of individual units.

9.5.6 Artificial illumination

We have already analyzed how mechanisms used for redistributing light and heat islands arising from cities could be detected. Here, we shall focus on

a different aspect—namely, identifying sources of artificial illumination as delineated in A. Loeb and Turner (2012).

Let us consider sources of artificial illumination within our Solar system at distances of $a \gg 1$ AU. The flux detected by an observer on Earth falls off as a^{-2} for artificial illumination, whereas in the case of natural objects illuminated by reflected sunlight it would decline as a^{-4}. The latter arises because the stellar flux incident on the object scales as a^{-2} while the flux received by an observer on Earth introduces another factor of a^{-2}. Hence, monitoring how the observed flux varies as a function of its changing orbital distance will permit us to differentiate between natural and artificially illuminated objects; the former will exhibit a slope of -4, whereas the latter are characterized by a slope of -2. Although this diagnostic is simple and powerful, we note that a number of complicating factors arise because the brightness also varies with phase angle, outgassing, and surface albedo, to name a few.

Let us suppose that the illumination produced by an artificial source has a flux of $\epsilon_{\rm light} S_\oplus$, where $S_\oplus \sim 1.4 \times 10^3$ W/m^2 is the solar constant. We will suppose that the Bond albedo of the natural object is A_p and that it is situated at a distance $a \gg 1$ AU from the Sun. Next, we denote the ratio of the size of the artificial source to that of a natural object at the same distance by δ_{NA}. If we demand that the flux received from both sources at the Earth is comparable, we obtain

$$\delta_{NA} \approx 5 \times 10^{-2} \left(\frac{a}{1\,\mathrm{AU}}\right)^{-1} \left(\frac{A_p}{0.07}\right)^{1/2} \left(\frac{\epsilon_{\rm light}}{0.01}\right)^{-1/2}. \tag{9.78}$$

Hence, given that Kuiper Belt Objects of sizes $\sim 10^3$ km were found by telescopes at distances of ~ 50 AU, it is conceivable that putative objects with artificial illumination on the order of 100 km might be detectable at the same distance.

The situation is rendered very different when we consider *specular* reflection from mirrorlike objects in place of diffuse reflection; we briefly tackled specular reflection in Section 6.2.1 when covering the glint effect. We will adopt the methodology presented in Lacki (2019) henceforth. The distance of the artifact from the Sun is denoted by $d_\odot$, whereas the corresponding distance to an Earth-based observer is $d_\oplus$. When the object is very far from the Sun, it is evident that $d_\oplus \approx d_\odot$. As seen from the artifact, the Sun resembles a disk with an angular size of $\theta_\odot = R_\odot / d_\odot$. In the idealized

limit, we can compute the time taken by an observer to traverse the angular region $2\theta_\odot$, assuming that the relative velocity between the artifact and observer is $V_{\rm rel}$. This time represents the duration over which the glint is visible from Earth ($t_{\rm glint}$) and is given by

$$t_{\rm glint} \sim \frac{2\theta_\odot d_\oplus}{V_{\rm rel}} \sim 12.9\,{\rm hr}\left(\frac{V_{\rm rel}}{30\,{\rm km/s}}\right)^{-1}\left(\frac{d_\oplus}{d_\odot}\right). \tag{9.79}$$

The flux ($\Phi_{\rm spec}$) detected by an observer at Earth is expressible as

$$\Phi_{\rm spec} = \frac{A_p L_\odot \mathcal{A}_{\rm obj}}{4\pi R_\odot^2 d_\oplus^2}, \tag{9.80}$$

where $\mathcal{A}_{\rm obj}$ is the projected area of the object from the viewpoint of the observer, and A_p represents the reflectivity in this instance. In contrast, if we consider a diffuse (i.e., isotropic) object with the same albedo and area situated at the same distance, its flux ($\Phi_{\rm diff}$) will be

$$\Phi_{\rm diff} \approx 2.2 \times 10^{-5}\left(\frac{d_\odot}{1\,{\rm AU}}\right)^{-2}\Phi_{\rm spec}. \tag{9.81}$$

Lastly, we can assess the relative prospects of detecting specular and dull (i.e., diffuse) reflectors. On the one hand, as seen from the above equation, the maximum distance to which a specular reflector is detectable is typically higher than its diffuse counterpart. On the other hand, because the reflection is focused into a solid angle of $\sim\pi\theta_\odot^2$, the probability of the glint being aligned with the observer's field of view is $\sim\pi\theta_\odot^2/(4\pi)$. Lacki (2019) determined the ratio of the effective survey volumes for specular and dull reflectors (denoted by $\delta_{\mathcal{V}}$) by combining these two factors and arrived at

$$\delta_{\mathcal{V}} \approx \frac{1}{4}\left(\frac{d_{\rm diff}}{R_\odot}\right)\left(\frac{d_{\rm diff}}{d_{\rm spec}}\right)^2, \tag{9.82}$$

where $d_{\rm diff}$ and $d_{\rm spec}$ represent the distances between the Sun and the dull and specular artifacts, respectively. Broadly speaking, it is found that small and medium-size specular artifacts are favored in terms of detection (i.e., higher $\delta_{\mathcal{V}}$) over their diffuse counterparts, whereas the converse is true for large objects.

In looking beyond the Solar system, it is apparent that the strategy of monitoring variations in the observed flux with distance does not work properly because the flux received at Earth falls off as $d_\star^{-2}$ to leading order regardless of whether the illumination of the source is artificial or natural. It might be feasible, instead, to discern nightside illumination by measuring the orbital phase modulation of the observed flux. However, in order for this technique to become viable, one requires either a nightside illumination that is comparable to that of the dayside over a given bandwidth or telescopes with kilometer-sized apertures for detecting current levels of artificial illumination on Earth (Schneider et al. 2010).

In closing, we note that the search for artificial objects is one that can be realistically undertaken within our Solar system. However, searches for technosignatures have generally tended to direct their attention outside the Solar system, with only a handful of exceptions. Robert Freitas and Francisco Valdes (1985) undertook a fairly detailed survey of the Earth-Moon Lagrange points in the 1980s and found no evidence for artifacts of sizes $\gtrsim 1$–10 m with geometric albedos $\gtrsim 0.1$. Some preliminary searches for interstellar probes in the vicinity of the Earth have also been carried out at optical wavelengths but did not yield any conclusive results (Ansbro 2001).

9.5.7 Transiting megastructures

The reader will have noticed that we already tackled specialized cases of transiting megastructures: starshades, space-based mirror fleets, and satellite belts. We will concern ourselves here with describing some *general* characteristics of megastructures in transit light curves and the ensuing implications.

Arnold (2005) presented the first quantitative results underscoring the fact that transiting megastructures could give rise to light curves whose properties are very different from natural objects. Hypothetical Jupiter-sized megastructures at 1 AU around a star of $1.15\,M_\odot$ were simulated, and residual differences in light curves between these objects and spheres with the same cross-sectional area were computed during transits. It was found that distinctive patterns emerged depending on the shape of the megastructures, with residual differences on the order of 10^{-4}, within the photometric precision of the *Kepler* mission. Furthermore, when multiple objects were incorporated into the study, the resulting light curves were argued to be unique and not reproducible by natural phenomena.

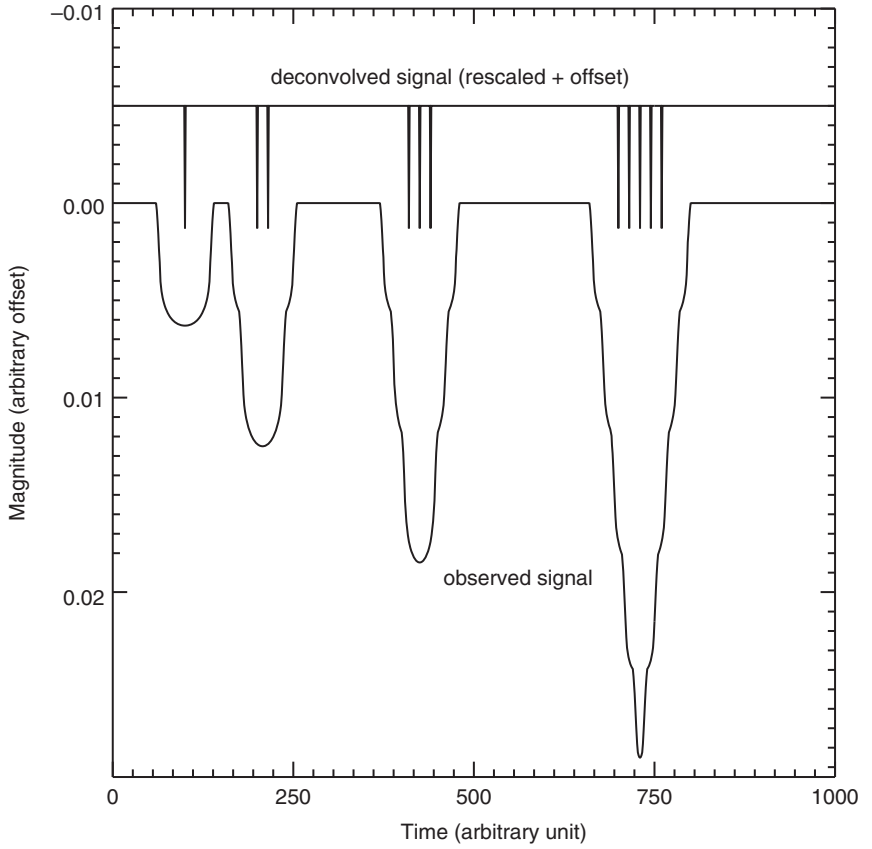


Figure 9.8 Multiple transits arising from a system comprising 11 objects in total arranged in groups of 1, 2, 3, and 5 objects. The time between transits also obeys a prime number sequence. Each object was assigned a cross section similar to Saturn and transits a 1.15 $M_{\odot}$ star at 1 AU. (© The American Astronomical Society. *Source:* Luc F. A. Arnold [2005], Transit light-curve signatures of artificial objects, *Astrophysical Journal* 627[1]: 534–539, fig. 8.)

Arnold (2005) also raised the intriguing point that multiple transiting objects present a natural means of encoding information; one such example is depicted in Figure 9.8. The transit depth is roughly proportional to the number of objects, implying that a sequence of transit depths that obeys 1 : 2 : 3 : 5 . . . probably has an artificial origin. Moreover, the spacing between the transits of successive objects (or successive groups of objects)

can be modulated to incorporate additional informational content. Lastly, the duration of transit associated with each object or group could be varied as it represents another avenue for encoding information. As per this line of reasoning, ETIs that seek to employ transits for the purpose of transmitting information are more likely to utilize many smaller objects rather than a few big ones.

If the number of bits conveyed in such a fashion is $\mathcal{B}$, the net data rate ($\dot{\mathcal{B}}$) is determined from the product of $\mathcal{B}$ and the solid angle encompassed by the transit zone divided by the orbital period of the megastructure(s), which yields

$$\begin{aligned}\dot{\mathcal{B}} &\sim 4\pi \left(\frac{R_\star}{a}\right) \frac{\mathcal{B}}{\mathcal{P}_{\rm pl}} \\ &\sim 1.9 \times 10^{-8} \text{ bps sr} \left(\frac{\mathcal{B}}{10}\right) \left(\frac{R_\star}{R_\odot}\right) \left(\frac{M_\star}{M_\odot}\right)^{1/2} \left(\frac{a}{1\,\text{AU}}\right)^{-5/2}\end{aligned} \tag{9.83}$$

where the orbital period $\mathcal{P}_{\rm pl}$ has been expressed in terms of a and $M_\star$ using Kepler's Third Law. Arnold (2005) suggested that $\dot{\mathcal{B}}$ for transiting megastructures and electromagnetic signals might be comparable, although in both instances there are a number of free variables. Instead of using the data rate as the metric, the energetic costs associated with different types of beacons can be evaluated. Arnold (2013) developed a simple theoretical model and found that the power consumed per star is only about an order of magnitude higher for transiting megastructures relative to radio signals, provided that the former are long-lived.

Catching the attention of other technological species during transits does not necessitate the construction of megastructures. As an observer in the transit zone will observe the characteristic dip in brightness when monitoring the light curve of transiting exoplanets, it is possible to alter the features of the light curve using lasers. This can be effectuated in many ways by producing unusual spikes, but starspots could also give rise to similar results. Hence, Kipping and Teachey (2016) proposed that cloaking the ingress and egress of a transit while preserving the rest of the transit serves as an unambiguous signal of ETIs since no natural phenomenon is known to date that yields this feature. The total energy expenditure ($\mathcal{E}_t$) required to accomplish this feat is

$$\mathcal{E}_t \sim 2.5 \times 10^{10}\,\mathrm{J} \left(\frac{R}{R_\oplus}\right)^3 \left(\frac{\lambda}{600\,\mathrm{nm}}\right)^2 \left(\frac{D_t}{10\,\mathrm{m}}\right)^{-2} \times \left(\frac{L_\star}{L_\odot}\right)^{5/4} \left(\frac{R_\star}{R_\odot}\right)^{-2} \left(\frac{M_\star}{M_\odot}\right)^{-1/2}, \tag{9.84}$$

which can be further simplified by using $L_\star \propto M_\star^3$ and $R_\star \propto M_\star^{0.8}$, which leads to $\mathcal{E}_t \propto M_\star^{1.65}$. As opposed to broadcasting their presence, ETIs may opt to cloak the transit light curves of their host planets. The energy expenditure incurred is straightforward to calculate, yielding

$$\mathcal{E}_t \sim 2.7 \times 10^{12}\,\mathrm{J} \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{\lambda}{600\,\mathrm{nm}}\right)^2 \left(\frac{D_t}{10\,\mathrm{m}}\right)^{-2} \times \left(\frac{L_\star}{L_\odot}\right)^{5/4} \left(\frac{R_\star}{R_\odot}\right)^{-1} \left(\frac{M_\star}{M_\odot}\right)^{-1/2}. \tag{9.85}$$

By substituting the mass-radius and mass-luminosity scaling relationships introduced above, we arrive at $\mathcal{E}_t \propto M_\star^{2.45}$. Strictly speaking, both (9.84) and (9.85) are valid only when the transit impact parameter (b) is much smaller than unity.

If transiting megastructures do exist, the next question is: What are the anomalies that would serve to distinguish them from natural objects? A detailed analysis of this issue was undertaken by J. T. Wright et al. (2016), who identified ten anomalies arising from unusual light curves, orbits, masses, transmission spectra, presence of swarms, and complete obscuration of stars. With regard to the latter, we note that a comprehensive search for missing stars and galaxies ($\sim 10^7$ objects in total) using databases from different epochs was undertaken by Villarroel et al. (2016), who concluded that none of them were strong candidates for follow-up studies, with the possible exception of one dubious object; further observations of this source in an extended survey (comprising $\sim 10^9$ objects) did not yield any unusual results (Villarroel et al. 2020). The false positives linked with each of the aforementioned ten anomalies were simultaneously explored by J. T. Wright et al., and Table 9.1 delineates the ensuing predictions.

The next point to address is the *information* content of transit light curves. As we have seen, it may be feasible for putative ETIs to encode

Table 9.1 Ten anomalies of transiting megastructures that could distinguish them from planets or stars

Anomaly	Artificial mechanism	Natural confounder
Ingress and egress shapes	Non-disk aspect of the transiting object or star	Exomoons, rings, planetary rotation, gravity and limb darkening, evaporation, limb starspots
Phase curves	Phase-dependent aspect from nonspherical shape	Clouds, global circulation, weather, variable insolation
Transit bottom shape	Time-variable aspect during transit, e.g., changes in shape or orientation	Gravity and limb darkening, stellar pro/oblateness, starspots exomoons, disks
Variable depths	Time-variable aspect during transit, e.g., changes in shape or orientation	Evaporation, orbital precession, exomoons
Timings/Durations	Nongravitational accelerations, co-orbital objects	Planet-planet interactions, orbital precession, exomoons
Inferred stellar density	Nongravitational accelerations, co-orbital objects	Orbital eccentricity, rings, blends, starspots, planet-planet interactions, very massive planets
Aperiodicity	Swarms	Very large ring systems; large debris fields; clumpy, warped, or precessing disks
Disappearance	Complete obscuration	Clumpy, warped, precessing, or circumbinary disks
Achromatic transits	Artifacts could be geometric absorbers	Clouds, small scale heights, blends, limb darkening
Very low mass	Artifacts could be very thin	Large debris field, blends

Notes: Observational signatures that are potentially attributable to transiting megastructures, and false positives that might give rise to such anomalies. *Achromatic transits* refers to the absence of wavelength-dependent spectral features in transmission spectroscopy. (© The American Astronomical Society. *Data source:* Jason T. Wright, Kimberly M. S. Cartier, Ming Zhao, Daniel Jontof-Hutter, and Eric B. Ford [2016], The search for extraterrestrial civilizations with large energy supplies. IV. The signature and information content of transiting megastructures, *Astrophysical Journal* 816[1]: 17, table 1.)

information by using appropriately chosen megastructures during transit. To quantify the information content, J. T. Wright et al. (2016) developed an ingenious diagnostic, dubbed the normalized information content (NIC) statistic,

$$\mathrm{NIC} = \frac{K_m - K_0}{K_{\max} - K_0}, \tag{9.86}$$

where K_m signifies the Kullback-Leibler (KL) divergence of a given signal. The KL divergence is defined as

$$K_m = \int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx, \tag{9.87}$$

where $p(x)$ constitutes the signal's probability distribution function and $q(x)$ represents the probability distribution function that approximates the signal. When a signal has a high information content, $p(x)$ is characterized by a large number of possible values, whereas the converse is true for low content. In quantitative terms, the maximal and minimal signals correspond to a uniform probability distribution and a delta function for $p(x)$, respectively. The KL divergence for the maximal and empty signals is denoted by $K_{\max}$ and K_0, respectively. In calculating $K_{\max}$ and K_0, it is vital to ensure that the maximal and empty signals feature the *same* SNR and signal length as the measured signal used to determine K_m.

The time series of the transit depth was analyzed by J. T. Wright et al. (2016) for a number of cases to assess whether NIC functions as a reliable diagnostic. When this method was applied to the transit light curves of Kepler-4b and Kepler-5b, it was found that $\mathrm{NIC} \approx 0.0483$ and $\mathrm{NIC} \approx 0.0215$, respectively. In contrast, for signals with low SNR ($\lesssim 10$) derived from a couple of representative artificial megastructures, values of $\mathrm{NIC} \approx 0.8806$ and 0.7516 were obtained. Hence, at first glimpse, it would appear as though this metric is fairly reliable. Yet, when this measure was applied to Kepler-1520b, it was determined that $\mathrm{NIC} \approx 0.6976$. Theoretical models that explain the observed variations by means of planetary evaporation and disintegration have provided a satisfactory explanation for this light curve. Therefore, even if NIC lacks the requisite sensitivity to differentiate between natural and artificial objects, it may still play a valuable role in highlighting objects of interest, characterized by high values of NIC, that merit follow-up observations.

In summary, none of the $\sim 2 \times 10^5$ stars surveyed by *Kepler* have revealed unambiguous signatures of megastructures, but no formal studies undertaken to date have precisely derived upper limits on their prevalence. Before bringing our analysis to a close, we briefly comment on megastructures apart from the ones encountered in this chapter that might be constructed by putative ETIs.

1. If part of the stellar disk is obscured by a Shkadov thruster (see Section 9.4.3), it will cause changes in the transit light curve. Forgan (2013) studied the transit light curve for the composite system (exoplanet and Shkadov thruster) and determined that it exhibits an asymmetry compared to the scenario where only the exoplanet exists. More specifically, the transit is characterized by an early egress in the presence of a Shkadov thruster, but the same effect could arise from starspots. Forgan also estimated that the probability of detecting a Shkadov thruster via the transit method is very low, with an upper bound of $< 10^{-6}$.
2. In order to mitigate the threats posed by high-energy astrophysical phenomena (e.g., supernovae and Gamma-Ray Bursts), Ćirković and Vukotić (2016) suggest that ETIs may opt to construct swarms of small objects held together by electromagnetic forces for shielding purposes. The identification of such swarms is theoretically feasible through measurements of the mass and radius of this structure, arriving at very low densities.
3. A major technological threat is posed by large coronal mass ejections (CMEs) when they impact planets inhabited by ETIs. The resulting electromagnetic disturbances can disrupt electrical power grids, satellite communications, and supply chains; the worldwide economic damage to humanity has been predicted to exceed $1 to $10 trillion (US). Hence, ETIs could opt to place a magnetic shield near the Lagrange point $\mathcal{L}_1$ to deflect the incoming charged particles. As the resulting engineering design conceivably resembles a planet-sized starshade (see Section 9.5.3) in some respects, similar signatures might be produced in the transit light curves.

9.5.8 Searches for Stapledon-Dyson spheres

In Section 9.4.2, we outlined how Stapledon-Dyson spheres may be constructed by hypothetical advanced ETIs seeking to harness stellar-level energy on a massive scale and how these structures are potentially detectable by searching for signatures of waste heat in the infrared.

The number of searches for Stapledon-Dyson spheres by means of this method have, however, been few and far in between. Most of them relied on analyzing the data collected by the Infrared Astronomical Satellite (IRAS) in the 1980s. IRAS conducted a virtual all-sky search (coverage of 96 percent) at wavelengths of 12, 25, 60, and 100 μm with a sensitivity of ~ 0.1–1 Jy. As seen from (9.45), IRAS might be capable of detecting IR radiation from Stapledon-Dyson spheres up to kiloparsec distances. One of the first searches for Stapledon-Dyson spheres using IRAS was carried out by Slysh (1985), who identified six anomalous IR sources. However, it was also pointed out that distinguishing between red giants with thick dust envelopes and Stapledon-Dyson spheres was not possible with the limited data available. In the 1990s, the launch of the Infrared Space Observatory (ISO) raised the prospects of detecting macroengineering projects (Tilgner & Heinrichsen 1998), but this endeavor did not bear fruit.

Timofeev et al. (2000), building on research reported in Kardashev (1997), also made use of the IRAS catalog to search for Stapledon-Dyson spheres by implementing blackbody temperature fits and determining whether the resultant sources had a temperature of $\lesssim 300$ K. A total of fourteen candidates were identified from 3000 sources, but none of them were deemed sufficiently anomalous. Jun Jugaku and Shiro Nishimura carried out a series of searches for partial or complete Stapledon-Dyson spheres with $\gamma_{\rm ETI} \gtrsim 10^{-2}$, where the latter quantity is defined in Section 9.4.4. The data from the Two Micron All-Sky Survey (2MASS) was analyzed, and it was concluded that none of the 384 Solar-type stars within 25 pc in this sample displayed any concrete evidence emblematic of Stapledon-Dyson spheres (Jugaku & Nishimura 2004).

The most thorough search to date for Stapledon-Dyson spheres, between temperatures of 100 and 600 K up to distances of 300 pc from Earth via their waste heat emission was conducted by Carrigan (2009). Out of the $\sim 2 \times 10^5$ sources initially investigated by IRAS, around 1.1×10^4 were subjected to further analysis by the Calgary Low Resolution Spectrometer (LRS), which provided more accurate and detailed data. The latter

served as the sample for Carrigan, and the sources were run through a number of filters, including blackbody temperature fits. The winnowing process yielded sixteen quasi-plausible candidates. Of these, the three most promising contenders were subjected to further scrutiny, and it was found that they could either be explained through natural phenomena or lacked sufficient information to derive conclusive results.

Apart from waste heat, an interesting alternative approach for discerning near-complete Stapledon-Dyson spheres was outlined by Zackrisson et al. (2018). This methodology relies on seeking discrepancies between the spectrophotometric and parallax distances, thereby discovering the existence of putative Stapledon-Dyson spheres. First, given a source with optical luminosity $L_{\rm opt}$ and the measured optical flux $\Phi_{\rm opt}$ (after correcting for dust), the spectrophotometric distance ($d_{\rm sp}$) is given by

$$d_{\rm sp} = \left(\frac{L_{\rm opt}}{4\pi \, \Phi_{\rm opt}} \right)^{1/2} . \tag{9.88}$$

Suppose that the source (i.e., the star) is encompassed by a partial Stapledon-Dyson sphere that blocks a fraction $\alpha_{\rm opt}$ of the optical radiation and functions as a gray (wavelength independent) absorber. In this event, the sole impact of this megastructure is that the optical flux is subjected to dimming without changing the spectral energy distribution. The actual distance $d_\star$ is therefore calculated via

$$d_\star = \left(\frac{L_{\rm opt} \left(1 - \alpha_{\rm opt}\right)}{4\pi \, \Phi_{\rm opt}} \right)^{1/2} = d_{\rm sp} \sqrt{1 - \alpha_{\rm opt}}. \tag{9.89}$$

If $\alpha_{\rm opt} \approx 1$, corresponding to a near-complete Stapledon-Dyson sphere, it follows that the true distance is smaller than what is determined from spectrophotometric observations. However, an alternative method for computing $d_\star$ that does not rely on stellar brightness is through measurement of the trigonometric parallax. The basic idea underpinning the parallax method is the parallax angle p: the angle between the Earth's position at a given time and its position six months later as seen from the perspective of a nearby star. Using trigonometry, it is easy to show that the distance to the star is expressible as $d_\star \, ({\rm pc}) = 1/p \, ({\rm arcsec})$, where $1 \, {\rm arcsec} \approx 4.8 \times 10^{-6}$ radians.

Hence, for sources with sufficiently accurate spectrophotometry and parallax, it is possible to seek anomalies that are characterized by $d_\star < 0.5 d_{\rm sp}$, which translates to $\alpha_{\rm opt} > 0.75$. Zackrisson et al. (2018) undertook a preliminary search for such sources that were included in both the Gaia and Radial Velocity Experiment surveys, thus comprising 2.3×10^5 objects in total. However, when only stars with accurate parallaxes were retained, the sample size was decreased to 8441 stars. By searching for $\alpha_{\rm opt} > 0.9$ and eliminating false positives, only a handful of outliers (two to six) were obtained. High-resolution spectroscopy of the most promising candidate, TYC 6111-1162-1, was carried out, and it was confirmed that $d_\star \approx 0.48\, d_{\rm sp}$, but no IR excess characteristic of Stapledon-Dyson spheres was detected. The advantage with this proposed technique is that it ought to facilitate the search for partial Stapledon-Dyson spheres across large sample sizes; for instance, the Gaia Data Release 3 will encompass $> 10^6$ stars.

Before forging ahead, we point out that searches for Stapledon-Dyson spheres ought not exclusively concentrate on main-sequence stars. From the standpoint of energetics as well as engineering considerations, there might be no definitive a priori reasons for ruling out the presence of Stapledon-Dyson spheres or artificial habitats around the likes of white dwarfs, neutron stars, and black holes (M. Inoue & Yokoo 2011; Osmanov 2016; Imara & Di Stefano 2018; Gertz 2019). Likewise, Stapledon-Dyson spheres need not exist only within the traditional habitable zone of main-sequence stars or other astrophysical sources. For example, if they are situated at sufficiently close-in distances, their emission may actually manifest at optical wavelengths (Osmanov & Berezhiani 2018).

9.5.9 Searches for Kardashev Type III ETIs

Recall from Section 9.4.1 that Type III ETIs are characterized by energy consumption on the scale of a typical galaxy. The necessity of extracting such high supplies of energy would call for engineering projects on a galactic scale, thus giving rise to potential technosignatures.

In an innovative publication, Annis (1999b) proposed that galaxies with Type III ETIs ought to be characterized by unusual galactic scaling relations. The gist of the argument is that astroengineering projects pursued by Type III ETIs would not disturb the gravitational potential and mass of the galaxy under consideration but would utilize the radiation from stars and thereby decrease the optical luminosity of the galaxy. One of

the most famous galactic scaling relations is the Tully-Fisher relationship, named after R. B. Tully and J. R. Fisher, who originated the central premise in 1977. The Tully-Fisher relationship states that the luminosity—or, equivalently, the absolute magnitude—and the rotation velocity (a proxy for the gravitational potential) of galaxies obey an empirical power-law scaling.

Annis (1999b) conducted a survey of 137 elliptical and disk galaxies by imposing the criterion that the actual luminosity should be at least four times smaller than the predicted value from the scaling relation, and he discovered no outliers. A similar study was undertaken by Zackrisson et al. (2015) for a sample of 1359 disk galaxies using the Tully-Fisher relationship. A total of eleven objects were significantly underluminous as per the Tully-Fisher relation, but further scrutiny of their infrared fluxes and morphologies (at optical wavelengths) failed to turn up any unusual features. Hence, Zackrisson et al. argued that the upper limit on the fraction of disk galaxies hosting Type III ETIs is $\lesssim 0.3$ percent.

Another sign of Type III ETIs is the production of waste heat, akin to the diagnostic employed for Stapledon-Dyson spheres but on much larger scales. The most comprehensive survey to date in this regard was undertaken by R. L. Griffith et al. (2015), who studied $\sim 10^5$ galaxies by utilizing the Wide-field Infrared Survey Explorer (WISE) in conjunction with supporting observations from 2MASS and the Spitzer Space Telescope. It was determined that none of the galaxies in this sample appear to host Type III ETIs that reprocess > 85 percent of the available starlight into mid-IR radiation (waste heat). When this lower bound was reduced to 50 percent, it was found that fifty extragalactic sources exhibit mid-IR luminosities consistent with this criterion, although the limited data makes it difficult to conclusively assess their nature. It is plausible, however, that many of them are starburst galaxies characterized by high rates of star formation and dust, with high fluxes of IR emission originating from the latter. R. L. Griffith et al. also concluded that none of the galaxies hosted ETIs whose nonstellar energy supplies were > 5.7 times higher than the available starlight in those galaxies.

Although we have discussed optical and IR wavelengths for identifying Type III ETIs, radio wavelengths (~ 1.4 GHz) are also well suited for this purpose. This is because robust empirical correlations have been documented for radio and IR fluxes, implying that sources with very high values of $q_{22} = \log\left(S_{22\,\mu\mathrm{m}}/S_{20\,\mathrm{cm}}\right)$—where $S_{22\,\mu\mathrm{m}}$ and $S_{20\,\mathrm{cm}}$ are the IR and radio

flux densities at wavelengths of 22 μm and 20 cm, respectively—could prove to be indicative of Type III ETIs. Garrett (2015) computed q_{22} for ninety-three sources identified in R. L. Griffith et al. as possessing high mid-IR emission, concluding that the majority (> 90 percent) of them do not meet the desired *anomalous* criterion of $q_{22} > 2$. Apart from being under- or overluminous in some regimes, the engineering activities of Type III ETIs may manifest themselves in other respects. For instance, consider the idealized scenario wherein they opt to build Stapledon-Dyson spheres only around stars below a particular luminosity threshold. As a result, stars of certain colors would be preferentially cloaked, thereby giving rise to galaxies characterized by unusual colors and luminosities in appropriate wavelength bands.

Hitherto, many of the proposed metrics rely on the galaxy comprising numerous Stapledon-Dyson spheres. The natural extension of this picture is to envision an entire galaxy cloaked in a galactic-scale Stapledon-Dyson sphere. Although the construction of these "blackboxes" appears extremely difficult, Lacki (2016) suggests that it could be undertaken via artificial dust to capture light. As the ensuing blackboxes would have blackbody temperatures only ~ 1 K higher than the cosmic microwave background (CMB) at most, they would be manifested as relatively large regions in all-sky searches at temperatures slightly higher than the CMB. The resultant patches, in principle, are detectable by analyzing the data collected by the *Planck* satellite. For a cosmological survey, Lacki derived an upper bound on the number of Type III ETIs that have built such blackboxes; the corresponding expression is given by

$$f_l \cdot f_i \cdot f_c \cdot f_{\rm III} \lesssim 7 \times 10^{-14}\, N_{\rm III} \left(\frac{\mathcal{V}_{\rm eff}}{\rm Gpc^3}\right) \left(\frac{L_{\rm III}}{\rm Myr}\right)^{-1} \left(\frac{N_e}{0.1}\right)^{-1} \times \left(\frac{\rho_{\rm SFR}}{0.05\, M_\odot\ {\rm yr}^{-1}\ {\rm Mpc}^{-3}}\right)^{-1} \left(\frac{\langle M_\star \rangle}{0.35\, M_\odot}\right), \tag{9.90}$$

where N_e, f_l, f_i and f_c, the parameters of the Drake equation, are delineated in Section 8.1. Note that $\rho_{\rm SFR}$ is the cosmic star-formation density normalized to its typical value at a redshift of ~ 0.5, while $\langle M_\star \rangle$ represents the average stellar mass. Here, $\mathcal{V}_{\rm eff}$ denotes the effective volume sampled by a survey, and $N_{\rm III}$ is the number of Type III ETIs found to build blackboxes within this

volume. The quantity L_{III} embodies the lifetime of a Type III ETI that must have a lower bound of $\sim 10^5$ yr because it has to span an entire galaxy, i.e., it cannot be shorter than the light travel time. The last factor on the left-hand side (f_{III}) is introduced as an extension of the Drake equation and signifies the fraction of all ETIs that go on to achieve Type III status. The crucial point worth appreciating with respect to the above inequality is that very low values of the right-hand side do *not* automatically imply that one or more of f_l, f_i, and f_c must approach zero. To put it differently, it is mathematically plausible that only $f_{\mathrm{III}} \to 0$ for certain reasons (e.g., sustainability) while the other factors are not excessively small, thereby ensuring that the inequality could still hold true.

9.6 THE RELATIVE PROSPECTS FOR DETECTING ETIs

> If you do not expect the unexpected, you will not find it; for it is hard to be sought out and difficult.
>
> —Heraclitus of Ephesus (ca. 535–475 BCE)

In Chapter 6, we explored the search for nontechnological life by means of biosignatures, whereas the theme in this chapter is the search for technological life via technosignatures. In both cases, the upcoming decades are expected to open up new avenues and enable access to historically unprecedented amounts of data in these fields. Hence, it is beneficial to develop a simple formalism to estimate the relative likelihood of *detecting* technological life with respect to nontechnological life through technosignatures and biosignatures, respectively. Naturally, it must be recognized that our results are dependent on (1) the survey methodologies employed and (2) the capabilities of current telescopes. In our subsequent analysis, we will closely mirror the framework and notation presented in Lingam and Loeb (2019e).

Regardless of whether we are contemplating technosignatures or biosignatures, the number of worlds with life that are detectable is determined by the product of the number of inhabited worlds in a given survey volume,[10] and the probability of actually detecting life by utilizing a suitable approach in conjunction with state-of-the-art observatories. When it comes

10. The *survey volume* represents the maximum possible volume that is covered by a particular telescope for detecting either biosignatures or technosignatures.

to biosignatures, therefore, we have

$$N_b \sim \mathcal{N}_b \cdot \mathcal{P}_b, \tag{9.91}$$

where N_b denotes the number of worlds with nontechnological life detectable through biosignatures, $\mathcal{N}_b$ is the number of worlds with nontechnological life within a survey volume, and $\mathcal{P}_b$ is the probability of life-detection via biosignatures for this particular survey. By the same token, for technosignatures, we are free to make use of

$$N_t \sim \mathcal{N}_t \cdot \mathcal{P}_t, \tag{9.92}$$

with $\mathcal{N}_t$ denoting the number of worlds with ETIs in the survey volume, $\mathcal{P}_t$ the probability of detecting technosignatures for a particular search strategy, and N_t the number of worlds with ETIs that are actually detectable through technosignatures. We are especially interested in the ratio δ_{tb} given by

$$\delta_{tb} = \frac{N_t}{N_b}, \tag{9.93}$$

because it measures the ratio of the number of worlds with life detectable via technosignatures and biosignatures; in consequence, it may be envisioned as the relative likelihood (RL) of detecting technological intelligence.

9.6.1 Inhabited worlds detectable through biosignatures

We will estimate N_b by using a Drake-type equation delineated in Seager (2018). As mentioned a few paragraphs earlier, the magnitude of N_b depends on two criteria, (1) and (2). We will suppose that transit spectroscopy is employed for identifying biosignatures using the James Webb Space Telescope (JWST). In principle, high-contrast direct imaging (outlined in Section 6.2.1) represents an attractive alternative, but the desired ground- and space-based telescopes are $\lesssim 10$ years in the future. Hence, we restrict ourselves to the identification of biosignature gases on M-dwarf exoplanets by means of transit spectroscopy.

With this set of assumptions, $\mathcal{N}_b$ is expressible as

$$\mathcal{N}_b \sim N_{\text{sur}} \cdot N_e \cdot f_l, \tag{9.94}$$

where N_{sur} denotes the number of stars spanned by a state-of-the-art telescope like JWST; N_e and f_l were defined in connection with the Drake equation in Section 8.1. The following point bears reiterating: N_e is not synonymous with the fraction of stars hosting Earth-sized planets in the habitable zone, because the latter is neither a necessary nor sufficient condition for guaranteeing habitability. Consequently, N_e is an unknown quantity given that our understanding of the requisite criteria for habitability is not complete. Next, $\mathcal{P}_b$ is given by

$$\mathcal{P}_b \sim f_q \cdot f_t \cdot f_{bd}, \tag{9.95}$$

where f_q represents the fraction of nonflaring (i.e., quiet) stars, f_t is the fraction of transiting planets observable by a telescope like JWST, and f_{bd} signifies the probability of detecting biosignatures, given that life is actually existent on a chosen planet. Our reason for highlighting nonflaring stars is motivated by the fact that active stars with frequent flares regularly perturb the concentrations of biosignature gases in the atmosphere (Grenfell 2017). However, when it comes to habitability *per se,* the situation is rendered more complex since flares give rise to both positive and negative ramifications, as outlined in Section 4.4. Lastly, f_{bd} is a measure of the datum that potentially habitable planets can produce both false negatives and false positives insofar as biosignatures are concerned. We furnished examples of these two phenomena in Chapter 6, owing to which we shall not tackle them further at this stage.

After combining (9.94) and (9.95), the final equation is essentially identical to the Biosignature Drake Equation proposed by Seager (2018).[11] In the event that planets at $\lesssim 30$ pc can be characterized by telescopes akin to JWST, we may work with $N_{\text{sur}} \sim 3 \times 10^4$ stars. We specify $f_q \sim 0.2$ and $f_t \sim 10^{-3}$ cautiously, since these parameters are fairly well constrained, but f_{bd} remains unknown. In our analysis, we will normalize this variable by a conservative fiducial value of 0.1. We have not provided estimates for N_e and f_l either because they will mostly drop out of our subsequent analysis.

However, before moving ahead, the reader will have observed that we have only dealt with prospects for detecting nontechnological life *beyond* our

11. https://www.cfa.harvard.edu/events/2013/postkepler/Exoplanets_in_the_Post_Kepler_Era/Program_files/Seager.pdf

Solar system. What about in situ and remote sensing searches for biosignatures within our Solar system? A number of targets are of astrobiological interest: Mars, icy worlds with subsurface oceans such as Europa and Enceladus (see Chapter 7), and even the clouds of Venus (Limaye et al. 2018). As the number of worlds surveyed by a given mission is typically unity—for example, the *Europa Clipper* mission—we note that the number of detectable worlds is expressible as

$$N_b' = f_l' \cdot f_{bd}', \tag{9.96}$$

implying that it constitutes the product of the target's likelihood of hosting life (assuming it already has habitable conditions) and the likelihood of detecting life, provided that it does exist. The use of the primed variables indicates that they are the Solar system counterparts of the unprimed variables specified earlier. Hence, the ratio of N_b' and N_b yields

$$\frac{N_b'}{N_b} \sim 1.7 \left(\frac{N_e}{0.1}\right)^{-1} \left(\frac{f_l'}{f_l}\right) \left(\frac{f_{bd}'}{f_{bd}}\right). \tag{9.97}$$

At this stage, it is worth recalling that the limitation of Drake-type equations is that they only admit point values in the classical formulation. Hence, although f_l and f_l' as well as f_{bd} and f_{bd}' are dependent on the specific choice of the habitable world, they must be replaced by their mean values. Seeing as how we have no knowledge of most of these variables a priori, it is not unreasonable to assign a uniform prior and therefore specify $f_l \sim f_l'$ and $f_{bd} \sim f_{bd}'$. Therefore, after simplification, we end up with

$$\frac{N_b'}{N_b} \sim 1.7 \left(\frac{N_e}{0.1}\right)^{-1}. \tag{9.98}$$

In other words, the key determining factor is the prevalence of worlds with *habitable* conditions in our Galaxy. If $N_e \ll 1$, then we see that N_b' is probably favored over N_b, whereas the converse is true when $N_e \gg 1$ is formally valid (albeit unlikely).

9.6.2 Inhabited worlds detectable through technosignatures

As elaborated throughout the chapter, the methods for identifying technosignatures are numerous. They can, however, be divided into two broad

and nonexclusive categories: electromagnetic technosignatures (e.g., radio and optical signals) and artifact technosignatures (e.g., megastructures). As a result, the survey volume also varies considerably depending on the chosen methodology. For now, we will restrict ourselves to searches for technosignatures within the Milky Way and defer the discussion of extragalactic SETI to Section 9.6.4.

To begin with, we will focus on the scenario wherein a conventional search for radio or optical signals is carried out. It is worth noting here that electromagnetic signals may be discernible due to the leakage of radiation while powering advanced propulsion systems, as explained in Section 9.5.2. Furthermore, it is appropriate to presume that the signals are being broadcast deliberately by putative ETIs, because even the largest human telescopes, such as the Square Kilometre Array (SKA), are predicted to have a very low probability of detecting accidental leakage emitted at human-level technology. In contrast, targeted signals broadcast by Arecibo are potentially detectable at a distance of $\lesssim 1$ kpc by receivers with similar capabilities. Benford et al. (2010) have estimated that it is not particularly difficult for humans to design beacons that would be visible throughout the Milky Way while incurring an expenditure of a few billion dollars (US).

Likewise, the detection threshold for continuous wave lasers is between a few kilowatts and megawatts (assuming a 10 m telescope) irrespective of the distance to the source, which falls within the bounds of current or upcoming human technology (Tellis & Marcy 2017). Instead, if optical pulses are employed, they will easily outshine the visible light of the Sun by > 4 orders of magnitude even with present-day lasers. Thus, viewed collectively, when it comes to radio and optical technosignatures entailing the deliberate broadcast of artificial signals, it is seemingly reasonable to presume that the survey volume comprises most of the Milky Way if the technological level of transmitting ETIs is more advanced than, or comparable to, humanity. In the future, the reach of these surveys may be considerably expanded if gravitational lenses are used to amplify electromagnetic signals.

For a technosignature survey encompassing the Milky Way, the value of $\mathcal{N}_t$ is determined by modifying the Drake equation as follows:

$$\mathcal{N}_t \sim \alpha_{\rm set} \cdot N_{MW} \cdot N_e \cdot f_l \cdot f_{ti} \cdot \frac{L}{t_{MW}}. \tag{9.99}$$

Note that $N_{MW} \sim 10^{11}$ is the number of stars in our Galaxy, and $t_{MW} \sim 10^{10}$ yr is the age of our Galaxy. In this equation, f_{ti} denotes the fraction of

life-bearing planets that eventually host ETIs with the capacity to partake in interstellar signaling. In the classical Drake equation, $f_{ti} = f_i \cdot f_c$, but we have collapsed the two factors into a single variable. The reason for doing so is threefold: (1) the definition of *intelligence,* interpreted sensu lato, is relatively nebulous with respect to *technological intelligence* (à la humans), (2) developing nontechnological intelligence might not be as difficult compared to technological intelligence (see Section 3.7), and (3) what really interests us is the potential for building technology that leaves distinctive imprints via technosignatures.

Lastly, we have introduced an extra factor $\alpha_{\rm set}$ to account for the fact that advanced ETIs need not exist solely on planets. They can spread outward and settle other worlds or even dwell in interstellar space; the latter appears to be particularly feasible if the technological species is post-biological.[12] Another possible phenomenon that is theoretically capable of contributing to $\alpha_{\rm set}$, and constitutes one of the chief subjects addressed in Chapter 10, is panspermia. In total, $\alpha_{\rm set}$ represents an averaged settlement factor akin to (8.9) that is present in some variants of the Drake equation. In writing down (9.99), we have implicitly assumed that the number of planets per star is around unity and that all stars are equally suitable for hosting life. The latter is not necessarily correct because several factors might degrade the habitability of M-dwarf exoplanets (Lingam & Loeb 2018f), as pointed out in Chapter 4. Yet, at the same time, it must be appreciated that M-dwarf exoplanets are ideally suited as outposts for ETIs since these stars are perhaps extremely long-lived, thereby providing plentiful energy for up to trillions of years.

We will henceforth employ the optimistic fiducial values of 0.1 for f_{ti} and $\sim 10^4$ yr for L. It is worth emphasizing the fact that our choices for both f_{ti} and L are by no means definitive, because we possess very little knowledge of these factors. There have been many proponents who have argued in favor of $f_{ti} \ll 1$ by drawing on the contingent nature of the evolutionary process on Earth, but others have appealed to evolutionary convergence instead and argued that $f_{ti} \sim 0.1\text{–}1$ holds true; we return to this point shortly hereafter. For instance, if we select $f_{ti} \sim 10^{-8}$, we end up with $\mathcal{N}_t \lesssim 1$ even for optimistic choices of the other parameters. In other words, the net effect

12. On the other hand, post-biological intelligences may find it more advantageous to send digital copies of themselves to other worlds, as opposed to undertaking physical travel (Scheffer 1994).

of $f_{ti} \ll 1$ is that searches for technological life are highly disfavored relative to searches for nontechnological life.

Next, the probability of detecting technosignatures ($\mathcal{P}_t$) is written as

$$\mathcal{P}_t \sim f_{cs} \cdot \varepsilon_d. \tag{9.100}$$

The first factor (f_{cs}) represents the fraction of current ETIs that are producing signals detectable by humans. If an ETI happens to be much more advanced than humanity, it may either refrain from signaling us or rely on channels that are not decipherable by us. The second factor (ε_d) quantifies the probability of intercepting an unambiguous signal from an ETI. In order to calculate ε_d, we will presume that the transmitter and receiver are cognizant of the optimal frequencies suitable for signaling; see Section 9.1.1.1 for a discussion of this nontrivial matter. If the signals are diffraction-limited beams and our search is undertaken over an interval denoted by Δt, we estimate ε_d using

$$\varepsilon_d \sim \Omega_r \cdot \Omega_t \cdot \frac{\Delta t}{\tau_{\mathrm{SNR}}}, \tag{9.101}$$

where Ω_r and Ω_t are the solid angles associated with the receiving and transmitting telescopes, respectively, whereas τ_{SNR} denotes the integration time required to achieve the minimum desired SNR.

It is apparent that (9.101) depends on many parameters that are dictated by the characteristics of the telescopes and mode of functioning. In consequence, assigning a point value to this variable is difficult. The magnitude of ε_d was investigated by Lubin (2016a) for optical SETI with survey parameters chosen to be commensurate with current human technology. Two different scenarios corresponding to "blind" and "intelligent" targeting were considered, and it was suggested that $\varepsilon_d \sim 10^{-6}$ for sources at distances of ~ 10 kpc in the case of blind targeting, whereas the case with intelligent targeting yielded $\varepsilon_d \sim 1$. Although isotropic beacons are very expensive from an energetic and economic standpoint, sufficiently advanced ETIs might set up such beacons, which would engender $\varepsilon_d \sim 1$ as a result.

An essential point to note is that $\varepsilon_d \propto \lambda^4$, implying that its value for radio SETI might be higher relative to optical SETI. Henceforth, we will adopt a fiducial normalization of 10^{-6} for ε_d, although we reiterate that its actual value could be much higher or lower. Lastly, we are confronted with the factor f_{cs} in (9.100). What fraction of ETIs wish to communicate? We have no true way of knowing, as the answer to this question is a function

of complex anthropological and sociological factors. The only example that we can draw on is our own. In 1974, the most powerful telescope on Earth (Arecibo) was deployed for transmitting an interstellar message. Hence, we will work with an optimistic fiducial estimate of unity for f_{cs}, but smaller values cannot be ruled out.

9.6.3 Implications of the relative likelihood function

We are now in a position to estimate δ_{tb} from (9.93) by utilizing the preceding results. Thus, our final expression is expressible as

$$\delta_{tb} \sim 0.02\, \alpha_{\text{set}} f_{cs} \left(\frac{f_{ti}}{0.1}\right) \left(\frac{L}{10^4\,\text{yr}}\right) \left(\frac{\varepsilon_d}{10^{-6}}\right) \left(\frac{f_{bd}}{0.1}\right)^{-1}. \tag{9.102}$$

One of the most striking points with respect to the above formula is that both N_e and f_l are absent. However, in lieu of these two factors, we have acquired two other variables (ε_d and f_{bd}), but the latter duo are conceivably easier to predict for given technologies and search strategies.

It bears repeating that one of the greatest advantages inherent to (9.102) is that we have eliminated both N_e and f_l. The latter, in particular, is extremely difficult to estimate because there are numerous unsolved mysteries as to how, when, and where life originated on our planet (see Chapter 2), to say nothing of other worlds. Although we have obviated the necessity of tackling N_e and f_l, we have also concomitantly sidestepped the issue of whether searches for biosignatures or technosignatures should be carried out in *absolute* terms; instead, we only deal with their relative merits herein. From a practical standpoint, this constitutes a pragmatic approach because the search for nontechnological life has come to be accepted as part of the scientific mainstream, whereas the search for technological life beyond Earth remains on the fringes.

It is worth unpacking the influence of f_l and f_{ti} on the quest for biosignatures and technosignatures. First, consider the scenario wherein $f_l \to 0$, which amounts to positing that life is extremely rare and abiogenesis is an incredibly difficult evolutionary step. In this event, we see that both $\mathcal{N}_b$ and $\mathcal{N}_t$ become infinitesimally small. Hence, mathematically speaking, the ratio δ_{tb} would remain indeterminate. Next, instead of supposing that the origin of life is the evolutionary bottleneck, let us transfer it to technological intelligence—that is, we will select $f_l \sim 1$ and $f_{ti} \to 0$. It follows that

$\delta_{tb} \to 0$, thereby implying that the search for technological life is heavily disfavored. Last, consider the idealized outcome in which $f_l \sim 1$ and $f_{ti} \sim 1$ as well as $L \gtrsim 10^5$ yr. In this regime, we see that $\delta_{tb} \sim 1$ is feasible, indicating that the searches for technosignatures and biosignatures merit roughly equal priority.

Returning to (9.102), it is clear that $\delta_{tb} \to 0$ whenever any of the parameters in the numerator approach zero. Of them, perhaps the variable subject to the most uncertainty is the likelihood of the advent of technological intelligence. One of the primary themes of Chapter 3 was that evolution may comprise a small ($\lesssim 10$) number of major events that have occurred only once (or a handful of times) on Earth, of which technological intelligence was the last of them in the current epoch. In particular, we argued that a total of four functionally analogous steps after abiogenesis might occur on a subset of inhabited worlds to yield technological intelligence. For the sake of illustrating our point, let us assign a probability of ~ 0.01 to each step, which does not come across as particularly small viewed in isolation. However, this choice would imply that $f_{ti} \sim (0.01)^4 = 10^{-8}$, which is clearly many orders of magnitude smaller than unity.

Yet, this does not necessarily imply that $f_{ti} \to 0$ is inevitable. Many scientists who argue in favor of a comparatively high value of f_{ti} have based their arguments on the premise that a large number of purportedly human traits—such as culture, high intelligence, theory of mind, tool making, and symbolic communication—have been documented, albeit controversially, to varying degrees in certain animals; for a more complete analysis of this issue, the reader may consult Section 3.7. Recent breakthroughs in evolutionary convergence have also been invoked to suggest that the emergence of technological intelligence has a high probability after the abiogenesis bottleneck has been successfully overcome (Ćirković 2018). Thus, in toto, the basic point that needs to be underscored is that we do not know whether $f_{ti} \to 0$ or $f_{ti} \sim 0.1$–1.

Next, we shall turn our attention to L and ε_d as there is a curiously common theme running through both of them. We will focus on L, because the same argument can be imported to ε_d with due modifications. Let us contemplate the hypothetical example wherein 1 percent of all ETIs are stable and long-lived with an average technological lifetime of $\sim 10^8$ yr, whereas the rest of them have a short lifetime of ~ 100 yr. After calculating the average lifetime, it is evident that $L \sim 10^6$ yr in this scenario. To generalize this result, there could be situations wherein the tail of the probability

distribution function of ETI lifetime might constitute the dominant contributor to the mean lifetime. This point has been appreciated since the 1960s (Shklovskii & Sagan 1966), thus representing a potential avenue by which L may be increased. Along similar lines, if a small fraction of the hypothesized ETIs were to opt for isotropic (as opposed to beamed) signals, despite the considerable power requirements involved, the average value of ε_d ought to be increased by a few orders of magnitude.

In contrast, if all ETIs are short-lived due to self-destruction—which was one of the most common explanations for Fermi's paradox during the Cold War (see Section 8.2.4)— it is clear that δ_{tb} will be correspondingly lowered. However, the destruction of ETIs does not automatically imply that all signs of their technology would be wiped out; it might be possible instead to search for artifacts of extinct ETIs, as explained later. Assessing the value of L, just as with f_{ti}, is remarkably difficult because our understanding of sociological and anthropological trajectories open to ETIs is highly limited. This theme is ubiquitous to many research topics within the domain of SETI, and one that we shall elaborate on at the end of the chapter.

We have yet to address the remaining three variables: f_{cs}, f_{bd}, and $\alpha_{\rm set}$. The former is evidently dependent on sociological factors that are hard to quantify a priori. From a qualitative standpoint, however, if the benefits of establishing contact with, or merely searching for, other technological species outweigh the risks, it seems conceivable that f_{cs} might approach unity once a particular ETI has attained a particular level of stability. Next, turning to f_{bd}, we presented several diagnostics for identifying false positives and negatives in Chapter 6, and a Bayesian framework developed by Catling et al. (2018) to handle the resultant uncertainty was described in Section 6.8. Hence, although $f_{bd} \sim 0.1$ comes across as being plausible, it appears relatively unlikely that $f_{bd} \ll 1$ would be valid.

Finally, we turn our attention to $\alpha_{\rm set}$, which represents the number of outposts established by ETIs per star. For our specific example, each outpost needs to possess the requisite technology to set up optical or radio beacons visible across the Milky Way. As our approach is analogous to the Drake equation, it bears mentioning that $\alpha_{\rm set}$ is the *average* number of sites settled. Hence, even if a small fraction of advanced ETIs were to have sent probes and settled a large number of sites, it is possible for $\alpha_{\rm set}$ to become much higher than unity. However, this scenario would bring us into conflict with Fermi's paradox, since we would need to explain the nondetection of probes

or their signals. Yet, as emphasized in Section 8.2.2.1, surveys of the Solar neighborhood have been few and far between, owing to which it is difficult to conclusively claim that interstellar probes in the Solar neighborhood do not exist.

In general, $\alpha_{\rm set}$ is expected to correlate with L because a longer ETI lifetime translates to more opportunities for interstellar travel and settlements. Suppose that the rate at which new sites are settled, or spacecraft with powerful enough radio or optical transmitters to be detectable by humans are launched, is constant; we denote it by $\Lambda_{\rm site}$. For this toy model, the maximum number of sites settled by a typical ETI corresponds to $\alpha_{\rm set} = 1 + \Lambda_{\rm site} L$, where $\alpha_{\rm set} \to 1$ when $L \to 0$, since the multiplicity has a lower bound of unity—that is, the home planet of a given ETI. Although the exact value of $\Lambda_{\rm site}$ remains unknown, a fairly conservative choice is $\Lambda_{\rm site} \sim 10^{-4}$/yr (Wright, Mullan, et al. 2014). For this ansatz, we arrive at

$$\alpha_{\rm site} \sim 1 + \left(\frac{L}{10^4 \, \rm yr} \right). \tag{9.103}$$

Hitherto, we have retained all possible factors that are not well constrained. However, on the basis of prior arguments, it seems reasonable to contend that the maximum variability is encoded in $\alpha_{\rm set}$, f_{ti}, and L compared to the other parameters in (9.102). Hence, after applying (9.103), the following formula constitutes a simplified version of (9.102) that retains sufficient complexity and the essential features:

$$\delta_{tb} \sim 0.02 \left(\frac{f_{ti}}{0.1} \right) \left(\frac{L}{10^4 \, \rm yr} \right) \left[1 + \left(\frac{L}{10^4 \, \rm yr} \right) \right]. \tag{9.104}$$

9.6.4 Artifact SETI and the relative likelihood of detection

Hitherto, we have restricted ourselves to only electromagnetic technosignatures, but we saw in Section 9.5 that the search for extraterrestrial technological artifacts is gaining in prominence. At this juncture, we shall differentiate between technological artifacts that are within the scope of Type I (and perhaps Type II) ETIs on the Kardashev Scale and those that fall under the domain of Type III ETIs. The reason for introducing this artificial distinction is because Type I ETIs are expected to undertake comparatively localized niche construction (i.e., on the scale of planetary systems), whereas

Type III ETIs can partake in astroengineering on galactic scales. As a consequence, the required survey volume to identify Type III ETIs is very different from that of Type I and possibly Type II ETIs.

Let us begin by considering the search for artifacts produced by ETIs that are somewhat more advanced than our own, i.e., we work with $\mathcal{K}_{\rm ETI} \sim 1$. Examples of such artifacts may include dense satellite belts, high levels of industrial pollution, and large-scale photovoltaic arrays. We presume that the putative artifacts are detectable while concomitantly carrying out searches for biosignatures. As we are dealing with searches using current-level technology on par with the JWST, we have

$$\mathcal{N}_t = N_{\rm sur} \cdot N_e \cdot f_l \cdot f_{ti} \cdot f_a, \tag{9.105}$$

where f_a denotes the fraction of the stellar lifetime ($\sim 10^{10}$ yr) that these artifacts are detectable. Upon comparing the above equation with (9.99), two major differences are manifested.

First, we have specified $\alpha_{\rm site} \sim 1$ because the hypothetical ETIs we investigate are not considerably more advanced than humans, implying that the settlement factor might be close to unity. Second, in lieu of L/t_{MW}, we have introduced f_a. This replacement was motivated by the fact that ETIs do not have to exist currently in order for their artifacts to still be detectable. In quantitative terms, it amounts to contending that $f_a \gtrsim L/t_{MW}$. The longevity of surface-based artifacts is dictated by a diverse array of geological processes, including the likes of plate tectonics, volcanism, and erosion by winds, glaciers, and flowing water. Hence, all traces of technology on the surface may be eliminated on a timescale of $\gtrsim 10^6$ yr (Schmidt & Frank 2019). In contrast, the longevity of space-based artifacts is primarily a function of their dynamical stability, suggesting that certain artifacts could potentially survive over Gyt timescales. We will adopt a normalization of 10^{-5} for f_a, which does not appear unreasonable at first sight.

Next, we need to assess the probability of detection ($\mathcal{P}_t$), which is expressed as

$$\mathcal{P}_t = f_q \cdot f_t \tag{9.106}$$

and therefore closely resembles (9.95); this is not surprising since the search for artifact technosignatures (of this technological level) would piggyback onto the search for biosignatures. We have opted to err on the side of caution and retain the factor of f_q, because frequent flares and superflares might not

be conducive to either complex surface-based life or technological artifacts; the underlying reason for the latter is that coronal mass ejections play havoc with unprotected electrical devices (Lingam & Loeb 2017e; Loper 2019).

The key difference between (9.106) and (9.95) is that the latter contains an extra factor of f_{bd} to account for the possibility of false positives and negatives. The reason for dropping the analog of f_{bd} stems from the conventional notion that technosignatures have a much lower likelihood of false positives. For instance, the most unambiguous signature indicative of life on Earth as observed by the *Galileo* spacecraft during a flyby mission was the detection of narrow-band radio signals. However, when it comes to advanced ETIs, discerning technosignatures from biosignatures could prove to be very difficult (see Section 9.5.4.1), implying that incorporating an extra variable in this formalism is warranted. In an intriguing publication, Raup (1992) mused that nonintelligent organisms might evolve means of communicating with one another at radio wavelengths, in which event the ensuing signals run the risk of being misinterpreted as signposts of technological activity.

Using the preceding equations, the relative likelihood turns out to be

$$\delta_{tb} \sim 10^{-5} \left(\frac{f_{ti}}{0.1} \right) \left(\frac{f_a}{10^{-5}} \right) \left(\frac{f_{bd}}{0.1} \right)^{-1}. \tag{9.107}$$

Upon comparing (9.107) with (9.104), the latter is ostensibly higher by about three orders of magnitude. However, it is worth reiterating that there are many unknown variables in both formulae, and the ratio may be flipped in some instances. Nonetheless, if we accept this result at face value, searching for electromagnetic signals might be more productive relative to artifacts. In mathematical terms, the primary reason for the dominance of electromagnetic technosignatures is attributable to the potential search volume being much larger, whereas detecting planet-based artifacts is limited by the extent of transit spectroscopy and direct imaging achievable with current telescopes. Yet, what our analysis fails to reveal is the economic advantages linked with artifact SETI, since it can piggyback onto biosignature searches, while the pursuit of electromagnetic technosignatures requires dedicated observation time and facilities for the most part.

Lastly, we turn our attention to Type III ETIs. We do not explicitly tackle Type II ETIs, but the results should fall in between Type I and Type III ETIs. One of the most striking aspects of Type III ETIs is that the search volume is so much larger. For example, the $\hat{G}$ survey for Type III ETIs

described in R. L. Griffith et al. (2015) and in Section 9.5.9 surveyed $\sim 10^5$ galaxies, which amounts to $\sim 10^{16}$ stars. Hence, it is tempting to conclude that δ_{tb} would be much higher. However, this optimism needs to be counterbalanced by two additional considerations. First, not all ETIs are guaranteed to eventually reach the Type III stage. Hence, we must work with the factor $f_{\rm III}$ that was first introduced in (9.90). Second, even if certain ETIs do attain this level, the subsequent lifetime of these entities and their artifacts remains unknown. After due simplification, we find that the analog of (9.107) for a survey of Type III ETIs comprising $\sim 10^{16}$ stars is given by

$$\delta_{tb} \sim 33.3 \left(\frac{f_{ti}}{0.1}\right) \left(\frac{f_{\rm III}}{10^{-6}}\right) \left(\frac{f_a}{10^{-4}}\right) \left(\frac{f_{bd}}{0.1}\right)^{-1}, \tag{9.108}$$

where the choice of normalization for $f_{\rm III}$ is based on the fact that arguably none of the $\sim 10^5$ galaxies in the $\hat{\rm G}$ survey exhibits unambiguous evidence of anomalously high waste heat production indicative of Type III ETIs. The fiducial value for f_a is chosen in light of causality constraints imposed on Type III ETIs; see the discussion following (9.90). Although (9.108) appears to be much higher than unity, we note that this is a consequence of our normalization scheme. If $f_{\rm III}$ is extremely small, which is potentially feasible on grounds of technological sustainability, $\delta_{tb} \ll 1$ is mathematically allowed.

9.6.5 Implications for funding different search strategies

An auxiliary advantage of this formalism is that we can employ it to gauge what level of funding should be assigned to different surveys. For the sake of simplicity, we shall focus on the United States (and use its system of currency for dollar amounts), but a similar analysis is feasible for other countries.

To begin with, we observe that the total astrobiology budget over the next decade is estimated to be $1 billion per year (Kite et al. 2018; Lingam & Loeb 2019e). We shall employ the simple ansatz wherein the amount of funding allocated to a particular search strategy is linearly proportional to the number of worlds that are detectable by this approach in principle; naturally, the predicted number will depend on the values we specify for the numerous free parameters in our model. This ansatz is potentially reasonable in qualitative terms since it assigns a higher priority and budget to surveys that have a higher probability of success. For starters, in Section 9.6.1, we saw that the

number of inhabited worlds detectable by telescopes like JWST might be comparable to that identifiable by astrobiological missions within our Solar system. Therefore, we shall work with a putative budget of \$500 million per year for detecting biosignatures outside our Solar system.

Our basic goal is to compare the efficacy of searching for technosignatures versus biosignatures beyond our Solar system. As per the above ansatz, the annual funding for technosignatures is given by $C_{\rm tech} \sim 0.5\,\delta_{tb}$ billion per year. There are a number of expressions for δ_{tb}, but we shall consider electromagnetic signatures henceforth for the sake of simplicity; the treatment can be easily generalized to artifact SETI. We are free to employ (9.102) or (9.104), but the latter is chosen because it retains the essential parameters f_{ti} and L and has a simpler form. Therefore, the magnitude of $C_{\rm tech}$ becomes

$$C_{\rm tech} \sim \$10\,{\rm million/yr}\left(\frac{f_{ti}}{0.1}\right)\left(\frac{L}{10^4\,{\rm yr}}\right)\left[1+\left(\frac{L}{10^4\,{\rm yr}}\right)\right]. \qquad (9.109)$$

A couple of aspects are worth highlighting with regard to the above formula. First, if either f_{ti} or L is very small, the justifiable budget for detecting technosignatures also experiences a steep decline. Second, if the fiducial values specified for f_{ti} and L are accurate, a total funding of \$100 million per decade could be allotted toward searching for technosignatures. Interestingly, this total is coincident with the budget recently allocated to the Breakthrough Listen initiative.

However, the above expression ignores a subtle point that is overlooked in many analyses. To elucidate this point, let us contemplate the following two scenarios: N worlds with microbial life were detected via biosignatures, and N worlds with technological life were discovered through technosignatures. Are the two results truly equivalent? The answer to this question depends on what we are seeking. If we merely concern ourselves with detecting extraterrestrial life in one form or another, the two outcomes are equal. On the other hand, there is a genuine case to be made that the *impact* of discovering ETIs is not identical to that of finding microbes and other forms of nontechnological life, even if they are complex multicellular organisms analogous to animals and plants. To a large extent, this originates from our anthropocentric bias: we are probably predisposed to rate technological intelligence as more important because we perceive this trait (rightly or wrongly) as being unique to human beings on Earth.

The impacts of detecting ETI are manifold (Dick 2018). They include not only potentially positive ramifications, such as the acquisition of novel scientific and technological knowledge, but also possible negative consequences (e.g., global existential crises). Although the range of outcomes spans a spectrum, it is likely that the net result will not be neutral—that is, the overall effect is probably considerable either way. On account of the above reasons, it is instructive to multiply the right-hand side of (9.109) with an extra factor ($\mathcal{I}_{\rm tech}$) that encapsulates the impact of discovering technological intelligence in comparison to nontechnological life. We need to make a choice for the fiducial value of $\mathcal{I}_{\rm tech}$. At the minimum, it appears reasonable to conjecture that it ought to be $\mathcal{O}(10)$ because of the impact on human culture, religion, and science. Thus, we find that $C_{\rm tech}$ is transformed into

$$C_{\rm tech} \sim \$100\,\text{million/yr} \left(\frac{\mathcal{I}_{\rm tech}}{10}\right)\left(\frac{f_{ti}}{0.1}\right)\left(\frac{L}{10^4\,\text{yr}}\right)\left[1+\left(\frac{L}{10^4\,\text{yr}}\right)\right]. \tag{9.110}$$

Alternatively, we can make use of the unorthodox ansatz $\mathcal{I}_{\rm tech} \sim 1/f_{ti}$, implying that the relative impact is inversely proportional to the prevalence of technological intelligence. Lingam and Loeb (2019e) motivated this scaling by drawing on classical economic theory: the law of downward-sloping demand loosely states that the price and the quantity of elastic commodities are inversely proportional to each other. Upon adopting this formulation for $\mathcal{I}_{\rm tech}$, we end up with

$$C_{\rm tech} \sim \$100\,\text{million/yr} \left(\frac{L}{10^4\,\text{yr}}\right)\left[1+\left(\frac{L}{10^4\,\text{yr}}\right)\right]. \tag{9.111}$$

The above equation is ideal from the standpoint that f_{ti} has dropped out, since it was one of the parameters subject to the most uncertainty.

If we adopt a lower bound of $L \sim 100$ yr, based on the duration over which human technology gave rise to detectable outputs, we arrive at $C_{\rm tech} \sim \$1$ million per year. This result is somewhat stable because it depends only on specifying a particular value for L and not f_{ti}. Similarly, if we apply the same bound to (9.110) in conjunction with $\mathcal{I}_{\rm tech} \sim 10$ and $f_{ti} \sim 0.1$, we end up with the same result. Therefore, on the basis of our analysis, it is potentially reasonable to conclude that the search for technosignatures merits a minimum budget of $1 million per year. In fact, since we

have chosen a fairly conservative value of L in (9.111) to obtain \$1 million per year, a budget of \$10 million per year is also not unreasonable.

Before rounding off this analysis of the relative merits of searching for technosignatures and biosignatures, some caveats are in order. Our formalism closely mirrors that of the Drake equation and consequently inherits many of the disadvantages associated with it. First, it depends on parameters like f_{ti} and L, which are poorly understood and constrained by theoretical studies. Second, the variables in δ_{tb} should be placed on a statistical footing, i.e., we must work with their probability distribution functions instead. Third, none of these parameters exhibits spatial and temporal variations as they are effectively treated as constants. In spite of these pertinent caveats, our approach might pave the way toward developing a heuristic framework for gauging the relative likelihood of success for different life-detection paradigms and the amount of funding that they would merit.

9.7 CONCLUSION

> By day, a mysterious wood, near the town,
> breathes out cherry, a cherry perfume.
> By night, on July's sky, deep, and transparent,
> new constellations are thrown.
>
> And something miraculous will come
> close to the darkness and ruin,
> something no-one, no-one, has known,
> though we've longed for it since we were children.
>
> —Anna Akhmatova, "Everything" (1921)

After its early heyday during the 1960s and 1970s, the field of SETI witnessed a downturn in scientific funding and legitimacy, and consequently found itself consigned to the backwater of mainstream science. Apart from these issues, it was constrained by its near-exclusive focus on the search for radio and optical signals, with a few notable exceptions. There are, however, genuinely promising signs that this state of affairs is experiencing an upsurge this decade. From the observational and economic standpoint the Breakthrough Listen initiative is expected to constitute the largest survey for technosignatures undertaken to date. There have also been concomitant glimmers of interest evinced by NASA and the United States Congress with regard to hunting for technosignatures. Thus, in the words of

Anna Akhmatova, perhaps "something miraculous will come" of the SETI enterprise in future decades.

From a theoretical perspective, many achievements are worth mentioning. First, the search for technological life clearly requires an understanding of planetary habitability and the conditions under which life originates, as seen from the factors N_e and f_l in the Drake equation. Much progress has been achieved in these areas, as summarized in the previous chapters of this book. Second, we have witnessed a blossoming of technosignature candidates beyond electromagnetic signals, many of which are described in Section 9.5, ranging from industrial pollutants in the atmosphere to localized heat islands resulting from urban conglomerations (i.e., cities). Lastly, we are beginning to witness the emergence of mathematical models that seek to predict the trajectories accessible to ETIs given basic energetic and raw material constraints and to explore the ensuing ramifications.

Thus, it is not an exaggeration to claim that the future of SETI looks brighter than before. Yet, at the same time, it is important to avoid getting carried away and appreciate the vast distance that we must travel before SETI can be said to enter its mature phase. From the viewpoint of observations and search strategies, several conundrums ought to be addressed (Sheikh 2020). Virtually all of the targeted searches for electromagnetic signals undertaken hitherto have operated on the premise that ETIs are situated around host stars. However, this assumption is questionable because advanced ETIs, especially of the post-biological kind, may spend most of their lives traversing the empty reaches of space. We will therefore need to broaden our search to encompass such possibilities. Moreover, most SETI searches are still directed toward finding electromagnetic signals. Although we are now seeing more surveys for Kardashev Type II and Type III ETIs via megastructures, both the frequency and reach of these surveys must be extended.

When it comes to modeling, there is arguably even more ground to cover. Our current understanding of the last two factors in the Drake equation—to wit, $f_{ti} = f_i \cdot f_c$ and L—remains at a very rudimentary level. In order to carefully assess the likelihood of the emergence of technological intelligence, it will be necessary to integrate insights from evolutionary biology, neuroscience, biophysics, and genomics. Likewise, anent L, we will need to synthesize perspectives from archaeology, linguistics, anthropology, and sociology to comprehend the cultural and biological evolution of ETIs

and whether any universal laws of complex societies are derivable thence. Finally, in the event that ETIs do exist and are currently signaling us, a host of unanswered questions still confront us (Lem 2020). What are the prospects for communication between ETIs with divergent evolutionary histories? What modalities of signaling are the most feasible across interstellar distances, and how can we discern deliberate signals from unintentional ones (i.e., leakage)? Answering these questions will evidently require us to draw on not only linguistics and information theory[13] but also nonhuman animal communication and ethology.

In the event that we detect potential technosignatures, it will be necessary to meticulously analyze them for false positives, take the environmental context into consideration, and thereby estimate the likelihood that the signals are genuinely indicative of ETIs. The Bayesian approach described in Section 6.8 is well suited for this purpose. The chief difference is that wherever *life* is alluded to, it ought to be replaced with *technological intelligence*; furthermore, the corresponding probabilities must be adjusted accordingly. By extending this logic, we can duly import and adapt the classification scheme for the identification of nontechnological life depicted in Table 6.1 to gauge the robustness of a given technosignature candidate. In passing, we also wish to note that Bayesian methodologies have been developed to infer the mean number of radio signals crossing the Earth in the case of detection or nondetection (Grimaldi & Marcy 2018).

Hence, it is apparent that SETI is not merely an interdisciplinary endeavor but also an intrinsically transdisciplinary one, which becomes evident upon inspecting Figure 9.9. The pressing need for a truly multidisciplinary approach toward SETI was expounded in Race et al. (2012) and Cabrol (2016). Yet, at first glimpse, it may seem as though the construction of such an imposing SETI superstructure is not only unnecessary but also foolhardy. Many scientists would probably argue, on the basis of the sole example of Earth, that technological intelligence is exceptionally rare and that we are likely the only technological species in the Milky Way or even in the Universe. However, regardless of whether this statement is true or false, what cannot be denied is that abstaining from searching for

13. The steps taken toward developing a bona fide theory of universal semantic communication and goal-directed communication, reviewed in Juba (2011) and Goldreich et al. (2014), are particularly invaluable for understanding the nature, content, and modality of putative signals from ETIs.

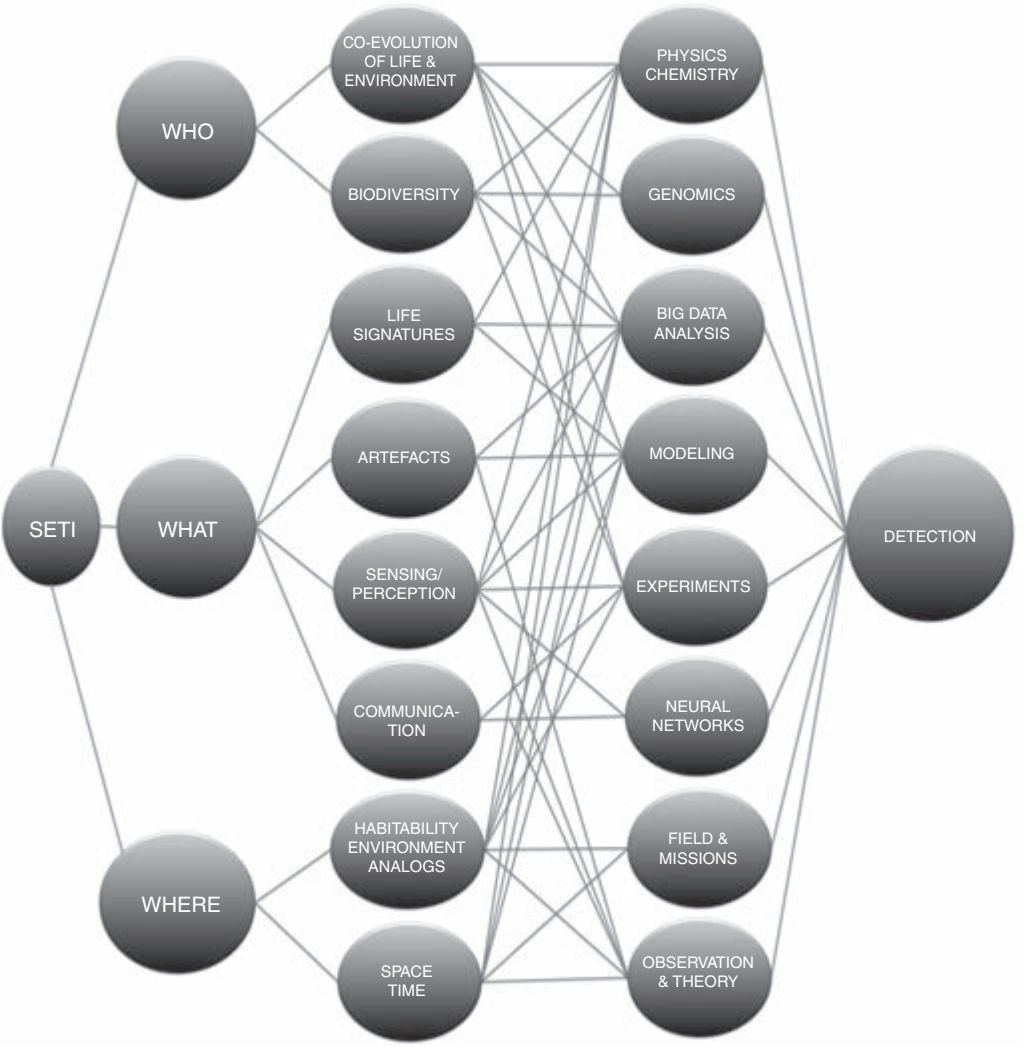


Figure 9.9 An integrated approach is essential for the detection of ETIs. This strategy requires us to not only address *how and where* we search for ETIs but also take into consideration *who or what* ETIs are; the former and latter are conceivably intertwined since they coevolve together. In turn, determining the *who or what* requires the synthesis of knowledge from fields as disparate as planetary science, molecular biology, ecology, and data science. (*Source:* Nathalie A. Cabrol [2016], Alien mindscapes—A perspective on the search for extraterrestrial intelligence, *Astrobiology* 16[9]: 661–676, fig. 1.)

technosignatures is a self-fulfilling prophecy: without seeking signatures of ETIs, we cannot hope to find them, except in the unlikely event that they deliberately and manifestly reveal their presence.

Of more significance, arguably, is the point that the SETI campaign is beneficial irrespective of whether we detect ETIs or not. If we do discover evidence of ETIs, the consequences are likely to be profound. Let us suppose, for instance, that we find signatures of an extinct ETI. Ideally, although not perhaps in actuality, this revelation may stimulate humanity to commence addressing present and future environmental catastrophes and eventually transition to a more sustainable and egalitarian future. Now, suppose that we consider the diametrically opposite (i.e., worst-case) scenario, wherein we fail to discover ETIs after decades or centuries of extensive surveys. Even in this case, at least two major positives would emerge. First, the observational techniques and theoretical models applied to this field can be gainfully transferred elsewhere. Second, a deeper recognition of the rarity of technological intelligence in the Universe might increase humanity's reverence for Earth's rich biosphere and underscore the necessity for assuming responsible stewardship of our planet.

Chapter 10

THE PROPAGATION OF LIFE IN THE UNIVERSE

> Though I do not believe that a plant will spring up where no seed has been, I have great faith in a seed. . . . Convince me that you have a seed there, and I am prepared to expect wonders.
>
> —Henry David Thoreau, *The Succession of Forest Trees and Wild Apples*

The Milky Way is around 10 billion years old. In this span of cosmic time, how has the distribution of life changed in our Galaxy and, by extension, in our Universe? As far as we know, there are only two distinct modes by which new worlds may be populated with life. The first is the most conventional and widely accepted paradigm, and the one that we have implicitly adopted throughout this book, for the most part. It involves the independent emergence of life on different worlds through a transition from prebiotic chemistry to biochemistry. As elucidated in Chapter 2, a multitude of unsolved issues—such as how and where this transition occurred—still persist.

If we look past the orthodox concept of abiogenesis transpiring independently on worlds over time, an intriguing heterodox possibility emerges. Perhaps life had originated on a particular world and hopped to a nearby world, and from thence to yet one more, and so on. This very premise, namely, that life could be transferred from one world to another, goes by the name of *panspermia,* with an intriguing history spanning thousands of years. Panspermia has mostly hovered on the fringes of respectable science, interspersed with occasional forays into the mainstream, as this mechanism is customarily deemed to have a low probability of occurrence owing

to multiple hazards encountered during the migration process. Moreover, several scientists have contended that panspermia does not explain the origin-of-life question but merely defers it to a different world. While this statement is correct sensu stricto, the real importance of panspermia may lie in its ability to boost the number of inhabited worlds. To be specific, many sites are potentially accessible for the initiation of life courtesy of this mechanism; therefore the widespread emergence of life might become more feasible in a statistical sense.

At this stage, it is necessary to distinguish between panspermia and pseudo-panspermia. By *pseudo-panspermia,* we will henceforth refer to the transfer of essential prebiotic molecules that comprise the building blocks of life from one world to another. While panspermia is hampered by the lack of tangible evidence, pseudo-panspermia has much more credibility on both empirical and theoretical grounds. In our Solar system, the analysis of meteorites has established that they are populated by a diverse array of prebiotic compounds. Meteorites known as carbonaceous chondrites are particularly rich in complex organics. The most famous among them is arguably the Murchison meteorite, which impacted the Earth in 1969. To date, roughly one hundred amino acids have been discovered within its interior (Elsila et al. 2016; Koga & Naraoka 2017); in contrast, only ~ 20 protein-building (proteinogenic) amino acids are prevalent on our planet. Likewise, the study of multiple carbonaceous chondrites has revealed the existence of nucleobase analogs (e.g., purine) that do not appear naturally in biological processes on Earth (Callahan et al. 2011; Burton et al. 2012). On the whole, it seems plausible that prebiotic synthesis can and does occur in the parent bodies of meteorites, thus permitting the exchange of these materials across worlds.

Hitherto, our allusions to panspermia have been predicated on the idea that the transfer of life across systems occurs via natural causes. In contrast, the existence of extraterrestrial technological intelligences (ETIs) is capable of effectuating novel avenues by which life may spread from one world to another. For example, ETIs could decide to deliberately seed other worlds with life. Alternatively, they might directly travel to other planets, terraform them—perhaps by deploying microbes (Lopez et al. 2019)—to meet their requirements and establish settlements. In both these instances, the capability to take part in interstellar travel (preferably at high speeds) through appropriate propulsion systems is necessary by definition.

When viewed in this spirit, we see that panspermia via natural carriers such as rocks or dust grains and interstellar travel by ETIs (or even humans) are two sides of the same coin. They constitute alternative paths toward a Universe that is increasingly populated by life. In this chapter, we have therefore opted to club panspermia and interstellar travel together. In case readers merely wish to delve into the former, they can truncate their reading at Section 10.3. On the other hand, if they are primarily interested in methods of interstellar travel, the relevant content commences from Section 10.4.

10.1 HISTORY AND PRINCIPLES OF PANSPERMIA

We will begin our voyage with a brief historical account and the general principles involved in the transfer of life via panspermia.

10.1.1 The history of panspermia

The long and storied history of panspermia spans more than two thousand years. The origin of the panspermia hypothesis is typically attributed to Greek philosophers such as Anaximander and Thales, but most notably Anaxagoras, who envisioned the Universe as being peopled with seeds (*spermata*) in the fifth century BCE.[1] However, a careful analysis of various mythologies reveals that proto-panspermia hypotheses were replete in the ancient cultures of the world, ranging from the "pointed seeds" broached in the Old Kingdom of ancient Egypt to the "universal germs" alluded to in the Vedic period of India. The reader should consult Temple (2007) for a detailed account of how proto-panspermia hypotheses evolved in myriad regions of the world at different times.

The first steps toward the current theory of panspermia were taken in the nineteenth century after the Darwinian revolution (Kamminga 1982; Demets 2012), although there were noteworthy predecessors such as the diplomat and natural philosopher Benoit de Maillet and Jacob Berzelius,

1. The excerpt from Henry David Thoreau at the beginning of the chapter is particularly apt, as it underscores both the biological potential and robustness of seeds (Tepfer & Leach 2006).

one of the founders of modern chemistry (Hollinger 2016). In 1865, to explain the origin of life, physicist Hermann Richter proposed that organic life had always existed in the Universe and that germs (termed *cosmozoa*) could travel across space in meteorites. By doing so, he anticipated many of the ideas that were subsequently espoused by Sir William Thomson (better known as Lord Kelvin) in his famous 1871 address to the British Association for the Advancement of Science. During the course of this lecture, Lord Kelvin asserted that (Thomson 1894, p. 202):

> we must regard it as probable in the highest degree that there are countless seed-bearing meteoric stones moving about through space. If at the present instance no life existed upon this Earth, one such stone falling upon it might, by what we blindly call *natural* causes, lead to its becoming covered with vegetation. . . . The hypothesis that some life has actually originated on this Earth through moss-grown fragments from the ruins of another world may seem wild and visionary; all I maintain is that it is not unscientific, and cannot rightly be said to be improbable.

It is therefore apparent that Lord Kelvin's reputation among the first proponents of lithopanspermia (i.e., the transfer of life between worlds via rocks) is well deserved. The latter half of the nineteenth century witnessed other notable scientists in this tradition, such as bacteriologist Ferdinand Cohn, but the work of noted physiologist and physicist Hermann von Helmholtz merits a special mention. A few months *prior* to Kelvin's more celebrated address, von Helmholtz delivered two remarkable public lectures wherein he highlighted the existence of hydrocarbons in meteorites and comets and consequently articulated the feasibility of lithopanspermia. His stance is well summarized by the following lines (1893, pp. 192–193):

> But carbon is the element, which is characteristic of organic compounds, from which living bodies are built up. Who knows whether these bodies, which everywhere swarm through space, do not scatter germs of life wherever there is a new world, which has become capable of giving a dwelling place to organic bodies? And this life we might perhaps consider as allied to ours in its primitive germ, however different might be the form which it would assume in adapting itself to its new dwelling place.

At the dawn of the twentieth century, the Nobel laureate Svante Arrhenius authored an insightful book, *Worlds in the Making,* building on earlier

papers, wherein he developed one of the first genuinely modern versions of panspermia (Arrhenius 1908). Although Arrhenius concurred with Richter and Kelvin that the notion of panspermia was definitely plausible, his model differed from that of Kelvin in certain vital respects. Arrhenius regarded the heat generated by atmospheric friction during meteorite entry as a potential showstopper and contended that collisions of two sizable bodies to produce fragments (the mechanism advanced by Kelvin) was very improbable. Instead, he argued that radiation pressure exerted by the host star could readily propel microscopic organisms or their spores across great distances, provided that they were small enough;[2] this hypothesis was christened *radiopanspermia*. Recently, Berera (2017) proposed that collisions with dust particles can facilitate the escape and dispersal of life-bearing particles from planetary atmospheres; simply put, space dust plays an operative role analogous to radiation in this instance.

It is a worthwhile exercise to assess the requirements for radiopanspermia. Let us model the microscopic object (either organism or spore) as a solid sphere with mass m_0, radius r_0, and density ρ_0. For an object with low reflectivity, by invoking (10.104) derived in Section 10.5.2 when tackling spacecraft known as light sails, the acceleration exerted by radiation pressure (a') is

$$a' \approx \frac{W_t}{m_0 c}, \tag{10.1}$$

where W_t is the power incident upon the microbe. If a star has a luminosity of $L_\star$ and the object is situated at a distance ℓ, we have $W_t \approx L_\star r_0^2/(2\ell)^2$. However, while the radiation pressure exerts an outward force, gravity exerts an inward force. We are not dealing with objects in orbit around the star, as otherwise the centrifugal force would also have to be accounted for. Therefore, the net acceleration a_n, takes on the form

$$a_n \approx \frac{GM_\star}{\ell^2}\left(\alpha_0 - 1\right), \tag{10.2}$$

where $M_\star$ is the stellar mass and α_0 is defined as

2. We shall anoint *spores* as shorthand notation for the functional analogs of bacterial endospores. The latter are typically dormant, resilient, and pared-down bacterial structures formed in response to harsh environmental conditions. Bacterial endospores exhibit high resistance to extreme heat, radiation, and desiccation, to name a few. Yet, in spite of the accompanying benefits, endospore-forming bacteria appear to be restricted to the phylum *Firmicutes* (Galperin 2013).

$$\alpha_0 = \frac{L_\star r_0^2}{4GM_\star m_0 c}. \tag{10.3}$$

Generalizing this result, we note that r_0^2/m_0 would need to be replaced by the ratio of the cross-sectional area to that of the mass. Clearly, a higher value of this quantity would lead to an enhancement of α_0. From an inspection of (10.2), we see that a net outward acceleration is feasible only when $\alpha_0 > 1$. Substituting $m_0 = 4\pi\rho_0 r_0^3/3$ into (10.3) and using this inequality yields

$$r_0 < 0.6\,\mu\text{m}\left(\frac{L_\star}{L_\odot}\right)\left(\frac{M_\star}{M_\odot}\right)^{-1}\left(\frac{\rho_0}{10^3\,\text{kg/m}^3}\right)^{-1}. \tag{10.4}$$

To begin with, the right-hand side is an increasing function of $M_\star$, implying that more massive stars enable larger particles to be transported via radiation pressure. Many microbes are characterized by a size of ~ 1 μm, which is commensurate with the upper limit derived above. Furthermore, microbes such as the bacterium *Pelagibacter* and the archaeon *Nanoarchaeum equitans* have volumes that are ~ 2 orders of magnitude smaller than the above limit. In case the microscopic object being propelled by radiation pressure resembles a spherical shell with thickness Δr_0, repeating the same calculation yields

$$\Delta r_0 < 0.2\,\mu\text{m}\left(\frac{L_\star}{L_\odot}\right)\left(\frac{M_\star}{M_\odot}\right)^{-1}\left(\frac{\rho_0}{10^3\,\text{kg/m}^3}\right)^{-1}. \tag{10.5}$$

Now, let us turn our attention to (10.2) and suppose that $\alpha_0 > 1$. Using $a_n = v\,dv/d\ell$, we can integrate this equation between the limits $\ell = \ell_0$ to $\ell = \infty$ and the corresponding velocities of $v = 0$ and $v = v_\infty$. By doing so, the terminal velocity (v_∞) is given by

$$v_\infty \approx \sqrt{\frac{2GM_\star}{\ell_0}\,(\alpha_0 - 1)}. \tag{10.6}$$

To simplify matters somewhat, let us suppose that the microscopic object is launched from a planet or moon situated in the habitable zone (HZ) of its host star that receives the same stellar flux as Earth. In this case, by utilizing (4.5), we obtain

$$v_\infty \approx 42\,\mathrm{km/s}\sqrt{\alpha_0 - 1}\left(\frac{M_\star}{M_\odot}\right)^{-1/4}. \tag{10.7}$$

Although this speed is not high, it suffices to traverse the diameter of the Milky Way in a timespan of < 1 Gyr. Hence, it may seem at first glimpse that Arrhenius's hypothesis is plausible. However, the major missing ingredient in his model was the sterilizing effect of high-energy electromagnetic radiation (ultraviolet, X-rays, and gamma rays) and cosmic rays. Experiments conducted by Paul Becquerel seemingly indicated that any microbes or their spores became extinct upon being irradiated by ultraviolet (UV) radiation. With the advent of additional studies in radiation biology that yielded similar results, Arrhenius's hypothesis of radiopanspermia was consigned to obscurity during the first half of the twentieth century.

The next major phase of panspermia was initiated by Sir Fred Hoyle and Chandra Wickramasinghe in the 1970s. While some aspects of the original Hoyle-Wickramasinghe "cometary panspermia" hypothesis were founded on robust science—the presence of organic molecules in interstellar dust being a noteworthy example—others concerning the transfer and ubiquity of life remain unproven. The most controversial facet of their writings was the conjecture that pathogens responsible for certain human diseases (e.g., influenza) were transported from space. As one might anticipate, this particular assertion was almost universally greeted with marked skepticism or outright dismissal by the scientific community. Owing to the substantial volume of publications by Hoyle and Wickramasinghe, a succinct summary is rendered impractical: the reader is instead referred to Hoyle and Wickramasinghe (2000), Wickramasinghe (2010), and Wickramasinghe et al. (2019) for in-depth expositions of the cosmic panspermia paradigm.

At around the same time as the Hoyle-Wickramasinghe model, a radical variant of panspermia was mooted by the Nobel laureate Francis Crick and Leslie Orgel, one of the doyens of abiogenesis. Crick and Orgel (1973) suggested that panspermia could be *directed,* i.e., the transfer of life from one world to another could occur by means of interstellar probes launched by ETIs.[3] In comparison to natural panspermia models, directed panspermia is not subject to biological constraints (e.g., radiation shielding) to the same

3. Naturally, the same logic applies to human beings (Mautner 1997). In principle, we may opt to seed other planetary systems with life. One possibility explored in some detail by Tepfer and Leach (2006) entails the use of plant seeds as vectors.

degree but, instead, is contingent on the existence of ETIs. Crick and Orgel also proposed that signatures of directed panspermia might be detected in the genetic code if life on Earth had been seeded through such a process. In a recent analysis, Makukov and Shcherbak (2018) claim to have identified artificial patterns in the genetic code, but it can be safely repudiated in the absence of any other corroborating evidence to date.

Panspermia has witnessed a modest revival since the 1990s, with a number of studies attesting to the fact that *interplanetary* panspermia is feasible in theory. Developments in celestial dynamics have shown that a small fraction of rocks expelled from one planet in our Solar system (e.g., Mars) can reach another planet (e.g., Earth) in a relatively short period of time, even if the mean transit times are long ($\sim 10^6$ yr). During the same period, space microbiology has grown by leaps and bounds and we possess a much deeper understanding of polyextremophiles—that is, microbes capable of concurrently tolerating multiple extremes such as radiation and temperature, although much still remains unknown on this front (Harrison et al. 2013). It is not merely microorganisms that deserve classification as polyextremophiles but also certain complex multicellular organisms such as the members of the phylum *Tardigrada*. By taking our cue from W. L. Nicholson (2009), we may confer the epithet of "micronauts" on the intrepid microbial astronauts potentially capable of journeying from one world to another.

To leading order, while it could be argued that interplanetary panspermia has steadily gained recognition as a genuine possibility (at least within the context of our Solar system), the status quo with regard to interstellar panspermia remains more complex and divisive. The chief difficulty with transporting life from one system to another via lithopanspermia was lucidly identified by Arrhenius—to wit, the low impact probabilities associated with life-bearing rocks from extrasolar systems. In toto, although the study of panspermia has been slowly reentering the mainstream, surprisingly few reviews of this subject exist: the reader is directed to Burchell (2004) and Wesson (2010) for qualitative conspectuses of panspermia.

Lastly, it is worth asking whether any evidence for panspermia has been detected so far: in turn, this requires us to identify the routes by which the hypothesis has been evaluated. The best-case scenario is to discover a sample of extinct life in rocky material transported from another world. Martian meteorites are well suited in this respect as they can be distinguished by means of their bulk elemental composition, isotopic ratios, and the analysis of trapped atmospheric gases within them. In 1996, David McKay and colleagues (1996) analyzed the Martian meteorite ALH84001

and proposed that carbonate globules discerned in this rock displayed evidence of fine-grained structures potentially indicative of biogenic activity. Naturally, the reported discovery sparked a media sensation at that time, with its influence extending all the way to the uppermost echelons of the US government, as embodied in the speech delivered by President Clinton.[4] Regardless of whether ALH84001 actually comprised traces of life, it rendered astrobiology an invaluable service by triggering a renewed interest in this field.

The evidence assembled over the past two decades indicates that most of the interesting features of this meteorite are explainable by an array of nonbiological processes, with arguably the most notable being the transformation of minerals into biomimetic structures via crystallization and binding to organic substrates (Martel et al. 2012). Nevertheless, it is particularly noteworthy vis-à-vis the preceding discussion that a few intriguing hypotheses favor the origin of life on Mars and its subsequent transfer to Earth. The advantages attributed to early Mars include (1) access to large-scale terrestrial geochemical environments (which might not have existed on Hadean Earth), and (2) relative abundance of oxidized boron and molybdenum minerals; the reader may peruse Benner and Kim (2015) for a pithy overview.

Another line of reasoning that was invoked as purported evidence for panspermia pertains to the increase of biological complexity on our planet. By treating the *functional* genome length as a reliable measure of biological complexity and modeling the former quantity in terms of exponential growth, Sharov (2006) determined that the inception of Earth-based life should have occurred at ~ 10 Ga, consequently implying that it should have originated elsewhere because our planet's age is only ~ 4.6 Ga. However, both the postulates underpinning this approach and the statistical techniques employed in the publication were critiqued, owing to which the credibility of this result is questionable.

10.1.2 Motivation for studying panspermia

Henceforth, we will employ the subscripts "I" and "II" to reference the systems from which the transfer of life is initiated and concluded, respectively.

---

4. NASA Jet Propulsion Laboratory (1996, August 7), President Clinton statement regarding Mars meteorite discovery, https://www2.jpl.nasa.gov/snc/clinton.html

Let us denote the probabilities that life originates spontaneously (i.e., through abiogenesis) on these two worlds by $\mathcal{P}_{\mathrm{I}}$ and $\mathcal{P}_{\mathrm{II}}$. Next, the probability of life being *successfully* transferred from I to II via panspermia is represented by $\mathcal{P}_{\mathrm{T}}$. If panspermia is the dominant factor responsible for the genesis of new life-bearing worlds, a useful heuristic is $\mathcal{P}_{\mathrm{I}} \cdot \mathcal{P}_{\mathrm{T}} \gtrsim \mathcal{P}_{\mathrm{II}}$. In other words, this criterion indicates that the joint probability of abiogenesis on the donor world and its transfer to the appropriate recipient world must exceed the likelihood of life originating de novo on the latter. We can rewrite this inequality to obtain

$$\mathcal{P}_{\mathrm{T}} \gtrsim \frac{\mathcal{P}_{\mathrm{II}}}{\mathcal{P}_{\mathrm{I}}}. \tag{10.8}$$

Although this equation is very simple, several interesting inferences may be drawn from it.

First, let us consider the scenario in which the probability of abiogenesis is nearly constant on all worlds; this choice of a flat prior is commonplace in the scientific literature. As this assumption corresponds to $\mathcal{P}_{\mathrm{I}} \sim \mathcal{P}_{\mathrm{II}}$, we see that (10.8) imposes the constraint $\mathcal{P}_{\mathrm{T}} \sim 1$. In other words, if all worlds have a roughly equal probability of abiogenesis, panspermia is likely to be a major player only if this mechanism has a high likelihood of occurrence. However, as we will see shortly, there are manifold steps involved and it seems improbable that $\mathcal{P}_{\mathrm{T}} \sim 1$, except perhaps for directed panspermia. Hence, either panspermia is unimportant in the emergence of new life-bearing worlds, or the flat prior ansatz is incorrect.

While the first explanation is undoubtedly a possibility, let us turn our attention to the alternative. We will therefore suppose that $\mathcal{P}_{\mathrm{I}}/\mathcal{P}_{\mathrm{II}} \gg 1$ holds true. In qualitative terms, this limit amounts to contending that the physicochemical conditions for the origin of life are much more favorable on the donor world relative to the recipient world. In prior chapters, we have delineated how a multitude of stellar and planetary factors mediate habitability. Hence, it is not inconceivable that some worlds are much more conducive to the emergence of life than others. Under the assumption of this nonuniform prior, it is thus apparent that $\mathcal{P}_{\mathrm{T}} \ll 1$ is mathematically allowed. As long as the destination worlds have sufficiently low probabilities of abiogenesis, panspermia might be regarded as a genuine mechanism for spreading life to them.

10.1.3 What factors govern the likelihood of panspermia?

Earlier, we commented on the requirements for radiopanspermia and suggested that the small size of the spores, microbes, or microbe-carrying grains pose difficulties to long-term transit. Hence, in our subsequent discussion, when we invoke the term *panspermia,* it will implicitly refer to lithopanspermia, unless explicitly stated otherwise. In order to understand how panspermia functions, it is helpful to break it down into various stages. By assessing the likelihood of each stage, one may thus gauge the likelihood of the panspermia process as a whole. In order to accomplish this task, we will adopt the framework presented in B. C. Clark (2001) and W. L. Nicholson (2009), which closely resembles the Drake equation described in Section 8.1.

The probability of life being successfully transferred from I to II is collectively given by

$$\mathcal{P}_{\rm T} = \mathcal{P}_{\rm BZ} \cdot \mathcal{P}_{\rm EM} \cdot \mathcal{P}_{\rm SL} \cdot \mathcal{P}_{\rm ST} \cdot \mathcal{P}_{\rm SE} \cdot \mathcal{P}_{\rm SI} \cdot \mathcal{P}_{\rm R} \cdot \mathcal{P}_{\rm SP}. \tag{10.9}$$

There are eight factors in total, which are described below, thereby making the overall equation rather difficult to handle. Even if we suppose that all of them have a relatively high probability of ~ 0.1, it is apparent that the overall probability becomes very small ($\sim 10^{-8}$). The only way to avoid this fate for $\mathcal{P}_{\rm T}$ is by demanding that many of the above probabilities are close to unity. In view of the multiple steps involved, it is hardly surprising some studies indicate that Mars-to-Earth panspermia is highly probable (Mileikowsky et al. 2000a), whereas others arrived at the diametrically opposite conclusion (B. C. Clark 2001). Assessing the magnitudes of the variables in (10.9) from an empirical standpoint necessitates in-depth laboratory experiments in a multitude of space- and ground-based environments. The reader is referred to Horneck et al. (2010), Moissl-Eichinger et al. (2016), Cottin et al. (2017), and Merino et al. (2019) for detailed surveys of this extensive field.

The first condition that needs to be fulfilled is that the impactor should not sterilize the region of the donor planet when it strikes the surface; the associated probability is $\mathcal{P}_{\rm BZ}$. At first glimpse, it is tempting to conclude that $\mathcal{P}_{\rm BZ} \ll 1$ because the impact of an object traveling at velocities of > 10 km/s is sufficient to melt or vaporize rocks, consequently rendering the impact zone sterile. However, factual data as well as theoretical models suggest that this is not the case. The analysis of several Martian meteorites has

revealed that they were subjected to maximum temperatures of < 373 K. For example, at temperatures > 313 K, the intensity of certain magnetization features would be suppressed. By confirming the *absence* of such signatures, Weiss et al. (2000) deduced that the interior of the famous Martian meteorite ALH84001 (introduced previously in this chapter) was not subjected to temperatures above this threshold.

A number of theoretical explanations have been proposed to explain why some rocks are expelled at low temperatures. If the object strikes the donor planet at an oblique angle, much of its energy could be converted into kinetic energy instead of thermal energy. One of the most widely accepted spallation mechanisms for those ejecta that have experienced relatively mild shocks was expounded by Melosh (1984). Intuitively, this phenomenon can be envisioned as follows. When the impactor strikes the surface, it triggers shock waves below the ground that are reflected back from the surface. The reflected shock wave is phase-shifted by 180° with respect to the incoming shock wave. Hence, when the two shock waves are superimposed, they nearly cancel each other out and reduce the net intensity of the shock; the surface layer at which the destructive interference occurs is known as the spallation zone.

The mechanics of spallation have been worked out in detail by several authors. We will restrict ourselves to providing only the final expressions, as the derivations are based on a combination of empirical, numerical, and analytical studies. The total mass of expelled ejecta ($M_{\rm ej}$) is expressible as (J. C. Armstrong et al. 2002)

$$M_{\rm ej} \approx \frac{0.75 P_{\rm max}}{\rho_t c_L v_{\rm imp}} \left[\left(\frac{v_{\rm imp}}{2\Delta v} \right)^{5/3} - 1 \right] M_{\rm imp}, \tag{10.10}$$

where $v_{\rm imp}$ and $M_{\rm imp}$ are the velocity and mass of the impactor, respectively, whereas Δv signifies the velocity of the ejecta. In the above formula, $P_{\rm max}$ represents the maximum shock pressure experienced by the ejecta, ρ_t denotes the density of the target struck by the impactor, and c_L is the longitudinal sound speed in the target; for basalt, we have $\rho_t \approx 2.9 \times 10^3$ kg/m^3 and $c_L \approx 6$ km/s. In order to ensure that the temperature of the ejecta does not exceed 373 K, $P_{\rm max} \lesssim 1$ GPa ($1\,{\rm GPa} \equiv 10^9$ Pa) is customarily imposed as a requirement. The average size of the ejected fragments $R_{\rm ej}$ is estimated by using

$$R_{\rm ej} \approx \frac{P_{\rm max} R_{\rm imp}}{4\rho_t \Delta v^{2/3} v_{\rm imp}^{4/3}}, \tag{10.11}$$

where $R_{\rm imp}$ is the radius of the impactor. The maximum size of the ejected fragments is $R_{\rm max} \approx 6.3 R_{\rm ej}$ for basalt-like rocks. The average number of fragments generated can be readily computed by dividing (10.10) by the mass of each fragment that is obtained from (10.11).

By inspecting (10.10), we see that Δv must obey $\Delta v < v_{\rm imp}/2$; that is, it cannot exceed 50 percent of the impactor's velocity. However, actual experiments have revealed that the upper bound may reach ~ 85 percent in some cases. If we invert this inequality, we find $v_{\rm imp} > 2\Delta v$, implying that the impactor must have a velocity that is ideally twice the escape velocity of the donor world to ensure that the ejecta escapes the surface. The escape velocity ($v_{\rm esc}$) for a world with radius R and mass M is

$$v_{\rm esc} = \sqrt{\frac{2GM}{R}} \approx 11.2\,{\rm km/s} \left(\frac{R}{R_\oplus}\right)^{1.35}. \tag{10.12}$$

The second equality follows from the mass-radius relationship $M \propto R^{3.7}$. As seen from (10.12), for an Earth-sized planet, we require $v_{\rm imp} > 22.4$ km/s. This velocity is higher than the average velocities of impacts, as they tend to be < 15 km/s in most instances. However, if we consider donor worlds smaller than the Earth, it becomes apparent from (10.12) that the escape velocity declines quite appreciably with radius; for example, the escape velocity for Mars is only 5 km/s. The lower escape velocities for icy objects (e.g., Europa and Enceladus) in our Solar system offer a valid reason to hypothesize that one (or more) of them may have served as the donor world, especially since icy ejecta with microbial loads appear to represent viable transports as per laboratory experiments (Burchell et al. 2003).

Apart from the preceding constraint on Δv, there is also an additional restriction on the size of the vertically ejected fragments. If the fragments are too small, they will not escape the donor world due to the drag exerted by the atmosphere. Artemieva and Ivanov (2004) derived an analytical estimate for the minimum fragment size ($R_{\rm min}$), which was found to exhibit reasonable agreement with numerical simulations,

$$R_{\rm min} \approx \frac{3P_s}{8\rho_t g} \left(\frac{\Delta v + v_{\rm esc}}{\Delta v - v_{\rm esc}}\right), \tag{10.13}$$

where P_s is the surface atmospheric pressure and g is the planet's surface gravitational acceleration. As the formula illustrates, $R_{\min}$ becomes very large when we consider objects with $\Delta v \approx v_{\rm esc}$.

The second step is actually an essential precondition for the first step as it signifies the existence of endoliths—namely, organisms that live within rocky materials—capable of withstanding harsh enviroments; in probabilistic terms, it is represented by $\mathcal{P}_{\rm EM}$ in (10.9). A number of endoliths have been documented, ranging from bacteria to algae, fungi, and lichens, and they encompass both photoautotrophs and heterotrophs. The Atacama Desert merits a special mention because it is unique in two respects: it not only is the driest region (in terms of rainfall) on Earth but also holds the world record for the maximum UV irradiance. Yet, it is known to possess low-complexity communities of endolithic microbes that are polyextremophiles; a review of this subject can be found in Wierzchos et al. (2018). In this context, we note that a number of species belonging to the bacterial genus *Bacillus* are endoliths and polyextremophiles, inhabiting niches as diverse as Antarctic soils and deep-sea sediments. This is an encouraging sign because *Bacillus* species have been employed in several laboratory experiments designed to test the efficacy of different stages in the panspermia process.

The next variable that we ought to evaluate in (10.9) is $\mathcal{P}_{\rm SL}$, which encapsulates the probability of surviving the launch. Multiple factors must be taken into account at this stage. First, we remark that the ejected rocks are accelerated to their final speeds of Δv over a span of $\sim 10^{-3}$ s. In other words, they will experience an acceleration of $\sim 10^6\, g_\oplus$ (where $g_\oplus \approx 9.8$ m/s^2). However, it is not merely the acceleration in isolation that plays a critical role but also its rate of change—known in scientific parlance as the *jerk*—whose mathematical definition is given by da/dt.

Experiments involving spores of *Bacillus subtilis* and *Deinococcus radiodurans* have yielded survival rates of 40 to 100 percent when they were subjected to accelerations and jerks that were 2.5–25 times higher than the corresponding values experienced during ejection from Mars (Mastrapa et al. 2001). Apart from spores, a wide number of microbes on Earth, such as *Escherichia coli, Paracoccus denitrificans,* and *Shewanella amazonensis,* not only are capable of tolerating acceleration of $\sim 4 \times 10^5$ g but also reproduce and proliferate under these conditions (Deguchi et al. 2011). The nematode *Caenorhabditis elegans,* which is an *animal* of size $\sim 10^{-3}$ m (i.e., clearly not a microbe), is also adept at tolerating high acceleration up to $\sim 4 \times 10^5$ g (de Souza & Pereira 2018).

Second, the ejecta produced by the spallation mechanism are subjected to very high shock pressures ranging between ~ 1 and 100 GPa. Laboratory experiments indicate that *B. subtilis* spores and the components of the lichen *Xanthoria elegans* can survive shock pressures of ~ 5–50 GPa, albeit the survival fraction drops to $\sim 10^{-4}$ for pressures of 40–50 GPa (Stöffler et al. 2007). In a similar vein, at shock pressures of ~ 25 GPa, the yeast *Saccharomyces cerevisiae* exhibits a survival fraction of $\sim 10^{-4}$ (M. C. Price et al. 2013). The corresponding shock pressure range for the survival of the cyanobacterium *Chroococcidiopsis* is comparatively limited, with an upper bound of ~ 10 GPa. Other investigations of *B. subtilis* spores have obtained a survival fraction of $\sim 10^{-5}$ at even higher shock pressures of ~ 60–80 GPa. Likewise, when cells of the bacterium *Rhodococcus erythropolis* were subjected to shock pressures of ~ 78 GPa, the survival fraction was determined to be $\sim 10^{-7}$ (Burchell et al. 2004).

Moving on, the next stage in the panspermia process is surviving the transit from the donor to the recipient planet, whose associated probability is $\mathcal{P}_{\mathrm{ST}}$. This field has been subjected to intense experimental scrutiny using a multitude of Earth's extremophiles ranging from bacteria and archaea to lichens and tardigrades (Veras et al. 2018). Many studies have yielded survival fractions on the order of 10 percent in the presence of suitable shielding, although it is essential to recognize that these experiments were conducted over a period of ~ 0.1–1 yr. Furthermore, most of these investigations utilized only one (or a few) species, with an intriguing exception being the series of experiments undertaken at the EXPOSE facility on board the International Space Station in low Earth orbit (LEO), which subsequently established the survival of endolithic microbial communities in space over the span of 1.5 yr (Onofri et al. 2012; de Vera et al. 2019). It is vitally important for future experiments to magnify their focus on microbial communities as opposed to isolated species because of the manifold benefits (e.g., synthesis of complementary metabolic products and growth factors) conferred by the former.

There are numerous challenges encountered by microbes traveling in space. A few of the more notable ones are very cold temperatures, desiccation in the ambient near-vacuum of space, ionizing radiation (comprising cosmic rays and high-energy electromagnetic radiation), and ultraviolet photons. Owing to the complexity arising from multiple environmental extremes, we will only analyze a select few variables that are conventionally deemed the most pertinent ones. In all cases, the survival fraction (χ) is generically expressed as

$$\chi = \exp\left(-\frac{t_0}{\tau_{\rm crit}}\right), \tag{10.14}$$

where t_0 is the total time spent in transit and $\tau_{\rm crit}$ is the decay time over which the survival fraction drops by a factor of e. This timescale can be further simplified as $\tau_{\rm crit} \approx \min\{\tau_k\}$, where τ_k is the decay timescale for the k-th process.

The first pivotal issue in need of addressing is desiccation. It is well-known that extreme desiccation causes DNA damage and disruption of cellular metabolism, functioning, and integrity due to the inevitable degradation of biomolecules. *B. subtilis* spores were exposed to vacuum, albeit with UV shielding, over 5.77 yr at the Long Duration Exposure Facility situated in LEO. It was found that the survival fraction after this period was relatively low ($\lesssim 1$ percent), but the addition of salts or sugars dramatically increased the survival fraction to ~ 70 to 90 percent (Horneck et al. 2010). The formation of biofilms (viz., microbial aggregates) could also render additional protection against dehydration effects. On a related note, the effects of desiccation-induced damage were found to be severe for the lichen *X. elegans* (A. Brandt et al. 2016). It is, however, worth mentioning that most experiments were carried out at higher temperatures than deep space. As a consequence, the exponential dependence on the inverse temperature in reaction kinetics may lead to longer survival times.

With these caveats in mind, if we solve for the decay timescale $\tau_{\rm vac}$ using (10.14) and a survival fraction of ~ 80 percent over 5.77 yr, we obtain $\tau_{\rm vac} \approx 25.8$ yr. Although this timescale is very short, it is crucial to recognize that these studies were conducted for microbes subjected to vacuum. In actuality, it is plausible that the rocks flung out into space would retain some trapped air (i.e., host finite pressure) after ablation and solidification of the outer layers. In this event, the moisture present in the interior will not undergo sublimation, and the microbes might avoid desiccation. Hence, DNA degradation via exposure to vacuum constitutes, in all likelihood, the limiting factor for those rocks with air leakage to space, but its significance for sealed ejecta is not precisely understood.

The second major factor that needs to be accounted for is DNA damage due to hydrolysis. Of course, this mechanism implicitly presupposes the existence of ice or even liquid water, thus implying that it is only applicable to sealed ejecta whose interiors are protected from the vacuum. The reason DNA hydrolysis is relatively unimportant for Earth-based organisms is

because efficient DNA repair processes exist. At the low temperatures anticipated in outer space, metabolism and repair machinery are both significantly impeded. The survival fraction of bacteria such as *E. coli* is very sensitive to the fraction of DNA bases degraded during hydrolysis. For example, base loss of 0.1 percent yields a survival fraction of ~ 10 percent, whereas base loss of 0.3 percent translates to the survival of ~ 0.001 to 0.1 percent of bacteria. The characteristic time required for a given fraction of DNA base damage is very sensitive to the ambient temperature experienced by the ejecta. Based on the experimental data tabulated in Mileikowsky et al. (2000a), the characteristic hydrolysis time (τ_{H_2O}) can be modeled as

$$\tau_{H_2O} \sim 4.2 \times 10^4 \text{ yr} \exp\left[1.6 \times 10^4 \left(\frac{1}{T_a} - \frac{1}{253}\right)\right], \tag{10.15}$$

where T_a is the ambient temperature inside the target. Note that this timescale will probably become increasingly inaccurate as we move toward lower temperatures because of additional deleterious physiological and biochemical effects engendered by low temperatures.

The third factor that should be accounted for is the cumulative damage caused by ionizing radiation during transit (Dartnell 2011). The effects of ultraviolet radiation and X-rays are comparatively unimportant because microbes would be completely shielded by a rock thickness of ~ 1 cm. Galactic Cosmic Rays (GCRs) are more problematic in this regard, as they possess high energies and can deposit their energy at greater depths. The relationship between the characteristic decay time for GCRs (τ_{GCR}) and the shielding thickness (Δr) is not linear. For example, if we consider the radioresistant bacterium *D. radiodurans,* the following scalings are derived from Mileikowsky et al. (2000a):

$$\begin{aligned}
\tau_{GCR} &\sim 1.2 \times 10^5 \text{ yr} && 0 < \Delta r \lesssim 0.4 \text{ m}, \\
\tau_{GCR} &\sim 3 \times 10^5 \text{ yr} \left(\frac{\Delta r}{1 \text{ m}}\right) && 0.4 \text{ m} \lesssim \Delta r \lesssim 1 \text{ m}, \\
\tau_{GCR} &\sim 3 \times 10^5 \text{ yr} \left(\frac{\Delta r}{1 \text{ m}}\right)^4 && \Delta r \gtrsim 1 \text{ m}.
\end{aligned} \tag{10.16}$$

As per the above formula, shielding starts to play a prominent role when $\Delta r \gtrsim 1$ m, implying that microbes in meter-sized rocks and larger should

be most immune to the effects of GCRs. B. C. Clark (2001) also studied the damage from GCRs and concluded that rocks of sizes ~ 1 m and larger could offer adequate shielding over a timescale of ~ 1–10 Myr, which is in reasonable agreement with (10.16).

Apart from GCRs, another major source of radiation is the ejecta itself. If the ejecta is made up of rocks, as generally assumed, it will possess radioactive minerals comprising the likes of uranium-238, uranium-235 and thorium-232. The difficulty with estimating the contribution from natural radioactivity is that it depends on both the age of ejected rocks as well as their composition. Granites on our planet commonly exhibit a high abundance of radioactive isotopes, whereas ultramafic rocks from the Earth's mantle exhibit lower abundances. If we consider rocks with composition akin to granites from Earth, the total decay timescale including contributions from GCRs and natural radioactivity ($\tau_{\rm rad}$) is roughly expressible as (Mileikowsky et al. 2000a)

$$\begin{aligned} \tau_{\rm rad} &\sim 8.4 \times 10^4 \,{\rm yr} \qquad\qquad 0 < \Delta r \lesssim 0.7 \,{\rm m}, \\ \tau_{\rm rad} &\sim 1.2 \times 10^5 \,{\rm yr} \left(\frac{\Delta r}{1 \,{\rm m}} \right) \quad \Delta r \gtrsim 0.7 \,{\rm m}. \end{aligned} \tag{10.17}$$

In contrast, we note that a more recent study by Valtonen et al. (2009) favors a longer timescale for $\tau_{\rm rad}$, whose upper bound is given by

$$\tau_{\rm rad} \sim 7.5 \times 10^7 \,{\rm yr} \left(\frac{\Delta r}{1 \,{\rm m}} \right)^2 . \tag{10.18}$$

Therefore, from inspecting (10.15), (10.16), (10.17), and (10.18), we determine that the overall decay timescale $\tau_{\rm crit}$ might be on the order of 10^5 yr. However, in the event that neither DNA hydrolysis nor vacuum-induced desiccation play a major role, it is conceivable that meter-sized boulders may support lifeforms for $\sim 10^7$–10^8 yr.

The next three steps in the panspermia process entail microbes surviving atmospheric passage and landing on the recipient world. In (10.9), $\mathcal{P}_{\rm SE}$ denotes the probability of surviving atmospheric entry, $\mathcal{P}_{\rm SI}$ quantifies the likelihood of surviving impact with the surface, and $\mathcal{P}_{\rm R}$ signifies the probability of safely releasing microbes from the target. Relatively few empirical results are available concerning the likelihood of microorganisms surviving

this passage. One of the points worth singling out is that frictional heating melts the surface of the meteorite, and the resultant vaporized material transports away the heat. It is therefore suspected that at least some meteorites (e.g., ALH84001) ought not experience significant heating in their interiors (Weiss et al. 2000), where microbes could reside.

It is instructive to elaborate on and quantify the above statement concerning the ablation (i.e., erosion) of the meteorite. We will carry out a heuristic derivation that captures the salient details of the analytical model presented in Whipple (1938). The drag force experienced by an object is roughly expressible as the product of the ram pressure and the area of the object. The ram pressure is given by $\rho_a v^2$ and the area is expressible as $\sim (m/\rho_t)^{2/3}$; note that ρ_a is the density of air, while m and v are the mass and velocity of the meteorite. Hence, we see that the drag force is proportional to $m^{2/3}v^2$. The power exerted by this force is approximately equal to the product of the force and velocity. Now, let us suppose that this power is utilized for the purpose of vaporization. As the heat required for vaporization is the product of the mass m and the latent heat of vaporization, it follows that the heating rate is proportional to dm/dt. Therefore, by combining all of these results, we obtain

$$\frac{dm}{dt} = -\mathcal{C}_0 \rho_a m^{2/3} v^3, \tag{10.19}$$

where $\mathcal{C}_0$ is a constant whose exact value does not matter for our analysis. Let us consider the idealized limit wherein v does not change by more than a factor of order unity. This premise is valid when the velocity of the meteorite far away from the recipient world (v_∞) is comparable to the planet's escape velocity. Next, we will use $dz = v dt$ and $\rho_a \propto \exp(-z/H_a)$, where H_a refers to the atmospheric scale height of an isothermal atmosphere. By plugging these relations into (10.19), we arrive at

$$m^{1/3} \sim m_\infty^{1/3} - \mathcal{B}_0 v_\infty^2, \tag{10.20}$$

where m_∞ is essentially the initial mass of the meteorite, and we have opted to replace v with v_∞ for reasons explained earlier. From this equation, two interesting results are deducible. If the initial mass of the object is sufficiently high, subsequently it reaches the surface without much ablation; when evaluated numerically, this amounts to requiring $m_\infty \gtrsim 10^5$ kg for velocities on

the order of 10 km/s. On the other hand, if m_∞ is small, we find that $m/m_\infty \to 0$. If we specialize to the case where the final mass is a small fraction of the initial mass (i.e., $m \propto m_\infty$), it is found that $m \propto v_\infty^{-6}$. Both of these results are qualitatively consistent with observational and laboratory studies as well as theoretical and numerical models (Ceplecha et al. 1998).

Of course, one of the chief limitations of the preceding approach is that the object is treated as a single entity. Au contraire, it is plausible that the meteorite will undergo fragmentation as a result of aerodynamic drag forces, and the resultant debris will be scattered over a large area. If the size of the debris (r_d) is small enough, they will reach the surface at the terminal velocity, which can be relatively low. The terminal velocity ($v_{\rm term}$) is found by balancing the upward viscous drag force with the downward gravitational force (with buoyancy force neglected), thereby yielding

$$v_{\rm term} \approx 0.65\,{\rm km/s} \left(\frac{\rho_t/\rho_a}{3000}\right) \left(\frac{g}{g_\oplus}\right) \left(\frac{r_d}{1\,{\rm mm}}\right)^2 \left(\frac{\nu_a}{10^{-5}\,{\rm m^2/s}}\right)^{-1}, \quad (10.21)$$

where ν_a refers to the kinematic viscosity of the atmosphere. As this expression illustrates, to achieve low terminal velocities, the fragmented debris must possess very small sizes.

Owing to the relative paucity and limited scope of empirical data, we do not have sufficient information to determine whether the survival during the entry phase onto the recipient world is feasible. In one of the notable early experiments of its ilk, Fajardo-Cavazos et al. (2005) investigated the response of *B. subtilis* spores using sounding rockets and found that 1 to 4 percent of them survived at an atmospheric entry velocity of 1.2 km/s. However, the tide appeared to turn subsequently with the results from the STONE-5 and STONE-6 experiments, which investigated the survival of several microbes—such as the cyanobacterium *Chroococcidiopsis,* the spores of *B. subtilis,* the fungus *Ulocladium atrum,* and the lichen *Rhizocarpon geographicum*—at an atmospheric entry velocity of 7.6 km/s. Despite the presence of 1–2 cm shielding, it was concluded that none of these species seemed to have survived atmospheric entry (de la Torre et al. 2010).

This result is consistent with numerical and analytical modeling indicating that the thickness of the shield must be > 5 cm in order to ensure that microbes are sufficiently protected (Foucher et al. 2010). Although the viability of organisms during atmospheric passage remains under question, more recent experiments are beginning to yield promising results once again (Barney et al. 2016). For example, the thermophilic and spore-forming

anaerobic bacterium *Thermoanaerobacter siderophilus* was embedded in an artificial meteorite with a shielding of 1.4 cm and imparted an atmospheric entry velocity of 7.6 km/s. The experiment yielded a fairly high survival rate of 8 to 25 percent for *T. siderophilus* (Slobodkin et al. 2015). Note, however, that this study chose to deploy *T. siderophilus* on account of its thermophilic nature, but its capacity to withstand the other stages of the panspermia process remains indeterminate.

The final step in the successful transfer of life concerns the capability of the transported microbes to survive and grow on the recipient world, whose probability is embodied by $\mathcal{P}_{\mathrm{SP}}$ in (10.9). This phase brings to mind the exquisite words of the Persian poet Jalāl ad-Dīn Muhammad Rūmī (2004, pg. 15):

> Be a spot of ground where nothing is growing,
> where something might be planted
> a seed, possibly, from the Absolute.

This factor is often omitted or neglected in the lion's share of publications, since the recipient world is automatically assumed to be habitable for the selfsame organisms. However, as we saw in Section 10.1.2, panspermia constitutes a potentially important mechanism when $\mathcal{P}_{\mathrm{II}} \ll \mathcal{P}_{\mathrm{I}}$—to wit, when the probability of abiogenesis on the recipient world is much lower than the donor world. Hence, this very same factor could act to suppress $\mathcal{P}_{\mathrm{SP}}$, although it is crucial to appreciate the differences between *habitable* and *inhabited*. In other words, the recipient world may be sufficiently habitable but not have witnessed an independent origin of life for one reason or another (Cockell 2014).

To properly assess $\mathcal{P}_{\mathrm{SP}}$, the physicochemical and geological constraints of the recipient world will need to be taken into consideration. Suppose, for the sake of argument, that microbes from Earth had been successfully transported to modern Mars. On the Martian surface, they would face a hostile cocktail of low atmospheric pressure and surface temperature, scarcity of liquid water, high doses of ionizing radiation, and many other hindrances. Yet, there are still localized "oases" where Martian life might have survived unto this day, most notably sequestered away in sheltered caves, ices, and the subsurface (Onstott et al. 2019; Stamenković et al. 2019; Carrier et al. 2020). If we instead consider transfer between Earth and Mars around 4 Ga, this choice increases $\mathcal{P}_{\mathrm{SP}}$, as both worlds probably hosted habitable conditions and possibly life.

Another valuable point to bear in mind is that, after landing on recipient worlds with relatively benign but not entirely hospitable environments, organisms could either remain dormant or maintain low-level metabolic activity until they experience more favorable conditions for growth (Lennon & Jones 2011), after which they can rapidly proliferate. Over the past few decades, a number of intriguing studies claimed to have revived bacterial spores or isolated live strains of microbes that are $\sim$ 10–100 Myr old (Moger-Reischer & Lennon 2019). Striking results in this realm include the apparent revival of 25 to 40 Myr bacteria spores from Dominican amber (Cano & Borucki 1995), the extraction of putative $>$ 100 Myr live bacteria and archaea from salt crystals (Vreeland et al. 2000; Jaakkola et al. 2016), and the purported renascence of aerobic microbes in $\sim$ 4 to 100 Myr seafloor sediments of the South Pacific Gyre (Morono et al. 2020).

Naturally, in light of the samples' alleged great antiquity, the central results espoused in most such publications were contested, vehemently so on occasion. At the minimum, three major classes of explanations have been propounded to explain the published ages of the samples—namely, the microbes and spores. First, despite the numerous precautionary protocols adopted, the prospects of laboratory contamination cannot be conclusively ruled out. Second, even if the geological samples recovered are millions of years old, it does not ineluctably follow that the organisms are of similar ages, as they may have settled in these habitats at a later stage. Third, theoretical and empirical analyses indicate that DNA cannot readily survive beyond timescales of $\sim$ 0.1–1 Myr (Hebsgaard et al. 2005).

If the claims of microbial and spore viability across timescales of $\sim$ 10–100 Myr are definitively established in the future, they may amplify the prospects for panspermia due to the underlying implication that lifeforms possess the capacity to persist over geological timescales. Aside from the aforementioned microbes ostensibly revitalized after unusually long temporal intervals, similar claims have been advanced for more complex organisms. Sediments in the Arctic permafrost host a wide range of biota, of which some might retain viability over timescales of $\sim 10^4$ yr. A study by Shatilovich et al. (2018) reportedly demonstrated that nematodes are capable of long-term cryobiosis (suspended metabolism in cold temperatures)—as per the study, they were revived after being frozen for $\sim$ 30 to 40 kyr—but this claim remains unsubstantiated because it is unclear, among other things, whether the nematodes in question were actually living amid the permafrost.

With this, we shall wrap up our analysis of the multiple stages of panspermia and move on. Clearly, in order to rigorously assess the viability

of panspermia, an integrated synthesis of laboratory experiments, field studies of polyextremophiles, and numerical and analytical modeling is of paramount relevance.

10.2 INTERPLANETARY AND INTERSTELLAR PANSPERMIA

Hitherto, we have dealt with the history and multiple stages involved in (litho)panspermia. We will now explore a few select models that have investigated the prospects for interplanetary and interstellar panspermia.

10.2.1 Interplanetary panspermia

Two of the most crucial *dynamical* factors that govern the likelihood of panspermia are the fraction of ejecta that impact the recipient world (planet II) and the transit time from the donor world (planet I) to the recipient world. Clearly, if the first factor is low, it will reduce the prospects for panspermia. Likewise, if the transit time (t_0) is long, it will lower the chances for panspermia because the survival fraction is exponentially dependent on t_0, as seen from (10.14).

A number of numerical models have been developed to study the two aforementioned issues. The reader may consult Gladman et al. (1996, 2005), which serve as representative publications in this field. Early simulations indicated that the fraction of Martian ejecta delivered to Earth ($f_{M\to E}$) could be approximated as

$$f_{M\to E} \approx 6 \times 10^{-3} \left(\frac{t_0}{10^6\ \mathrm{yr}}\right), \tag{10.22}$$

but this scaling was found to apply only when $t_0 < 10$ Myr; at longer timescales, the ejecta population was depleted due to other collisional and dynamical effects. At this stage, it is essential to recognize that panspermia is *not* a symmetric process. To illustrate this point, the fraction of Earth ejecta delivered to Mars ($f_{E\to M}$) was determined to be

$$f_{E\to M} \approx 1.6 \times 10^{-4} \left(\frac{t_0}{10^6\ \mathrm{yr}}\right), \tag{10.23}$$

with the formula losing applicability at $t_0 \gtrsim 10$ Myr. The result that $f_{M\to E} \approx f_{E\to M}$ does not hold true arises from the fact that the effective cross-sectional area of the Earth is much higher than that of Mars.

As there have been several numerical studies of interplanetary panspermia, we will restrict ourselves to describing the results from the state-of-the-art simulations undertaken by Worth et al. (2013). In this work, the trajectories of ejecta were tracked over a span of ~ 10 Myr, and the transfer of rocks from one world to another was evaluated. Unlike previous studies, the probabilities of ejecta from Mars and Earth impacting the major moons of Jupiter and Saturn were also calculated. On the basis of these results, the total amount of mass delivered to the recipient worlds was also estimated using empirical data from cratering rates. The major findings are summarized in Figure 10.1 and Table 10.1. In the latter, the total number of ejecta impacting the recipient world over the span of 3.5 Gyr should be regarded as a lower bound of sorts, because only rocks with sizes > 3 m were treated as having the potential to transport viable organisms.

At this stage, it is useful to develop heuristic scalings that enable us to gain a better understanding of the impact probability $\mathcal{P}_{\mathrm{I}\to\mathrm{II}}$ and the average transit time $\langle t_0 \rangle$ introduced in Table 10.1 for different planetary systems. A rigorous analytical model to calculate these parameters was developed by Veras et al. (2018), but the derivation and final results are both rather complex, owing to which we shall not pursue this approach. Let us commence our analysis by studying $\mathcal{P}_{\mathrm{I}\to\mathrm{II}}$. If the ejecta are emitted in an isotropic fashion, we may estimate $\mathcal{P}_{\mathrm{I}\to\mathrm{II}}$ as

$$\mathcal{P}_{\mathrm{I}\to\mathrm{II}} \sim \frac{\pi b_{\mathrm{II}}^2}{4\pi\,(\Delta a)^2} \sim \left(\frac{b_{\mathrm{II}}}{2\Delta a}\right)^2, \tag{10.24}$$

where πb_{II}^2 serves as the effective cross-sectional area associated with the recipient world and Δa is the separation between the donor and recipient worlds; b_{II} represents the impact parameter. If πb_{II}^2 was purely the geometric cross section, we would have $\pi b_{\mathrm{II}}^2 = \pi R_{\mathrm{II}}^2$, where R_{II} is the radius of the recipient world. However, the situation is rendered more interesting because of gravitational focusing and other gravitational interactions. In qualitative terms, the net effect of gravitational focusing is to enhance the probability of two particles colliding as a result of their mutual gravitational attraction.

In quantitative terms, we can derive b_{II} by enforcing conservation of energy and angular momentum. Let us assume the ejecta (test particle) is approaching from infinity with a speed v_∞. We will posit that the distance of closest approach, at which the gravitational capture might be initiated, is the Hill radius of planet II (instead of adopting R_{II}). The Hill radius is, sensu lato, the gravitational sphere of influence exerted by the planet. Let us

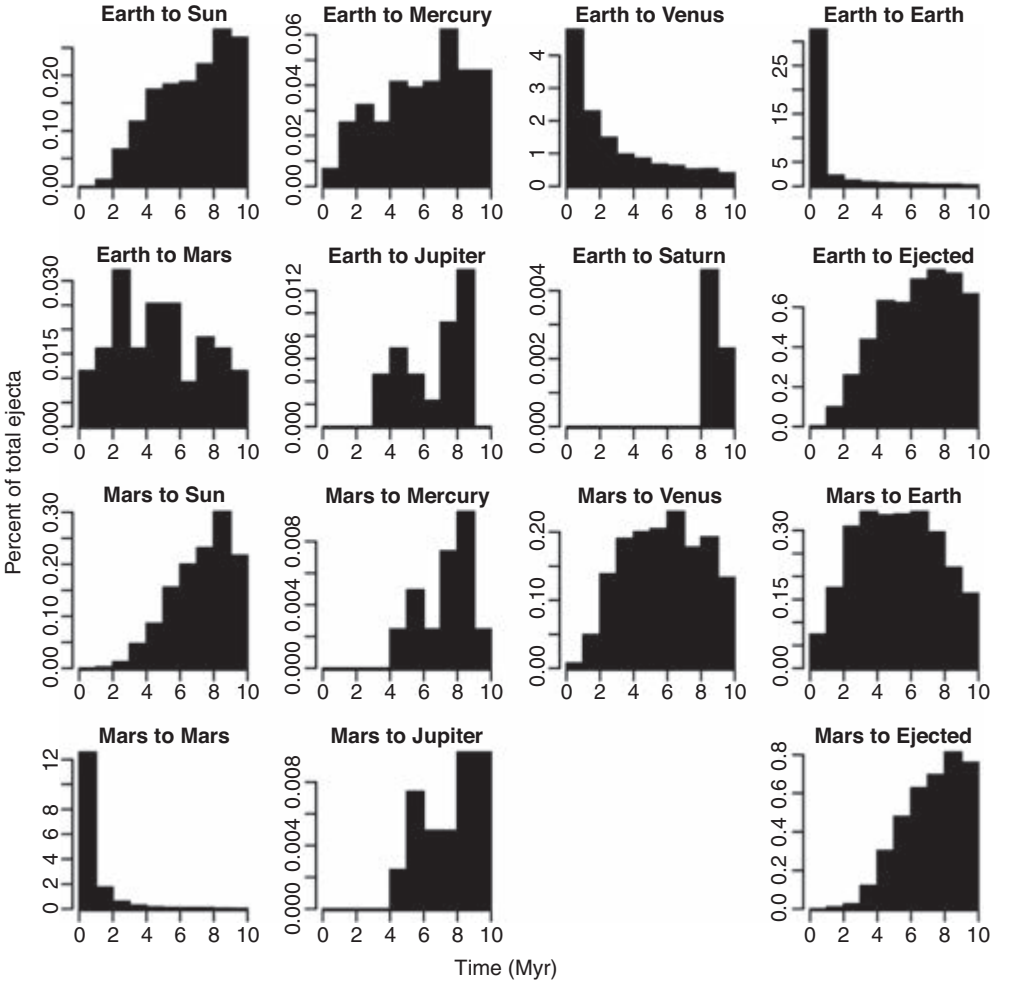


Figure 10.1 The fraction of ejecta that impact a chosen recipient world from a particular donor world as a function of time. (© Mary Ann Liebert, Inc. *Source:* R. J. Worth, Steinn Sigurdsson, and Christopher H. House [2013], Seeing life on the moons of the outer planets via lithopanspermia, *Astrobiology* 13[12]: 1155–1165, fig. 1.)

suppose that the recipient world has a mass $M_{\rm II}$ and orbits a star with mass $M_\star$ at a distance $a_{\rm II}$. A rough derivation of the Hill radius (R_H) follows from demanding that the mass density arising from the star is comparable to that of the planet at the Hill radius:

$$\frac{M_\star}{a_{\rm II}^3} \approx \frac{M_{\rm II}}{R_H^3} \quad \Rightarrow \quad R_H \approx a_{\rm II}\left(\frac{M_{\rm II}}{M_\star}\right)^{1/3} \tag{10.25}$$

Table 10.1 General characteristics of interplanetary panspermia within our Solar system

Donor	Recipient	$t_{\rm first}$ (in yr)	$\langle t_0 \rangle$ (in Myr)	$\mathcal{P}_{\rm I\to II}$ (%)	$N_{\rm tot}$
Earth	Venus	8.3×10^3	1.7	13	2.6×10^7
Earth	Earth	0.3	9.6×10^{-2}	40	7.9×10^7
Earth	Mars	1.1×10^5	4.7	0.18	3.6×10^5
Earth	Europa	Unknown	Unknown	2.8×10^{-6}	~ 6
Earth	Ganymede	Unknown	Unknown	3.1×10^{-6}	~ 6
Earth	Enceladus	Unknown	Unknown	5×10^{-7}	~ 0.1-1
Earth	Titan	Unknown	Unknown	7.5×10^{-7}	~ 0.1–1
Mars	Venus	4.6×10^5	5.8	1.5	1.2×10^7
Mars	Earth	2.8×10^5	5.2	2.6	2.1×10^7
Mars	Mars	0.6	0.21	16	1.3×10^8
Mars	Europa	Unknown	Unknown	2.7×10^{-6}	~ 20
Mars	Ganymede	Unknown	Unknown	3.0×10^{-6}	~ 20
Mars	Enceladus	Unknown	Unknown	$< 1.8 \times 10^{-7}$	< 1.4
Mars	Titan	Unknown	Unknown	2.7×10^{-7}	< 2.1

Notes: $t_{\rm first}$ represents the minimum amount of time taken by an object to reach the recipient world after ejection, while $\langle t_0 \rangle$ is the median transit time computed from the 10 Myr simulations with residual velocities (v_∞) of $\sim$0–1 km/s. The probability $\mathcal{P}_{\rm I\to II}$ quantifies the fraction of ejecta that reach the recipient world after the 10 Myr time span. Lastly, $N_{\rm tot}$ is the total number of ejecta with sizes > 3 m impacting the recipient world over the past 3.5 Gyr. (© Mary Ann Liebert, Inc. *Data source:* R. J. Worth, Steinn Sigurdsson, and Christopher H. House [2013], Seeding life on the moons of the outer planets via lithopanspermia, *Astrobiology* 13[12]: 1155–1165, table 2.)

A rigorous calculation of the Hill radius yields a factor of 3 in front of $M_\star$ inside the cube root but will otherwise leave the expression unchanged.

In our analysis, we treat the recipient world as a massive object at rest. By invoking the conservation of energy, we have

$$\frac{v_\infty^2}{2} = \frac{v_{\rm max}^2}{2} - \frac{GM_{\rm II}}{R_H}, \tag{10.26}$$

where $v_{\rm max}$ is the velocity of the particle at the distance of closest approach. Next, the law of angular momentum conservation yields

$$v_{\rm max} R_H = v_\infty b_{\rm II}. \tag{10.27}$$

After utilizing the above two equations and simplifying the resultant expression, we obtain

$$b_{\rm II}^2 = R_H^2 \left(1 + \frac{2GM_{\rm II}}{R_H v_\infty^2} \right). \tag{10.28}$$

In Section 10.1.3, we saw that the velocity of the objects in the immediate aftermath of ejection is constrained by the impactor's velocity. Moreover, if the process of panspermia can only occur through comparatively large ejecta (with sizes of $\gtrsim 1$ m), ceteris paribus, it is seen from (10.11) that lower values of ejection velocity from the donor world are expected. Hence, the residual velocity of the ejecta after leaving the atmosphere of the donor world is generally much smaller than the donor world's escape velocity. This limit is also conventionally employed in most numerical studies (e.g., see Worth et al. 2013), as the residual velocities tend to be $\lesssim 1$ km/s. In this regime, we might consequently work with $v_\infty^2 \lesssim 2GM_{\rm II}/R_H$, provided that the recipient world is not very small. Therefore, after making use of (10.28), we find

$$b_{\rm II}^2 \propto a_{\rm II} M_{\rm II}^{4/3} M_\star^{-1/3}. \tag{10.29}$$

The next two factors that we must account for are Δa and $a_{\rm II}$. In order for panspermia to successfully occur for life-as-we-know-it, the two worlds ought to be ideally located close to the habitable zone (HZ), where liquid water can exist on the surface. If the effective temperature and albedo of a world are specified, it is evident from (4.4) that the orbital radius is proportional to the square root of the luminosity. If we adopt this logic, the separation between the two worlds might also exhibit the same scaling, thus implying $a_{\rm II} \propto L_\star^{1/2}$ and $\Delta a \propto L_\star^{1/2}$, where $L_\star$ represents stellar luminosity. We rewrite $L_\star$ in terms of $M_\star$ using the generic mass-luminosity relationship $L_\star \propto M_\star^3$ from the previous chapters.

After synthesizing the preceding scaling relations, (10.24) transforms into

$$\mathcal{P}_{\rm I\to II} \sim 2.6 \times 10^{-2} \left(\frac{M_{\rm II}}{M_\oplus}\right)^{4/3} \left(\frac{M_\star}{M_\odot}\right)^{-11/6}, \tag{10.30}$$

where we have adopted the normalization to preserve consistency with Worth et al. (2013). We are now in a position to test the validity of the formula. Let us consider Earth-to-Mars and Mars-to-Earth transfers; the ratio of the impact probabilities is then expressible as

$$\frac{\mathcal{P}_{\rm E\to M}}{\mathcal{P}_{\rm M\to E}} \sim \left(\frac{M_{\rm Mars}}{M_\oplus}\right)^{4/3} \sim 5.1 \times 10^{-2}. \tag{10.31}$$

If we use the data from Table 10.1, the corresponding ratio is approximately 6.9×10^{-2}, which is close to the above analytical result.

However, one of the issues with this model is that it yields problematic results for low-mass stars. To illustrate our point, let us consider the famous TRAPPIST-1 system. We will tackle the fraction of objects ejected from TRAPPIST-1f that will land on TRAPPIST-1d. Upon substituting $M_{\rm II} \approx 0.297\, M_\oplus$ and $M_\star \approx 0.089 M_\odot$ in (10.30), we obtain $\mathcal{P}_{\rm f\to d} \approx 0.43$. In comparison, numerical simulations performed by Krijt et al. (2017) have yielded $\mathcal{P}_{\rm f\to d} \sim 0.1$, with the exact value exhibiting a dependence on the residual velocity of the ejecta. For other planets of the TRAPPIST-1 system, we end up with probabilities even higher than unity, which is clearly nonsensical. This mathematical artifact arises because sufficiently small ejection velocities (v_∞) around low-mass stars drive $b_{\rm II} > 2\Delta a$, thus yielding $\mathcal{P}_{\rm I\to II} > 1$ as seen from (10.24).[5] In this regime, a more accurate formula for the capture probability is given by (10.34).

The next quantity that we wish to tackle is the average transit time, $\langle t_0 \rangle$. Let us start by defining the difference in the orbital velocities of the recipient and donor worlds by $v_{\rm rel}$, and the distance between them is Δa. Thus, we can interpret the orbit crossing time as $\tau_{\rm rel} \sim \Delta a / v_{\rm rel}$. However, only a small fraction of the ejecta that are capable of crossing over to planet II will actually impact it. If we specialize to orbits in the *same* plane (i.e., coplanar orbits), the probability of intersecting with the planet's gravitational sphere of influence is the ratio of the impact parameter to the separation between the donor and recipient planets—that is, it is given by $\mathcal{P}_{\rm orb} \sim b_{\rm II}/\Delta a$. Note that this expression closely resembles the transit probability derived in Section 6.1. By combining these relations, the average transit time is interpreted as being proportional to the time taken for orbital transfer divided by the likelihood of intersecting the planet, which yields

$$\langle t_0 \rangle \propto \frac{\tau_{\rm rel}}{\mathcal{P}_{\rm orb}} \sim \frac{(\Delta a)^2}{b_{\rm II} v_{\rm rel}}. \tag{10.32}$$

Finally, we make use of (10.29) and $\Delta a \propto M_\star^{3/2}$ from earlier, and utilize the scaling $v_{\rm rel} \propto M_\star^{-1/4}$ for the relative orbital velocity, which is motivated by (10.71) and the subsequent text. By substituting these relations in (10.32), we obtain

5. In fact, by inspecting (10.28), we see that $b_{\rm II} \to \infty$ in the limit $v_\infty \to 0$. It is thus apparent that the formula cannot be employed for arbitrarily low values of v_∞

$$\langle t_0 \rangle \sim 5.2 \times 10^6 \text{ yr} \left(\frac{M_{\rm II}}{M_\oplus} \right)^{-2/3} \left(\frac{M_\star}{M_\odot} \right)^{8/3}, \tag{10.33}$$

with the normalizing factor chosen to maintain consistency with Worth et al. (2013). Let us consider the ratio of the transit times from Mars-to-Earth and Mars-to-Venus. After applying the above formula, the ratio is determined to be ~ 0.87. If we utilize the data from Table 10.1, the corresponding ratio is around 0.9. Next, let us consider the TRAPPIST-1 system and hold the planetary mass fixed, as most of the planets are reasonably close to our planet's mass. However, due to the $M_\star^{8/3}$ factor in (10.33), we find that the average transit time is $\sim 10^3$-10^4 yr for the TRAPPIST-1 system. This result is compatible with the simulations of Krijt et al. (2017), who concluded that the accretion of test particles by the recipient planets reached saturation at timescales of $\sim 10^2$–10^3 yr.

Before proceeding further, we note that Lingam and Loeb (2017a) proposed a slightly different theoretical model in which $b_{\rm II}$ was specified to be the Hill radius sans any gravitational focusing. Under this assumption, it can be verified that the impact probability is transformed into

$$\mathcal{P}_{\rm I \to II} \sim 2.6 \times 10^{-2} \left(\frac{M_{\rm II}}{M_\oplus} \right)^{2/3} \left(\frac{M_\star}{M_\odot} \right)^{-2/3}, \tag{10.34}$$

whereas the expression for the average transit time, derived by following the same procedure outlined earlier, is given by

$$\langle t_0 \rangle \sim 5.2 \times 10^6 \text{ yr} \left(\frac{M_{\rm II}}{M_\oplus} \right)^{-1/3} \left(\frac{M_\star}{M_\odot} \right)^{25/12}. \tag{10.35}$$

It is left as an exercise to the reader to derive these scalings and check their accuracy against the numerical results delineated in Table 10.1.

One of the major qualitative conclusions that can be inferred thus far from our analysis is that dynamical constraints appear to strongly favor tightly packed planetary systems around M-dwarfs such as TRAPPIST-1 with its multiple planets in the HZ. The reasons are twofold: the probability of ejecta impacting the recipient world is higher by ~ 2 orders of magnitude with respect to the Earth-to-Mars transfer and the transit times are ~ 3 orders of magnitude shorter. Yet, we caution that these two factors do not automatically imply that M-dwarfs are more conducive to interplanetary panspermia.

The existence of sufficiently high impactor fluxes and the likelihood of surviving the harsher radiation environments in the vicinity of M-dwarfs are two potential obstacles that merit further study.

It is necessary to reiterate at this juncture that our toy model has been deliberately pared down to retain dependence on merely two parameters, namely, the masses of the recipient planet and of the host star. In actuality, the orbital parameters of planets (e.g., eccentricity and inclination), the orbital resonances, and the velocities of ejecta will play crucial roles. Nevertheless, despite the preceding simplifications, we have seen that our results appear to yield qualitatively consistent results with more sophisticated dynamical studies.

Although we have hitherto focused on the general principles, we will bring our discussion of interplanetary panspermia to a close by delving further into the transfer of putative microbes from Mars to Earth. We will adopt the methodology and numbers described in Mileikowsky et al. (2000b) as they serve as a useful benchmark for future studies. The number of viable microbes (N_V) landing on Earth is given by

$$N_V \sim \bar{n} \cdot m_\infty \cdot N_{\rm ej} \cdot f_{M\to E} \cdot \chi . \tag{10.36}$$

The first factor on the right-hand side ($\bar{n}$) quantifies the unknown number of Martian microbes per kilogram of rock (units of kg^{-1}). The second (m_∞) and third ($N_{\rm ej}$) variables on the right-hand side represent the mass of a single ejecta and the number of ejecta, respectively. Moving on, we note that $f_{M\to E}$ was defined in (10.22) and measures the probability of ejecta impacting Earth. The last term on the right-hand side (χ) signifies the likelihood of survival of microbes per ejected rock and can be estimated via (10.14). It is apparent that N_V depends strongly on the transit time t_0 as well as the sizes of the ejecta, as longer transit times and small objects yield low survival fractions.

We will consequently work with ejecta that are meter-sized, as they offer shielding for the fiducial timescale of $t_0 \sim 10^5$ yr. In addition, objects of this size are often sufficiently massive to avoid the majority of their mass being lost during atmospheric entry via ablation. Hence, with this choice we have $m_\infty \sim 1.3 \times 10^4$ kg, and empirical data from cratering rates indicates that $N_{\rm ej} \sim 1.5 \times 10^{10}$ for Martian ejecta over the first ~ 0.5 Gyr of the Solar system. Next, after substituting $t_0 \sim 10^5$ yr in (10.22), we obtain

$f_{M\to E} \sim 6 \times 10^{-4}$. For the last factor—namely, the survival fraction—we adopt $\chi \sim 7.3 \times 10^{-2}$ (Mileikowsky et al. 2000b). By collecting these numbers together, we arrive at

$$N_V \sim 8.5 \times 10^{20} \left(\frac{\bar{n}}{10^{11}\,\mathrm{kg}^{-1}} \right), \tag{10.37}$$

where the fiducial value of $\bar{n}$ is the geometric mean of endolithic microbial densities sampled in the hyperarid regions of the Atacama Desert (Ziolkowski et al. 2013). Other studies might yield even higher values of N_V if the survival fraction is modeled as being close to unity.

In the same spirit, a sizable fraction of ejecta from Earth will land on the surface of the Moon, where they will be subjected to mixing by subsequent impacts. By implementing an approach analogous to the one presented above, Armstrong et al. (2002) determined that the mean abundance of material derived from Earth ejecta on the lunar surface is a few parts per million. Hence, in principle, it is feasible and expedient for future missions to extract and analyze lunar samples, thereby yielding valuable information concerning the Earth's geochemical and perhaps biological features during the Hadean and Archean eons.

10.2.2 Interstellar panspermia

There exists general consensus that interstellar panspermia has a reduced likelihood relative to its interplanetary variant. Qualitatively speaking, this line of reasoning appears reasonable as the probability of the ejecta from one planetary system impacting a planet orbiting another star is low. Second, owing to the much longer distances involved, the transit time might increase by orders of magnitude relative to interplanetary panspermia, except for some special cases that we shall encounter shortly. In quantitative terms, if we replace Δa in (10.24) and (10.32) with the interstellar distance $d_\star$, we see that the impact probability could be significantly suppressed while the converse is true for the transit time; recall that a longer transit time dampens the prospects for microbial survival in accordance with (10.14).

The number of investigations undertaken in interstellar panspermia are comparatively fewer. We will first summarize the results from orthodox lithopanspermia studies before addressing some potential alternatives.

10.2.2.1 Lithopanspermia across interstellar distances

One of the first modern explorations of interstellar lithopanspermia was carried out by Melosh (2003). The process of panspermia as envisioned by most authors actually comprises two distinct stages. In the first stage, the recipient planetary system is assumed to either possess a Jupiter-sized planet or be a stellar binary. The Jupiter-star (or star-star) system acts as a gravitational "fishing net" that traps incoming particles. After they are captured in bound orbits, the second stage entails gravitational scattering and instabilities that collectively drive the collision of these objects with a terrestrial planet.

On the basis of a suite of numerical simulations, Melosh (2003) concluded that the number of objects ejected from our Solar system that could have impacted a terrestrial exoplanet in a system with a Jovian planet was $\lesssim 10^{-4}$ over the past ~ 1 Ga. It was consequently argued that the likelihood of interstellar panspermia is extremely low. However, this result was predicated on two nontrivial assumptions: (1) the ejection velocities of objects from our Solar system were typically on the order of a few kilometers per second, and (2) the stellar density ($n_\star$) in the solar neighborhood is low ($n_\star \sim 0.1$ pc^{-3}).

However, neither of these postulates is strictly valid. In particular, when it comes to (2), the majority of stars are expected to have formed in stellar clusters that are characterized by high stellar densities and lower relative velocities, thereby increasing the prospects for panspermia. The timescale for the dispersal of a stellar cluster obeys an empirical law of the form (Adams 2010)

$$t_{\rm disp} \sim 2.3 \times 10^6 \text{ yr} \left(\frac{M_{\rm cluster}}{M_\odot} \right)^{0.6}, \tag{10.38}$$

where $M_{\rm cluster}$ represents the mass of the stellar cluster. If we consider a cluster with $\sim 10^3$ stars, the cluster dispersal timescale might be on the order of ~ 100 Myr. In comparison, the timescale for terrestrial planet formation is ~ 10–100 Myr. Hence, if life originated sufficiently fast ($\lesssim 100$ Myr) on a terrestrial planet, it opens up the possibility of life being transferred to other planetary systems in the same cluster.

This scenario was explored in detail by Adams and Spergel (2005) using a combination of analytical and numerical models. This study determined that the number of lithopanspermia events per cluster was ~ 0.001–1.6; the corresponding estimate per star is found by dividing these bounds by

the number of stars in the cluster (~ 300). An interesting calculation in Adams and Spergel concerns the likelihood of the Earth seeding other stellar systems with life *after* the dispersal of the cluster. The likelihood ($\mathcal{P}_{\rm cap}$) of a stellar binary capturing an object ejected from Earth is

$$\mathcal{P}_{\rm cap} \approx f_B n_\star \langle \Sigma \rangle \sigma_\star \Delta t_a, \tag{10.39}$$

where f_B is the fraction of stars in binary systems, $\langle \Sigma \rangle$ denotes the effective cross-sectional area of the binary for gravitational interactions, $\sigma_\star$ is the velocity dispersion between stars, and Δt_a represents the time lapsed since the origin of life on Earth. An intuitive means of obtaining this equation is as follows. The number of collisions expected is merely the length traversed divided by the mean-free path because the latter signifies the length covered between collisions. Using the fact that the length is $\sigma_\star \Delta t_a$ and the mean-free path is $\left(f_B n_\star \langle \Sigma \rangle\right)^{-1}$, we arrive at (10.39). By substituting $f_B \approx 0.4$, $n_\star \sim 0.1$ pc^{-3}, and $\langle \Sigma \rangle \approx 3$ AU^2 for $\sigma_\star \approx 40$ km/s, and adopting $\Delta t_a \approx 4$ Gyr from Adams and Spergel, we find $\mathcal{P}_{\rm cap} \approx 4.6 \times 10^{-7}$.

The number of rocks potentially bearing life that have been ejected from Earth in this period has been estimated to be $N_B \sim 4 \times 10^{10}$. Therefore, the total number of rocks captured by extrasolar systems is $N_B \mathcal{P}_{\rm cap}/2$ with the factor of 2 introduced to account for the fact that newly ejected objects have traveled over a shorter distance with respect to their predecessors. However, capture into a planetary system is not the same as impacting a terrestrial planet. The above equation must be multiplied by the impact probability ($f_{\rm imp}$), which is assumed to be $\sim 10^{-4}$. In other words, the expected number of lithopanspermia events is $\sim 0.5 f_{\rm imp} \mathcal{P}_{\rm cap} N_B$, and we consequently end up with ~ 0.9. In other words, it is conceivable that life on Earth might have been transferred to another world in a binary system, although its subsequent survival on that world is by no means guaranteed.

The prospects of lithopanspermia among stellar clusters was also investigated by Valtonen et al. (2009), whose approach is delineated below. This work relied on a two-stage scheme to capture objects. In the first stage, let us calculate the cross-sectional area of a planetary system by assuming that the object is captured into an orbit with an orbital radius that is smaller than $a_0 \sim 30$ AU. For a planetary system with an architecture like our own, the choice of 30 AU is motivated by the fact that all of the four giant planets are roughly situated within this distance. As measured in the rest frame of the target system, the velocity of the incoming object at infinity is denoted

by v_∞. We will suppose that $v_\infty^2 \lesssim GM_\odot/a_0$, which is not always guaranteed to be correct. In this event, the corresponding impact parameter is estimated from (10.28) using a_0 and $M_\odot$ in lieu of R_H and $M_{\rm II}$. Given that the corresponding cross-sectional area (Σ_0) is πb_0^2, we find

$$\Sigma_0 \sim 4 \times 10^{-6}\,{\rm pc}^2 \left(\frac{v_\infty}{1\,{\rm km/s}}\right)^{-2}. \tag{10.40}$$

The next step to address is the probability of collision with an Earthlike planet in the sphere of radius ~ 30 AU within the time t'. On the basis of numerical simulations, the probability has been estimated to be $\sim 1.5 \times 10^{-9}\,\left(t'/1\,{\rm Myr}\right)$. By combining this result with the previous expression, the effective cross-sectional area (Σ_T) for impact with the Earth-analog is

$$\Sigma_T \sim 6 \times 10^{-15}\,{\rm pc}^2 \left(\frac{t'}{1\,{\rm Myr}}\right)\left(\frac{v_\infty}{1\,{\rm km/s}}\right)^{-2} \quad 0 < v_\infty \lesssim 0.5\,{\rm km/s}. \tag{10.41}$$

At higher velocities, the following expression extracted from simulations should be utilized instead:

$$\Sigma_T \sim 1.5 \times 10^{-15}\,{\rm pc}^2 \left(\frac{t'}{1\,{\rm Myr}}\right)\left(\frac{v_\infty}{1\,{\rm km/s}}\right)^{-4} \quad v_\infty \gtrsim 0.5\,{\rm km/s}. \tag{10.42}$$

Note that (10.41) and (10.42) are not continuous, as one may have expected. The last piece of information we require before assembling the puzzle is the distribution of v_∞. It is well-known that stars in the solar neighborhood exhibit a Maxwellian distribution for their *relative* velocity with dispersion $\sigma_\star$. If the objects were ejected from the donor system at low velocities, their relative velocity can be approximated by the relative stellar velocity:

$$f\,(v_\infty)\;d^3v_\infty \sim \sqrt{\frac{2}{\pi}}\frac{v_\infty^2}{\sigma_\star^3}\exp\left(-\frac{v_\infty^2}{2\sigma_\star^2}\right)\,dv_\infty \tag{10.43}$$

Let us suppose that the transfer of life is occurring over a timescale t_0. Of this duration, $t_0 - t'$ is spent in interstellar space, whereas t' is expended in the donor and recipient systems. The number density of donor systems is $n_\star$, whereas $\dot{\mathcal{R}}$ represents the ejection rate of objects per system. The number density of objects with ages younger than $t_0 - t'$ is consequently

given by $n_\star \dot{\mathcal{R}} \left(t_0 - t'\right)$. Next, we note that the infinitesimal collisional volume encompassed by the recipient system over the time Δt is $v_\infty \Delta t \, d\Sigma_T$, where $d\Sigma_T \equiv \partial \Sigma_T / \partial t' \, dt'$ is the infinitesimal cross-sectional area. Multiplying this volume with the above number density of ejecta and integrating over t' as well as the velocity yields the number of ejecta ($N_{\rm litho}$) impacting an Earthlike planet. In mathematical parlance, this amounts to solving

$$N_{\rm litho} \sim n_\star \dot{\mathcal{R}} \Delta t \int_0^\infty \left[\int_0^{t_0} \frac{\partial \Sigma_T}{\partial t'} \left(t_0 - t'\right) \, dt' \right] f\left(v_\infty\right) v_\infty \, d^3 v_\infty . \quad (10.44)$$

There are two contributions to this double integral, one from (10.41) and one from (10.42), but it is the former that dominates, provided that $\sigma_\star < 30$ km/s. Hence, to leading order, we may opt to neglect (10.42) in our analysis. The final expression derived by Valtonen et al. (2009) relied on the intriguing coincidence that the product $\dot{\mathcal{R}} t_0^2$ is nearly independent of the size of the ejecta, provided that their sizes are $\sim$ 1–10 m, thus yielding

$$N_{\rm litho} \sim 4 \times 10^{-3} \left(\frac{\dot{\mathcal{R}} t_0^2}{10^{13} \, \text{Myr}} \right) \left(\frac{\Delta t}{1 \, \text{Myr}} \right) \left(\frac{n_\star}{1 \, \text{pc}^{-3}} \right) \left(\frac{\sigma_\star}{1 \, \text{km/s}} \right)^{-2} . \quad (10.45)$$

If we consider the Hyades cluster, we may work with $\Delta t \sim 625$ Myr, $n_\star \sim 2 \, \text{pc}^{-3}$, and $\sigma_\star \sim 0.25$ km/s. After substituting these values into (10.45), we obtain $N_{\rm litho} \sim 10^2$. In the case of long-lived and dense clusters,[6] it is conceivable that this value could be enhanced by about two orders of magnitude. In place of a stellar cluster, let us consider dispersed stars in the solar neighborhood. After specifying $\sigma_\star \sim 40$ km/s, $\Delta t \sim 10$ Gyr, and $n_\star \sim 0.1 \, \text{pc}^{-3}$, we find $N_{\rm litho} \sim 2.5 \times 10^{-3}$.[7] In other words, the prospects for panspermia might be moderate in stellar clusters but improbable in the Galactic field.

---

6. Taking the great age and high stellar density of globular clusters into account, Di Stefano and Ray (2016) suggest that they represent promising sites for panspermia.

7. If we compare this result with the one calculated in the discussion following (10.39), it appears as though the two contradict one another. However, this apparent discrepancy is explained by the fact that the study by Adams and Spergel (2005) was undertaken for stellar binaries (with higher cross-sectional areas), whereas Valtonen et al. (2009) investigated single stars with giant planets.

One of the major implicit assumptions in the above treatment was that the velocities of the objects immediately after ejection from the donor system are sufficiently low (< 1 km/s). At first glimpse, this premise seems problematic since numerical simulations indicate that the mean velocity of ejecta is ~ 5 km/s. However, at the same time it must be recognized that the distribution function of ejecta speeds has a sizable spread, implying that a reasonable fraction of objects would manifest velocities of ~ 0.1–1 km/s. It is precisely this channel that was explored in a recent study by Belbruno et al. (2012). This process is known as *weak escape* and is based on the combination of parabolic-type trajectories and mathematical chaos. In order for weak escape to be initiated, the object must pass close to the largest planet in the donor system, with the final outcome of weak escape being the slow meander of the object away from the system.

Weak capture essentially operates as the inverse of weak escape, thereby enabling the capture of the object at low velocities by the recipient system. The chief advantage with weak escape is that the resultant low velocity greatly enhances the cross-sectional area and probability of capture. Earlier, we have alluded to the fishing net analogy for three-body gravitational interactions involving two massive objects (i.e., the binary) and a massless test particle. The capture cross section for three-body gravitational interactions was worked out in a seminal paper by Heggie (1975). In simplified form, the effective cross-sectional area is expressible as

$$\Sigma_0 \sim 10^8 \ (\mathrm{AU})^2 \left(\frac{a_{12}}{1\ \mathrm{AU}}\right)^{-1} \left(\frac{m_1 m_2}{M_\odot^2}\right)^2 \left(\frac{m_1 + m_2}{M_\odot}\right)^{-1} \left(\frac{v_\infty}{1\ \mathrm{km\,s^{-1}}}\right)^{-7}, \tag{10.46}$$

where a_{12} is the semimajor axis of the binary, while m_1 and m_2 are the masses of the two components of the binary system; for our Solar system, they correspond to the Sun ($m_1 = M_\odot$) and Jupiter ($m_2 \sim 10^{-3} M_\odot$). From inspecting this formula, it is apparent that the cross section exhibits a strong dependence on v_∞, thus greatly favoring the capture of objects endowed with low velocities.

Belbruno et al. (2012) carried out comprehensive numerical simulations to assess the prospects for weak gravitational capture by other systems inhabiting the same cluster in which the Sun formed. We can estimate the number of ejecta that impact an Earth-sized planet orbiting a Solar-type star in the same cluster as follows

$$N_{\mathrm{litho}} \sim N_B \cdot f_{WT} \cdot \mathcal{P}_{WC} \cdot f_{\mathrm{imp}}, \tag{10.47}$$

where f_{WT} is the fraction of life-bearing objects that are amenable to weak escape, and $\mathcal{P}_{WC}$ is the probability of weak capture. Let us adopt $f_{WT} \sim 10^{-2}$, $\mathcal{P}_{WC} \sim 1.5 \times 10^{-3}$, and $f_{\rm imp} \sim 10^{-4}$ in conformity with the values espoused by Belbruno et al. However, unlike this study, we will fix the lower mass bound of ejecta to be $\sim 10^4$ kg instead of 10 kg because the former corresponds to meter-sized objects that not only can shield microbes for a longer period of time but also are less subject to ablation upon atmospheric entry, as noted earlier. By working with this lower limit, we find $N_B \sim 2 \times 10^{11}\,(L_{\rm crust}/1\,{\rm km})$, with $L_{\rm crust}$ representing the mean thickness of the Earth's crust ejected during the window of opportunity prior to the dispersal of the stellar cluster. By substituting these estimates into the above equation, we have

$$N_{\rm litho} \sim 300 \left(\frac{L_{\rm crust}}{1\,{\rm km}}\right). \tag{10.48}$$

Hence, provided that ~ 1 km of the Earth's crust was excavated by impactors after the origin of life and before the dispersal of the cluster, it is plausible that panspermia may have played a vital role in the transfer of life between the constituent planetary systems of a stellar cluster.

10.2.2.2 Other mechanisms for interstellar panspermia

The basic premise underlying lithopanspermia is that rocks derived from the donor planet impact the surface of the recipient planet and thereby assist in seeding life. A few models of interstellar panspermia can be perceived as variants of the classical lithopanspermia model. We will briefly summarize some of their salient features herein.

A theoretical model developed by Napier (2004) bridges lithopanspermia and radiopanspermia, as it combines elements of both. At the first stage, rocks are ejected from the planet after an impactor crashes into it, in accordance with the principles of lithopanspermia. Subsequently, it was suggested that the boulder is subject to erosion due to repeated impacts with dust particles, thus ending up being completely eroded after a timescale of

$$t_{\rm loss} \sim \frac{4R_{\rm ej}\rho_t}{v_\infty \Gamma_d \rho_d}, \tag{10.49}$$

where ρ_d is the density of dust grains in the ambient medium and Γ_d represents the ratio of the mass excavated from the rock after impact with a dust particle of mass m_d. It was argued that $t_{\rm loss} \sim 10^4$–10^5 yr for meter-sized

boulders in the interplanetary medium of the Solar system, with the final result of repeated collisions being the production of small ($\sim 1\ \mu$m) particles that are transported away by means of radiopanspermia. In this model, the Solar system comprises a steady-state population of $\sim 10^{21}$ microbes in a cloud extending to 5 pc. During the passage of the Sun through molecular clouds, the microbes are assumed to be deposited in star-forming nebulae where they seed other systems with life. In light of the constraints imposed by shielding against cosmic rays adumbrated in Section 10.1.3, it is unclear whether the survival of viable microbes over long timescales is feasible for small grains of the kind described above.

In some respects, the proposal by Wallis and Wickramasinghe (2004) mimics conventional lithopanspermia, but it deviates from the norm when it comes to the mode of microbial delivery to the recipient world(s). In this stage, the ejecta are subjected to fragmentation into millimeter-sized particles after hypervelocity collisions with grains in protoplanetary discs. These particles are subsequently accreted by planets during the process of their formation. Let us suppose that $\dot{M}_{\rm ej}$ is the total mass of biogenic material departing the Solar system per year. If a protoplanetary disc with an effective radius (R_{pp}) passes by the Sun at a distance of $d_\star$ with relative velocity $\sigma_\star$, the time over which the dispersal of biogenic material occurs is modeled as $\sim d_\star/\sigma_\star$, implying that the total expelled mass from the Solar system in this period is $\sim \dot{M}_{\rm ej}\, d_\star/\sigma_\star$. As a consequence, the total mass of material delivered to the target (M_t) is estimated as

$$M_t \sim \left(\frac{\pi R_{pp}^2}{4\pi\, d_\star^2}\right) \times \left(\frac{d_\star \dot{M}_{\rm ej}}{\sigma_\star}\right) \sim \frac{R_{pp}^2 \dot{M}_{\rm ej}}{4 d_\star \sigma_\star}. \tag{10.50}$$

By choosing $R_{pp} \sim 50$ AU, $d_\star \sim 1$ pc, $\dot{M}_{\rm ej} \sim 4 \times 10^3$ kg/yr, and $\sigma_\star \sim 20$ km/s, a value of $M_t \sim 3$ kg is obtained. A number of uncertainties arise with regard to this model, including the risks posed by radiation, hydrolysis, and desiccation to microbes in millimeter-sized fragments as well as the efficiency of delivery from the protoplanetary disc to the eventually formed planets.

A different tactic was adopted in Ginsburg et al. (2018). The discovery of the interstellar object 'Oumuamua in 2017 (Meech et al. 2017), in conjunction with the interstellar comet 2I/Borisov in 2019 (Guzik et al. 2020), permitted a tentative estimate of the number density of objects with

radii $\gtrsim 0.1$ km (i.e., sizes comparable to 'Oumuamua). Rather unexpectedly, the inferred number density of $\sim 2 \times 10^{15}$ pc^{-3} (Do et al. 2018) is much higher than prior theoretical predictions.[8] Motivated by the possibility that interstellar objects are more common than expected, Ginsburg et al. developed an analytical model to determine the number of life-bearing interstellar objects that could have been captured by stellar binaries in the Milky Way over its total age of $\sim 10^{10}$ yr. There are two important limitations to bear in mind: the number calculated was a loose upper bound, and the estimate only corresponds to the objects being captured into the stellar system and *not* automatically impacting a terrestrial planet therein. The maximum number of potentially life-bearing objects ($N_{\rm max}$) estimated by these authors was

$$N_{\rm max} \sim 10^5 \left(\frac{\tau_{\rm crit}}{10^6\,{\rm yr}} \right)^4 \left(\frac{\sigma_{\rm ISO}}{100\,{\rm km\,s^{-1}}} \right)^{-3} \left(\frac{R_{\rm ISO}}{0.1\,{\rm km}} \right)^{-3}, \qquad (10.51)$$

where $\sigma_{\rm ISO}$ and $R_{\rm ISO}$ are the velocity dispersion (i.e., characteristic velocity) and radius of the interstellar objects, respectively; we previously introduced the microbial survival time $\tau_{\rm crit}$. This formula exhibits a higher degree of accuracy for $\tau_{\rm crit} \lesssim 10^6$ yr and $\sigma_{\rm ISO} \gtrsim 100$ km s^{-1}. Outside these limits, Figure 10.2 should be utilized, but this plot was calculated for objects with sizes larger than 'Oumuamua by Ginsburg et al. In (10.51), the power-law exponent of -3 governing the cumulative size distribution was interpolated based on observational studies of interstellar dust grains (Landgraf et al. 2000), interstellar meteors (Siraj & Loeb 2019a), and 'Oumuamua and 2I/Borisov (Rice & Laughlin 2019; Jewitt et al. 2020); we will deploy the same exponent hereafter.

It is advantageous to carry out a related calculation along the following lines. The number density of interstellar objects will clearly be a function of their size and velocity because larger objects are fewer in number, and ejection at velocities of $\gtrsim 100$ km/s is rare. For the sake of simplicity, let us assume that the overwhelming majority of objects are ejected with velocities on the order of $\lesssim 10$ km/s, and we may adopt $\sigma_{\rm ISO} \sim 20$ km/s after accounting for the additional contribution from stellar dispersion. As we have posited that most objects have similar velocity, their number

8. A caveat is that the empirical number density was derived under the assumption of isotropy, owing to which it may be incorrect.

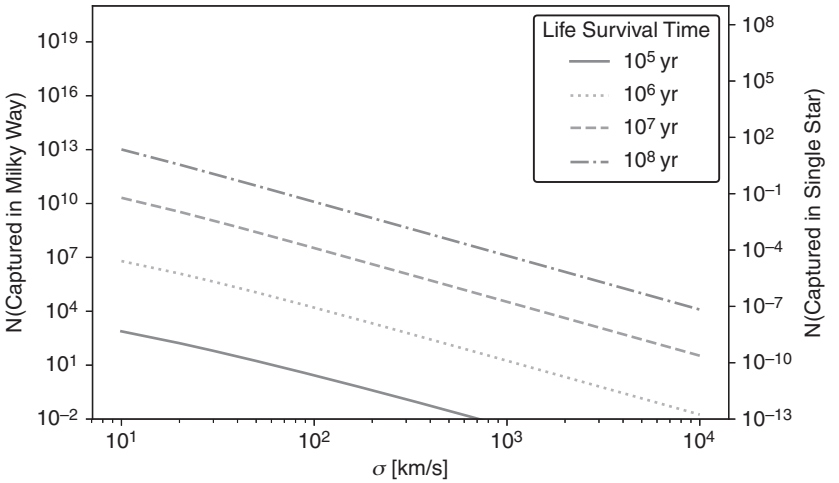


Figure 10.2 Number of captured interstellar objects with sizes > 0.1 km that might potentially host life as a function of their total velocity dispersion. The various curves correspond to different choices of the characteristic microbial population survival time (denoted by $\tau_{\rm crit}$). One can observe that, broadly speaking, higher velocities and shorter survival times result in fewer captured objects. (© The American Astronomical Society. *Source:* Idan Ginsburg, Manasvi Lingam, and Abraham Loeb [2018], Galactic panspermia, *Astrophysical Journal Letters,* 868[1]: L.12, fig. 1.)

density ($n_{\rm ISO}$) only depends on the size. We will work with the ansatz $n_{\rm ISO} \sim 2 \times 10^{15}\ {\rm pc}^{-3}\,(R_{\rm ISO}/0.1\,{\rm km})^{-3}$, where the power-law dependence on $R_{\rm ISO}$ is incorporated on the basis of (10.51). Note that $n_{\rm ISO}$ measures the number density of objects whose sizes are $\gtrsim R_{\rm ISO}$, with the normalization specified by invoking Do et al. (2018).

Let us tackle the case where these free-floating interstellar objects strike an Earth-sized planet. For the sake of convenience, we shall ignore the effects of gravitational focusing and other gravitational interactions. The cross-sectional area is therefore merely the geometric cross section of $\pi R_\oplus^2$. The number of collisions over a time of $\Delta t \sim 10^{10}$ yr is obtained from the product of the capture rate and Δt; the final expression is approximately equal to $\pi n_{\rm ISO} R_\oplus^2 \sigma_{\rm ISO} \Delta t$. We have not grappled with the particular mechanisms that drove the formation and ejection of these interstellar objects, as it has not directly concerned us thus far. Without tackling this detail explicitly, let us suppose that a fraction $f_{\rm bio}$ of these objects actually host life; this fraction can depend on $R_{\rm ISO}$. Lastly, when the object impacts the

planet, it will enter through the atmosphere and strike the surface, during which microbes in life-bearing objects could be subjected to extinction. We explored this issue in Section 10.1.3, wherein it was indicated that a shielding on the order of $\gtrsim 1$ m might suffice to protect the majority of microbes. We shall denote the survival fraction of biologically active objects after the collision by $f_{\rm surv}$.

By assembling all of the above facets, the total number of life-bearing objects that are captured by the planet ($N_{\rm bio}$) over its entire habitable duration is

$$N_{\rm bio} \sim \pi R_{\oplus}^2 \cdot n_{\rm ISO} \cdot \sigma_{\rm ISO} \cdot \Delta t \cdot f_{\rm bio} \cdot f_{\rm surv}. \tag{10.52}$$

A necessary, but *not* sufficient, condition to facilitate panspermia through this avenue is $N_{\rm bio} \gtrsim 1$. By rearranging the expression and choosing a fiducial value of ~ 0.1 for $f_{\rm surv}$, we find

$$f_{\rm bio} \gtrsim 1.8 \times 10^{-7} \left(\frac{R_{\rm ISO}}{1\,{\rm m}}\right)^3 \left(\frac{\sigma_{\rm ISO}}{20\,{\rm km/s}}\right)^{-1} \left(\frac{\Delta t}{10\,{\rm Gyr}}\right)^{-1} \left(\frac{f_{\rm surv}}{0.1}\right)^{-1} \tag{10.53}$$

As per (10.53), at least one meter-sized object out of ~ 10 million must carry viable microbes for interstellar panspermia to be deemed effective. The requisite fraction increases sharply with radius and exceeds unity for kilometer-sized objects; the threshold at which $f_{\rm bio} > 1$ occurs is $R_{\rm ISO} \gtrsim 175$ m. In contrast, the requirement is potentially much less stringent for meter-sized objects, which may still possess the capacity to protect microbes for timescales as long as $\sim 10^8$ yr, as seen from (10.18). We note that (10.53) arguably represents a conservative estimate, seeing as how we have merely considered direct collisions and ignored subtler possibilities such as capture into a planetary system via the gravitational fishing net, followed by interplanetary panspermia.

In summary, there are reasonable grounds to believe that panspermia might be feasible on interstellar scales, albeit perhaps only in special environments. We have already mentioned stellar clusters, but another promising environment is the center of the Milky Way (H. Chen, Forbes, et al. 2018). The attractive feature of the Galactic center insofar as panspermia is concerned is the high stellar density, thereby yielding distances between neighboring rocky planets on the order of 500–5000 AU. As this range is comparable to the aphelion of trans-Neptunian objects such as 90377 Sedna (~ 1000 AU) as well as comets in the Oort cloud, it is conceivable

that interstellar panspermia in this region of the Milky Way would operate in a manner akin to interplanetary panspermia within our Solar system.

To round off our analysis, it is tempting to wonder whether intergalactic panspermia is feasible. In order to cover distances of ~ 1 Mpc over a span of ~ 1 Gyr, speeds of $\sim 10^3$ km/s are mandatory. Numerical simulations of black hole mergers and disruptions of binary stars by supermassive black holes have established that objects can be flung outward at velocities of $\gtrsim 10^4$ km/s (Ginsburg et al. 2012; Guillochon & Loeb 2015b). Thus, in principle, intergalactic panspermia is possible as far as temporal constraints are concerned. However, the number of objects produced at these velocities is relatively very small and the likelihood of capture is extremely low, thereby suggesting that this route is likely to remain only a mathematical possibility. Yet another potential route involves *directed* panspermia on intergalactic scales, although this scenario calls for the existence of technologically sophisticated ETIs with the ability and desire to send relativistic probes. We will not speculate on this matter further—instead, we shall proceed to tackle various relativistic propulsion technologies, and their prospects for actualization by humanity, in the upcoming sections.

10.3 SEEKING POTENTIAL SIGNATURES OF PANSPERMIA

In order to properly assess the viability of panspermia as a genuine scientific hypothesis, it is mandatory to identify diagnostics by which this conjecture may be rigorously tested. In this particular section, we will restrict ourselves to sketching prospective methods for appraising the feasibility of panspermia.

In the event that other samples of life are discovered within our Solar system, several avenues are suitable, in principle, for gauging whether the putative lifeforms share common ancestry with life on Earth. At the most basic level, we can analyze whether the constituent biomolecules of exolife exhibit chiral features that are opposite to Earth-based life—that is, finding right-handed amino acids and left-handed sugars might indicate that this exolife originated independently of life on our planet (C. P. McKay 2010). In a similar vein, we could compile an inventory of the biomolecules that these organisms are composed of. For example, if their nucleobases and amino acids differ from those found in Earth's biosphere, this distinction may serve as a smoking gun for independent abiogenesis.

Likewise, we could analyze whether basic functional characteristics universal to Earth-based biology are manifested in these samples. Examples of universality attributed to life-as-we-know-it include utilization of adenosine triphosphate as energy "currency," codon scheme (nucleotide triplets) comprising the genetic code, ribosomes as sites of protein synthesis, and numerous others. If extraterrestrial organisms do not possess the majority of these traits, it may indicate that their origin is not tied to life on our planet. However, in case all of the aforementioned traits are documented in extraterrestrial organisms as well, there are two possibilities: either they share a common origin with Earth-based life or they evolved independently.

In such an instance, however, we will perhaps have the chance to deploy molecular clocks and the techniques of phylogenetics to determine at what point in time these organisms appear to diverge from their counterparts on Earth. If the timing proves to be "recent" as per some appropriate rubric, it might favor shared ancestry via panspermia. On the other hand, if the deviation is traceable to many eons ago, it may bolster the hypothesis that life originated independently but adopted the same pathway. Naturally, as there remains a great deal that is unknown or unresolved about the origin and evolution of life on our planet and elsewhere, the preceding ideas must be imbibed with a healthy dose of skepticism.

It is, of course, patently obvious that all of the prior methods come into play only if we have the luxury of carrying out in situ explorations, optimally involving sample return back to our planet. In order for missions to be designed with this express goal in mind, the identification of suitable targets is of paramount importance. This issue is not particularly problematic within our Solar system, as we have a fairly good idea of what worlds seem most worthy of astrobiological exploration. The most favored candidates are currently Mars, Europa, Enceladus, and Titan. Ideally, one can visit all of these worlds, carry out thorough searches for life, and subsequently apply the metrics explicated in the previous paragraphs in the best-case outcome where extraterrestrial life is discovered.

When it comes to gauging the likelihood of interstellar panspermia, however, we cannot follow the same path espoused above. Either we must look for signatures within our Solar system or carry out remote-sensing surveys. In both instances, the selection of targets and coming up with suitable diagnostics present formidable challenges. We will therefore spend the rest of the section tackling these matters.

10.3.1 Interstellar objects in our Solar system

The statement that we should begin by seeking out interstellar objects in our Solar system to test the hypothesis of interstellar panspermia is fairly self-explanatory. It has been theorized since more than five decades ago that a significant population of free-floating interstellar objects might be trapped in our Solar system; the majority of them would, however, whiz through our neighborhood à la 'Oumuamua.[9]

Using the recently calculated number density of interstellar objects, Lingam and Loeb (2018a) estimated the population of trapped interstellar objects via the gravitational fishing net outlined earlier. By accounting for the capture of interstellar objects and their subsequent ejection due to gravitational instabilities, Lingam and Loeb determined that the steady-state population of captured interstellar objects ($N_{\rm trap}$) might be

$$N_{\rm trap} \sim 6 \times 10^3 \left(\frac{R_{\rm ISO}}{0.1\,{\rm km}} \right)^{-3}. \tag{10.54}$$

Hands and Dehnen (2020) carried out a series of numerical simulations to estimate the steady-state population of captured interstellar objects; in heuristic terms, their result is expressible as

$$N_{\rm trap} \sim 10^5 \left(\frac{R_{\rm ISO}}{0.1\,{\rm km}} \right)^{-3}, \tag{10.55}$$

of which only $\sim 3.3 \times 10^{-4}$ of them are within 6 AU. The discrepancy between the above formulae stems from the fact that the characteristic excess velocities of interstellar objects and residence times within the Solar system are very different for the two models. In absolute terms, these numbers seem quite large, but it must be noted that the population of asteroids with sizes $\gtrsim 0.1$ km in similar orbits (known as Centaurs) is $\sim 10^7$. Hence, locating these objects is truly analogous to searching for the proverbial needle in a haystack.

Fortunately, if the above estimate for $N_{\rm trap}$ is accurate, not all is doom and gloom. When dynamical simulations of the Jupiter-Sun system are

9. To chase down such objects is a difficult undertaking, but not an impossible one from an engineering standpoint (Seligman & Laughlin 2018; Hein et al. 2019).

carried out in conjunction with a population of interstellar objects, it is found that a substantial fraction of them exhibit unusual orbital parameters. The most prominent variable among them is the orbital inclination i—that is, the angle between the orbital plane and the ecliptic (which is coplanar with Earth's orbit). Numerical simulations carried out by Siraj and Loeb (2019b) indicate that objects at distances of $3 \lesssim a \lesssim 30$ AU with $i > 48°$ are about ten times more likely to be interstellar in nature relative to Centaurs. This study also contended that the total number of interstellar objects of sizes $\gtrsim 0.1$ km that meet the above criterion is $\sim 4 \times 10^3$. The upcoming Vera C. Rubin Observatory might therefore be able to detect > 100 objects of sizes 0.1–10 km at distances of $\sim$ 5–20 AU. A different conclusion was, however, reached by Hands and Dehnen (2020), as the majority of interstellar objects were determined to have orbits akin to those in the Oort cloud.

Apart from identifying interstellar objects through their orbital parameters, another method involves the analysis of their chemical composition (Lingam & Loeb 2018a). Oxygen isotope ratios have proven to be particularly useful in this context. Oxygen has three stable isotopes, thus yielding two isotope ratios: $^{17}O/^{16}O$ and $^{18}O/^{16}O$. The plot of these two quantities (i.e., three-isotope plot) in our Solar system has revealed that objects derived from a common source tend to exhibit similar slopes; for instance, meteorites known as carbonaceous chondrites are characterized by a slope of ~ 1. Hence, if the data points collected for putative interstellar objects deviate from the range detected in our Solar system, they may confirm an interstellar origin. This notion gains further credibility from the fact that the ratio of $^{17}O/^{18}O$ for the Solar system is lower than the average Galactic value; thus, finding higher-than-average $^{17}O/^{18}O$ could serve as a reliable indicator of interstellar objects.

Other isotope ratios like $^{12}C/^{13}C$ and $^{14}N/^{15}N$ are also potentially useful for demarcating interstellar objects. For instance, the ratio of $^{12}CO/^{13}CO$ ranges from ~ 85 to 165 in the vicinity of young stellar objects (YSOs), as opposed to average values of $\sim$ 65–70 in the interstellar medium (ISM). Hence, if interstellar objects had originated in, or spent an extended period near, YSOs, it is possible for their carbon isotope ratios to be skewed. Apart from isotope ratios, anomalies in CN-to-OH ratios and abundances of two- and three-carbon molecules have been conjectured to serve as markers for interstellar objects. Measuring chemical composition remotely requires interstellar objects to pass within very close distances of the Sun

(Sun-grazers), as it can facilitate the vaporization of volatiles and formation of long tails that are analyzable through spectroscopy (Forbes & Loeb 2019). However, if too much mass is lost from Sun-grazers, spectroscopic signatures will not be detectable; the well-studied comet ISON represents one such example.

In the event that we identify at least two independent lines of evidence implying that a particular object within our Solar system has an interstellar origin, it can be assigned a high priority for exploration. In situ exploration of interstellar objects would almost certainly yield a panoply of scientific gains if the missions are meticulously executed. At the very least, we would gain a better understanding of protoplanetary discs and planet formation, chemical composition of extrasolar systems, and the effects of cosmic rays, dust grains, and UV on interstellar objects traversing the ISM, to name a few. Given the ubiquity of prebiotic compounds discovered in meteorites, it is not unreasonable to contend that we may hope to find complex organic molecules in these objects. Last, but by no means the least, perhaps we might be lucky enough to find evidence of extant or extinct extraterrestrial life.

In addition to the aforementioned methods for distinguishing trapped interstellar objects, it is feasible to use the Earth's atmosphere as a detector for locating interstellar meteors. The identification of these objects is rendered possible by tracking their positions and velocities, thereby extrapolating their orbital trajectories via numerical simulations. If they are found to have hyperbolic orbits and asymptotic velocities much higher than those determined for Solar system objects, they may point toward an interstellar origin. Siraj and Loeb (2019a) utilized this approach to conclude that the ~ 0.45 m meteor detected at 2014-01-08 17:05:34 UTC was an interstellar object. As meteors burn up in Earth's atmosphere, spectroscopy of their gaseous debris can be carried out, thus enabling us to further place constraints on their origins along the lines prescribed in the prior paragraphs. Owing to the transient nature of meteor ablation in the atmosphere, it is unlikely that direct biosignatures could be discerned.

10.3.2 Statistical tests of panspermia

Hitherto, we have restricted ourselves to discussing the detection of panspermia (interplanetary or interstellar) in the context of our Solar system. A natural extension of this approach is to inquire whether signatures of

panspermia are manifested in exoplanet surveys. This issue is explored in Lin and Loeb (2015) and Lingam (2016a). Both of these studies reached similar conclusions, but we shall focus on the former due to its relative simplicity.

Let us consider two diametrically opposite cases. In the first case, we allow for the existence of panspermia. This situation is analogous to the simple percolation theory model described in Section 8.2.3. It is easier to envision percolation on a lattice. If $\mathcal{P}$ is the probability of life being transferred from one world to another, it is equivalent to assigning the same probability that a bond (i.e., life transfer) exists between two sites (i.e., habitable worlds). As one increases $\mathcal{P}$, it becomes apparent that more bonds start to emerge. One of the most striking and universal features of percolation theory is that there exists a critical probability $\mathcal{P}_c$ where the system transitions from the presence of multiple isolated small clusters to a giant cluster spanning the entire lattice.

Now, let us turn our attention to the second scenario, wherein no transfer of life is possible. In this setting, the only avenue by which the number of inhabited worlds increases is through independent abiogenesis events. Thus, in the first case, we witness the formation of smaller spatial clusters of worlds seeded by panspermia that merge to form larger clusters, whereas this dynamical behavior is noticeably absent in the second case. This difference can be quantified by using the standard two-point correlation function (C_2) defined as

$$C_2\left(\mathbf{x}-\mathbf{y}\right) \equiv \langle \eta_{\text{bio}}(\mathbf{x})\eta_{\text{bio}}(\mathbf{y})\rangle - \langle \eta_{\text{bio}}(\mathbf{x})\rangle^2, \tag{10.56}$$

where $\eta_{\text{bio}}(\mathbf{x})$ represents the density of inhabited planetary systems at location $\mathbf{x}$. Classical percolation models predict that $C_2\left(\mathbf{x}-\mathbf{y}\right)$ has an inverse exponential dependence on $|\mathbf{x}-\mathbf{y}|$, with a correlation length that is governed by the specifics of the panspermia mechanism under consideration (Lingam 2016b). In contrast, $C_2\left(\mathbf{x}-\mathbf{y}\right)$ would vanish in the absence of panspermia as no transfer of life and spatial correlations are expected.

However, the above formulation of C_2 ignores the well-known fact that the distribution of stars is itself not homogeneous as they exhibit signs of clustering. Hence, it is necessary to adjust (10.56) to account for this issue. As a result, one can work with the modified two-point correlation function ($\mathcal{C}_2$) given by

$$\mathcal{C}_2\left(\mathbf{x}-\mathbf{y}\right) \equiv \left\langle \frac{\eta_{\text{bio}}(\mathbf{x})\eta_{\text{bio}}(\mathbf{y})}{n_\star(\mathbf{x})n_\star(\mathbf{y})} \right\rangle - \left\langle \frac{\eta_{\text{bio}}(\mathbf{x})}{n_\star(\mathbf{x})} \right\rangle^2, \tag{10.57}$$

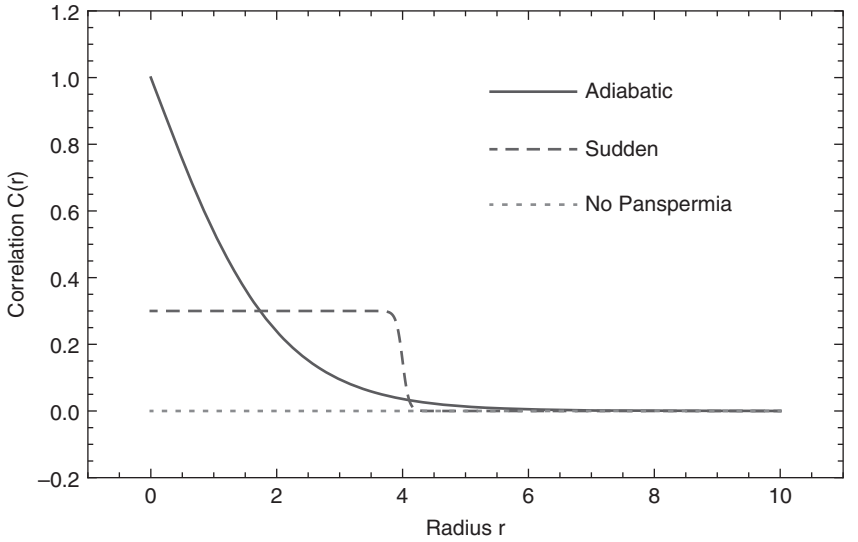


Figure 10.3 Modified two-point correlation function defined in (10.57) for three different cases. The no panspermia case serves as the baseline and is equal to zero, or very close to it. In the adiabatic case, spontaneous emergence of life via abiogenesis and transfer via panspermia are both permitted. In the sudden case, abiogenesis occurs at a single moment in time, and subsequent seeding of the galaxy with life entails panspermia. In both the adiabatic and sudden cases, the correlation function is nonzero and a finite correlation length can be associated with it. (© AAS. Reproduced with permission. *Source:* Henry W. Lin and Abraham Loeb [2015], Statistical signatures of panspermia in exoplanet surveys, *Astrophysical Journal Letters* 810: L3, fig. 2.)

where $n_\star(\mathbf{x})$ is the stellar density at location $\mathbf{x}$. In the case of independent emergence of life, $\eta_{\rm bio} \propto n_\star$, thus ensuring that we recover the baseline case of $\mathcal{C}_2 = 0$ in this limit. Figure 10.3 depicts how $\mathcal{C}_2(r)$ could serve as a useful means of distinguishing between the cases with and without panspermia.

In the future, if exoplanet surveys ideally reveal the existence of sufficiently unambiguous biosignatures on multiple worlds, it might become eventually feasible to estimate $\eta_{\rm bio}$ and thereby determine $\mathcal{C}_2$. In a favorable situation, the Sun may exist at the edge of one of the panspermia bubbles, in which case surveying nearby stars would reveal that about half of the sky is inhabited (with the rest being uninhabited). By investigating this admittedly special scenario, Lin and Loeb (2015) found that as few as ~ 25 targets with confirmed biosignatures of extraterrestrial life may suffice to establish the presence of panspermia with 5σ confidence.

In theory, it is thus possible to undertake investigations of panspermia hypotheses even via remote sensing. In closing, we observe that this test is not strongly dependent on what mechanisms facilitate panspermia. For instance, it may apply even to directed panspermia or terraforming other planets, both of which are a means of spreading life that require interstellar travel by ETIs (Lingam 2016b). In other words, if technosignatures were detected on multiple inhabited worlds, it might be feasible to identify whether these worlds are settlements of the same ETI or technological entities that evolved independently.

10.4 INTERSTELLAR TRAVEL VIA ROCKETS

Let us now turn our attention toward interstellar travel designed by human beings or ETIs. The agents under consideration may be either biological or *post-biological* (e.g., robots equipped with artificial intelligence and 3D printers); it is, however, likely that biological entities will face myriad hurdles due to physiological constraints. Over the past century, numerous propulsion systems have been propounded and actualized. Reviews of this rapidly growing field include Crawford (1990), Frisbee (2003), and Long (2012). In view of the manifold technologies that confront us, there are many valid categorization schemes. One of the most intuitive means of classification is by examining the fuel that powers a spacecraft. If a spacecraft carries its fuel aboard, we will loosely refer to it as a *rocket,* although this nomenclature is not accurate sensu stricto.

We will not address engineering challenges and specifics herein, as our primary focus is on the physical principles underpinning different propulsion systems. For instance, regardless of whether the spacecraft is relativistic or nonrelativistic, the waste heat generated during the burning of rocket fuel must be disposed efficiently; otherwise it may raise the local or global temperature above the melting point of spacecraft materials. Relativistic spacecraft face the additional issue of collisions with dust, gas, and cosmic rays as they traverse the ISM. However, as we demonstrated in Section 9.3.1, even small interstellar probes can surmount this issue by incorporating shielding layers of appropriate thickness.

Prior to embarking on the exploration of select propulsion methods, an interesting point should be borne in mind. If a spacecraft were capable of accelerating continuously at 1 gee (1 gee is equal to 9.8 m/s^2)—which would mimic being on the surface on Earth as per Einstein's theory of

general relativity—for a year, the final speed will be close to the speed of light. Doing the same over a quarter of a century, as observed in the frame of the passengers, will theoretically allow them to nearly traverse the entire observable Universe (Sagan 1963). However, in addition to risks posed by debris impacting the spacecraft at relativistic speeds, constant acceleration over this duration requires a massive amount of energy and is therefore not realizable in practice.

10.4.1 The relativistic rocket equation

Unlike most conventional treatments of the famous rocket equation, which is nonrelativistic, we opt to derive its relativistic counterpart Ackeret 1946); the nonrelativistic expression follows as a limiting case.

Let us suppose that the rocket with rest mass M ejects a reaction mass dm in the opposite direction with a constant *relative* velocity of $v_{\rm ex}$. The resultant velocity of the rocket and the expelled mass are denoted by U and u, respectively, as measured by a stationary observer. As per the laws of special relativity, we have

$$u = \frac{v_{\rm ex} - U}{1 - v_{\rm ex}U/c^2}, \tag{10.58}$$

and taking the limit $c \to \infty$ yields the familiar expression for the nonrelativistic relative velocity. Next, under the assumption that any external forces are negligible, we may apply the conservation of four-momentum from special relativity. The latter is further expressible in terms of the conservation of (relativistic) linear momentum and energy. From the conservation of linear momentum, we have

$$d\left(\frac{M}{\sqrt{1 - U^2/c^2}}U\right) = u\frac{dm}{\sqrt{1 - u^2/c^2}}. \tag{10.59}$$

The Lorentz factors are manifested in the left- and right-hand sides of this equation because the rest masses must be replaced by the relativistic masses. Next, the conservation of energy yields

$$d\left(\frac{M}{\sqrt{1 - U^2/c^2}}c^2\right) = -\frac{dm}{\sqrt{1 - u^2/c^2}}c^2. \tag{10.60}$$

After combining (10.58), (10.59), and (10.60), we end up with

$$d\left(\frac{MU}{\sqrt{1-U^2/c^2}}\right) = -\left[\frac{v_{\rm ex}-U}{1-v_{\rm ex}U/c^2}\right] d\left(\frac{M}{\sqrt{1-U^2/c^2}}\right). \tag{10.61}$$

This expression is simplified further by using the derivative chain rule and collecting terms involving dM and dU, thus giving rise to the unexpectedly simple differential equation

$$\frac{dM}{M} = -\frac{1}{v_{\rm ex}}\left[\frac{dU}{1-U^2/c^2}\right]. \tag{10.62}$$

After integrating the above differential equation from $M = M_i$ to $M = M_f$ and $U = 0$ to $U = \Delta v$, we obtain

$$\ln\left(\frac{M_i}{M_f}\right) = \frac{c}{2v_{\rm ex}} \ln\left(\frac{1+\Delta v/c}{1-\Delta v/c}\right). \tag{10.63}$$

Note that Δv quantifies the increase or decrease in the speed, depending on whether the rocket is accelerated or decelerated, whereas M_i and M_f represent the initial and final total mass of the rocket, respectively. By inverting the above equation and solving for $\Delta v/c$, we end up with

$$\frac{\Delta v}{c} = \tanh\left[\frac{v_{\rm ex}}{c} \ln\left(\frac{M_i}{M_f}\right)\right]. \tag{10.64}$$

At this juncture, let us take the nonrelativistic limit $\Delta v/c \ll 1$ by carrying out a Taylor expansion of the right-hand side. We find that (10.63) reduces to

$$\ln\left(\frac{M_i}{M_f}\right) = \frac{\Delta v}{v_{\rm ex}}. \tag{10.65}$$

Not surprisingly, this result is precisely the seminal rocket equation that is conventionally attributed to Konstantin Tsiolkovsky, one of the true pioneers of astronautics, in 1903. It is worth noting, however, that this equation was derived by several other scientists in the nineteenth century, with one such early notable example being William Moore in 1813. A meticulous historical account of the major milestones in rocketry can be found in von Braun and Ordway (1975).

By incorporating the constraint $v_{\rm ex} \leq c$ imposed by relativity, we find that the mass ratio M_i/M_f must satisfy

$$\frac{M_i}{M_f} \geq \sqrt{\frac{1+\Delta v/c}{1-\Delta v/c}}. \tag{10.66}$$

Of the mass dm that is expelled, let us suppose that a fraction $\epsilon_{\rm kin}$ is utilizable as kinetic energy (Sänger 1953). Therefore, using the fact that the relativistic kinetic energy is $(\gamma_{\rm ex}-1)$ times the available rest mass energy given by $(1-\epsilon_{\rm kin})\,dm\,c^2$, where $\gamma_{\rm ex} = \left(1-v_{\rm ex}^2/c^2\right)^{-1/2}$ represents the Lorentz factor associated with the exhaust in the rocket frame, we have

$$\epsilon_{\rm kin}\,dm\,c^2 = \left[\frac{1}{\sqrt{1-v_{\rm ex}^2/c^2}} - 1\right](1-\epsilon_{\rm kin})\,dm\,c^2. \tag{10.67}$$

Thus, after simplifying this equation, we arrive at

$$\frac{v_{\rm ex}}{c} = \sqrt{\epsilon_{\rm kin}\,(2-\epsilon_{\rm kin})}, \tag{10.68}$$

implying that $v_{\rm ex}/c$ attains a maximum of unity when $\epsilon_{\rm kin} = 1$, which is consistent with physical intuition.

Before embarking on a survey of rocket candidates, it must be noted that the above derivations are idealized because they neglected external forces. This assumption is realistic in regions like the ISM, where gravitational and drag forces are minimal. However, in the vicinity of planets and stars, gravitational forces become important; moreover, when passing through an atmosphere, the drag exerted is nonnegligible.

10.4.2 Chemical propulsion and implications

Rockets reliant on chemical propulsion are broadly classifiable into two categories, depending on whether the propellant is in solid or liquid form. From (10.63) or (10.65), we see that the logarithm of the ratio M_i/M_f is inversely proportional to $v_{\rm ex}$. Hence, increasing the latter by even a modest amount will translate to a significant reduction in the amount of fuel required. The reason is because M_f can be roughly envisioned as the mass of the payload alone, whereas M_i is the mass of the fuel and the payload.

The values of $v_{\rm ex}$ are typically higher for liquid-fuel rockets with respect to their solid-fuel counterparts, owing to which we shall focus on the former.

In his pioneering work, Tsiolkovsky (1903) presented a detailed description of how a suitable fuel (e.g., liquid hydrogen or hydrocarbons) and liquid oxygen could be combined to provide the desired thrust. The basic idea is that the two compounds are subjected to combustion and expelled through the rear of the rocket at supersonic speeds after passing through a nozzle chamber. The effective exhaust velocity under ideal conditions is rather complicated, owing to which we will reproduce only the final expression (Long 2012); it is given by

$$v_{\rm ex} = \sqrt{\frac{2\gamma_a}{\gamma_a - 1}\frac{k_B T_c}{m_s}\left[1 - \left(\frac{P_e}{P_c}\right)^{(\gamma_a - 1)/\gamma_a}\right]}, \tag{10.69}$$

where γ_a is the adiabatic index (i.e., ratio of heat capacities at constant pressure and volume), m_s is the weight of the chemical species used as fuel, P_e is the ambient pressure at the exit of the nozzle, while P_c and T_c are the pressure and temperature in the combustion chamber. It is possible to obtain a more transparent expression if we use $\gamma_a \approx 1.4$ for ideal diatomic gases and assume that $P_e \ll P_c$; the latter is reasonable in certain instances with $P_c \sim 100$ atm. By employing these simplifications, we find

$$v_{\rm ex} \sim 9.3\,{\rm km/s}\left(\frac{T_c}{3000\,{\rm K}}\right)^{1/2}\left(\frac{m_s}{2m_p}\right)^{-1/2}, \tag{10.70}$$

where our choice of normalization for m_s is the mass of molecular hydrogen. From inspecting (10.69) or (10.70), it is apparent that lighter fuels are favored in terms of yielding a higher exhaust velocity. In actuality, as the assumption $P_e \ll P_c$ is not wholly accurate, (10.70) is lowered by a factor of $\gtrsim 2$.

Let us explore the implications of the rocket equation for interstellar travel by ETIs using chemical rockets along the lines of Lingam and Loeb (2018e). If an ETI wishes to exit the planetary system, there are two barriers that must be overcome: the gravitational wells of the planet and the star. In the former case, the escape velocity $v_{\rm esc}$ is given by (10.12). In order to surmount the gravitational well of the star, note that the escape velocity $v_{\rm esc}^{(s)}$ is expressible as

$$v_{\rm esc}^{(s)} = \sqrt{\frac{2GM_\star}{a}} \approx 42.1\,{\rm km/s}\left(\frac{M_\star}{M_\odot}\right)^{-1/4} \tag{10.71}$$

where $M_\star$ is the stellar mass and a is the orbital radius of the planet. The second equality follows by assuming that the rocket is being launched from an Earth-analog in the HZ, whose orbital radius is given by (4.5). It is, however, possible to reduce the fuel costs by launching the rocket parallel to the motion of the planet. As the orbital velocity of the planet is $v_{\rm circ} = \sqrt{GM_\star/a}$, the desired velocity increment for a rocket far removed from the planet's gravity is $v_{\rm esc}^{(s)} - v_{\rm circ} = (1 - 1/\sqrt{2})v_{\rm esc}^{(s)}$. It is instructive to consider the following ratio:

$$\delta_{ps} \equiv \frac{v_{\rm esc}^{(s)} - v_{\rm circ}}{v_{\rm esc}} \approx 1.1\left(\frac{M_\star}{M_\odot}\right)^{-1/4}\left(\frac{R}{R_\oplus}\right)^{-1.35}. \tag{10.72}$$

Hence, if the planet has a radius smaller than or equal to that of the Earth and orbits either G-, K-, or M-type stars, we find that the stellar barrier is more important.

Let us restrict ourselves to worlds with $R \approx R_\oplus$ and determine the mass ratio required to escape the planetary and stellar gravitational wells. In this scenario, the requisite Δv is roughly estimated by summing $v_{\rm esc}$ and $v_{\rm esc}^{(s)} - v_{\rm circ}$.[10] After making use of (10.65), we arrive at

$$\frac{M_i}{M_f} = \exp\left(\frac{\Delta v}{v_{\rm ex}}\right) \approx \exp\left[2.5 + 2.8\left(\frac{M_\star}{M_\odot}\right)^{-1/4}\right], \tag{10.73}$$

where we have chosen $v_{\rm ex} \approx 4.4$ km/s for rockets that utilize liquid oxygen-hydrogen fuel. An inspection of Figure 10.4, depicting the mass ratio as a function of $M_\star$, reveals that M_i/M_f becomes very high for late M-dwarfs. A reasonable cutoff for interstellar travel is $M_i/M_f \approx 10^3$ because a payload mass comparable to the command module of the *Apollo* mission (~ 45 tons) would necessitate a fuel mass that is around 100 times the mass of the ISS. If we impose this constraint, we obtain $M_\star > 0.17M_\odot$.

---

10. This calculation implicitly presumes that the two phases—namely, escaping the gravity of the host planet and the star—are implemented independently of one another. A more accurate treatment would instead require us to specify $\Delta v \approx \sqrt{v_{\rm esc}^2 + v_{\rm circ}^2}$.

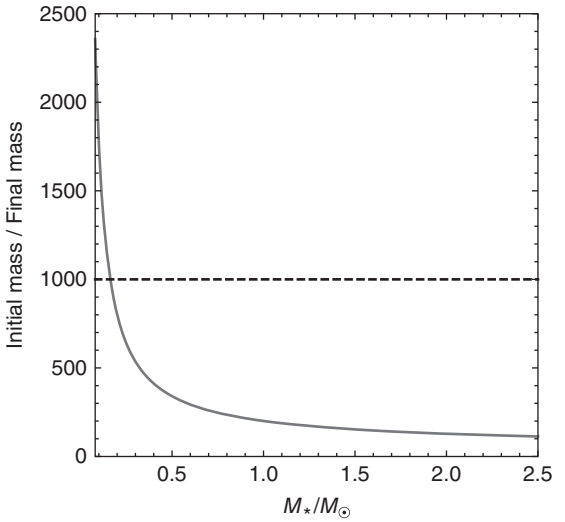


Figure 10.4 The ratio of the initial mass (dominated by fuel) to the final mass (primarily composed of payload) needed to escape from the surface of an Earth-analog in the habitable zone as a function of the host star mass $M_\star$, computed by using (10.73). The dashed line embodies the fiducial mass ratio threshold of 1000, with the intersection of curves taking place at $M_\star \approx 0.17 M_\odot$. (© Manasvi Lingam and Avi Loeb.)

Hence, the above calculation suggests that interstellar travel from late M-dwarfs characterized by $M_\star < 0.15 M_\odot$, a category that includes the famous stars Proxima Centauri and TRAPPIST-1 with planets in the HZ, is difficult by means of chemical propulsion. However, this does *not* imply that putative ETIs situated in such planetary systems would be incapable of interstellar travel. To begin with, our calculation was undertaken for a single-stage rocket, but the results will patently diverge for a multistage rocket (D. F. Spencer & Jaffe 1963). Furthermore, numerous rocket technologies are capable of facilitating high values of $v_{\rm ex}$ encountered subsequently that can greatly lower the fuel-to-payload mass ratio. Finally, alternative propulsion systems that do not require onboard fuel (e.g., light sails) are feasible, as we shall elucidate shortly.

10.4.3 Plasma propulsion

From an inspection of (10.65), for a given value of Δv, it is clear that a high exhaust velocity is desirable to minimize the fuel mass. As opposed to relying on chemical energy for generating the thrust, it is possible to deploy

electromagnetic forces instead. These methods are collectively grouped under the aegis of plasma propulsion. As with other propulsion technologies, the notion of plasma propulsion dates back to Tsiolkovsky's suggestion in 1911 of using electricity to eject particles from rockets; the reader is referred to Choueiri (2004) for an excellent historical account of the development of plasma propulsion. The driving mechanism behind the exerted thrust varies across plasma propulsion systems, with electrothermal, electrostatic, and electromagnetic devices constituting the three most widely investigated approaches. Reviews of this rapidly expanding field can be found in Goebel and Katz (2008), Mazouffre (2016), and Levchenko et al. (2020).

Of the trio, the easiest to understand is electrothermal propulsion. In this category, the role of electrical power is to heat the gaseous propellant to higher temperatures. The advantage in doing so is that the exhaust velocity scales as the square root of the temperature, as seen from (10.69), when all other factors are held fixed. A major limitation of this approach is that achieving very high temperatures may be permissible from the standpoint of electric power generation, but a practical upper bound is imposed by the materials utilized in the chamber. Exceeding temperatures above 3000 K generally calls for the maintenance of steep thermal gradients to avoid damage to the chamber and nozzle walls. Two of the standard thrusters placed in this category are resistojets and arcjets. Their exhaust velocities are comparable to those of chemical rockets with typical values of $\lesssim 10$ km/s.

The underlying principle behind electrostatic devices is also fairly simple. In this case, electric fields are used to accelerate charged particles to high velocities. If we denote the voltage supplied by $\mathcal{V}$ and the mass and charge of the ion by $m_{\rm ion}$ and $q_{\rm ion}$, respectively, the exhaust velocity is expressible as

$$v_{\rm ex} = \sqrt{\frac{2q_{\rm ion}\mathcal{V}}{m_{\rm ion}}}, \tag{10.74}$$

which follows from equating the kinetic and electrostatic energies. In principle, it would seem as though $v_{\rm ex}$ can be raised to arbitrarily high values by tuning $\mathcal{V}$. However, there is a theoretical limit on the current density J_c that is sustainable across a gap d_0. The corresponding value of J_c is known as the space-charge-limited current density and follows from Child's law.

A heuristic derivation of J_c is as follows. Suppose that we consider a parallel-plate capacitor with area A_0 and with a vacuum gap of d_0. The bound charge Q_b on this capacitor is given by $Q_b = \varepsilon_0 \mathcal{V} A_0 / d_0$, which follows from the product of the voltage and the capacitance; as per convention, ε_0 refers to the vacuum permittivity. Now, we will suppose that the free charge entering the gap is comparable to the bound charge. This charge will traverse the gap over the characteristic timescale $d_0/(0.5 v_{\rm ex})$, with $v_{\rm ex}$ given by (10.74). Note that we have employed $v_{\rm ex}/2$ in place of $v_{\rm ex}$, as the latter represents the final velocity, whereas we are interested in the average velocity. Thus, the current across the gap I_0 is the ratio of Q_b and the average time defined above. Finally, by recognizing that the current density J_0 is merely I_0/A_0, we find

$$J_0 \sim \frac{\varepsilon_0}{2} \sqrt{\frac{2 q_{\rm ion}}{m_{\rm ion} \mathcal{V}}} \left(\frac{\mathcal{V}}{d_0} \right)^2 . \tag{10.75}$$

A rigorous calculation of the current density limit yields $J_c \sim 9 J_0/8$, thus establishing that our simple estimate is very accurate. Next, we note that the thrust exerted ($\mathcal{T}$) is given by

$$\mathcal{T} = \dot{m}_{\rm ion} v_{\rm ex} \sim \frac{J A_{\rm jet} m_{\rm ion} v_{\rm ex}}{q_{\rm ion}}, \tag{10.76}$$

with J and $A_{\rm jet}$ denoting the current density and the area of the jet, respectively. The second relation on the right-hand side follows from $\dot{m}_{\rm ion} \sim m_{\rm ion}/\tau_{\rm ion}$, where the timescale $\tau_{\rm ion}$ is defined as $q_{\rm ion}/(J A_{\rm jet})$, with the denominator $J A_{\rm jet}$ representing the current. As there exists an upper bound on the current density derived above, the maximum value of $\mathcal{T}/A_{\rm jet}$ is duly constrained to be

$$\frac{\mathcal{T}}{A_{\rm jet}} \sim \frac{8 \varepsilon_0}{9} \left(\frac{\mathcal{V}}{d_0} \right)^2 . \tag{10.77}$$

An interesting feature of this formula is that it does not depend on the charge-to-mass ratio of ions but merely on the voltage and gap separation. In the presence of a dielectric, ε_0 should be replaced by the dielectric permittivity instead.

Gridded electrostatic ion thrusters (GEITs) are one of the best-known methods for propulsion using electrostatic forces, as described above. As the name suggests, GEITs rely on grids for the purpose of accelerating ions.

In fact, there are two grids present, of which the first is designed to screen out electrons. The second grid is responsible for the acceleration of ions, with a third grid sometimes being employed for refining ion extraction and acceleration. GEITs have been studied extensively and are well suited for interplanetary exploration missions as they are capable of achieving exhaust velocities of $\sim$10–100 km/s; the upper limits are more than an order of magnitude higher than the exhaust speeds of older chemical rockets.

The third, and most variegated, category comprises electromagnetic thrusters. As the name suggests, the ionization and acceleration of charged particles are achieved under the combined action of electric and magnetic forces. Owing to the greater flexibility accorded by electromagnetic forces (as opposed to just electrostatic forces), we will restrict ourselves to a brief picture painted in broad strokes. Magnetoplasmadynamic thrusters (MPDTs) provide a good example of the benefits accruing from electromagnetic propulsion. MPDTs rely on electric arc currents to ionize the incoming gases, and the resulting charged particles are subjected to the Lorentz force, thus experiencing acceleration and compression downstream. The exhaust velocity of MPDTs scales as $I_c^2/\dot{m}_{\text{ion}}$, where I_c represents the arc current. The typical exhaust velocities for MPDTs are $\sim$10–100 km/s, and they also yield high thrust. Despite these advantages, they are often characterized by a short operational life span and modest energy efficiency.

Hall-effect thrusters (HETs) offer a different means of implementing plasma propulsion. As these systems are reliant on the acceleration of charged particles using electric fields, they closely resemble electrostatic propulsion. Owing to this, they are sometimes classified under electrostatic devices. However, magnetic fields play a major role in HETs as they give rise to the Hall current, which is aligned in the direction of $\mathbf{E} \times \mathbf{B}$, where $\mathbf{E}$ and $\mathbf{B}$ denote the electric and magnetic fields, respectively; the Hall current is responsible for the "Hall" in HETs. The electrons spiral along magnetic field lines in HETs and drift in the $\mathbf{E} \times \mathbf{B}$ direction, thus leaving the ions free to be accelerated downstream. HETs are compact devices that can achieve moderate exhaust velocities of $\sim$15 km/s.

Apart from the three basic categories delineated above, there are several plasma propulsion concepts we have not addressed. In recent times, much attention has been directed toward utilizing electromagnetic radiation, either at radio or microwave frequencies, to produce and energize plasma. In particular, the Variable Specific Impulse Magneto Rocket (VASIMR) has garnered interest owing to its potential to reach exhaust velocities of

~ 100 km/s. VASIMR relies on an intricate combination of radio frequency and cyclotron resonance heating in conjunction with the use of magnetic fields to create a magnetic nozzle wherein perpendicular kinetic energy is converted to parallel kinetic energy to generate thrust.

10.4.4 Nuclear fusion and fission propulsion

We will explore the possibility of using either nuclear fusion or fission to design rockets that are theoretically capable of reaching weakly relativistic speeds.

10.4.4.1 Nuclear fusion

Nuclear fusion remains, after more than seven decades, one of the most ambitious (and vexing) challenges in physics and engineering. The actualization of nuclear fusion has been persistently and intensively pursued over the decades because it holds the promise of solving humanity's renewable energy requirements. An auxiliary benefit of achieving sustainable nuclear fusion is that it can be employed to construct rockets that might reach high speeds. To see why this is the case, let us consider one of the most classic fusion reactions:

$$\mathrm{D} + \mathrm{T} \rightarrow {}^{4}\mathrm{He} + \mathrm{n} \tag{10.78}$$

Here, deuterium (D) and tritium (T), two isotopes of hydrogen, combine to yield a helium-4 atom and a neutron. Let us assume that all of the energy generated during this reaction is accessible in the form of kinetic energy. In other words, the fractional mass released during this process equals ϵ_{kin}, where the latter was defined in Section 10.4.1. For the above reaction, we end up with

$$\epsilon_{\mathrm{kin}} \approx \frac{2.014 + 3.016 - 4.003 - 1.009}{2.014 + 3.016} \approx 3.75 \times 10^{-3}. \tag{10.79}$$

Next, we substitute the above value in (10.68) and solve for v_{ex}, thereby ending up with $v_{\mathrm{ex}}/c \approx 8.7 \times 10^{-2}$. Another well-known fusion scheme entails the use of deuterium and helium-3 to yield

$$\mathrm{D} + {}^{3}\mathrm{He} \rightarrow {}^{4}\mathrm{He} + \mathrm{p}. \tag{10.80}$$

Comparing (10.80) with (10.78), it is apparent that the former has the advantage of not producing neutrons, but this is offset by the fact that producing large quantities of helium-3 is difficult. For this reaction, it is easy to show that $\epsilon_{\text{kin}} \approx 3.9 \times 10^{-3}$, which yields $v_{\text{ex}}/c \approx 8.8 \times 10^{-2}$ after utilizing (10.68).

Thus, in both instances an exhaust speed that is a few percent the speed of light is achievable. In view of this datum, suppose that one wishes to attain a speed of $\Delta v = 0.2c$. Choosing $v_{\text{ex}}/c = 0.09$ and employing (10.63), we find $M_i/M_f \approx 9.5$. Hence, it is apparent that the estimated mass ratio is orders of magnitude smaller in comparison to the corresponding values required for chemical or plasma propulsion. The difficulty, however, lies in the practical implementation of nuclear fusion, as noted earlier. Hence, it is worth briefly addressing this issue below, although the reader is recommended to consult Freidberg (2008) for a meticulous treatment.

The two most widely studied approaches for achieving nuclear fusion are magnetic confinement fusion (MCF) and inertial confinement fusion (ICF). A useful heuristic worth mentioning at this juncture is the famous Lawson criterion, which in essence states that

$$n_{\text{pl}} \tau_{\text{pl}} T_{pl} \gtrsim 3.5 \times 10^{28}\ \text{K s m}^{-3}, \tag{10.81}$$

where n_{pl} and T_{pl} are the plasma density and temperature, while τ_{pl} denotes the energy confinement time. The numerical constant on the right-hand side is applicable to the deuterium-tritium reaction presented in (10.78). Numerous figures-of-merit have either extended or reformulated the above criterion, which we shall not describe here. Another useful quantity is the amplification factor $\mathcal{Q}$, which is defined as the ratio of net output power to input heating power. It is only when nuclear reactions obey $\mathcal{Q} > 10$ that they can be deemed practical. As of 2020, there haven't been any nuclear fusion reactors built that simultaneously satisfy the Lawson criterion or variants thereof and $\mathcal{Q} > 10$.

The basic idea behind MCF is that strong magnetic fields act to confine low-density plasma over comparatively large spatial and temporal scales. The magnetic fields trap the ions within the appropriate domain and mitigate thermal losses. The standard MCF device is the tokamak, which has the shape of a torus, but more exotic designs include stellarators (resembling the numeral 8) and the knot-shaped knotatrons (S. R. Hudson et al. 2014).

The Joint European Torus, the world's largest MCF experiment to date, has achieved $\mathcal{Q} \approx 1$. If and when the International Thermonuclear Experimental Reactor (ITER) is completed, it is expected to furnish the first evidence that $\mathcal{Q} \gtrsim 10$ is attainable.[11]

ICF is based on irradiation of the material by high-energy laser pulses. This process initiates ablation of the outermost layers, and the ablation-induced pressure leads to the formation of a central hot spot at very high temperatures. In turn, the hot spot region drives the ignition of the material and the onset of fusion by triggering a burn wave. In contrast to MCF, ICF is characterized by high densities and transient duration of fusion. Apart from MCF and ICF, a number of other alternatives are under active exploration: two examples include inertial electrostatic confinement and magneto-inertial fusion. A recent overview of fusion propulsion and the various methods under consideration for nuclear fusion is provided in Cassibry et al. (2015).

Not many detailed engineering studies have been carried out with regard to fusion propulsion. The most comprehensive among them until recently was Project Daedalus, which arose from detailed investigations conducted between 1973 and 1978 by several members of the British Interplanetary Society; the reader may consult A. Bond and Martin (1986) for a review. Some of the salient features of Project Daedalus are as follows:

- The use of deuterium and helium-3 as fuel was advocated, with the corresponding fusion reaction given by (10.80).
- Two-stage mission design was proposed, with the velocity increments being $\Delta v/c \approx 0.07$ and $\Delta v/c \approx 0.05$. The associated exhaust velocities were $v_{\rm ex}/c \approx 0.035$ and $v_{\rm ex}/c \approx 0.03$.
- The propulsion system relied on the ignition of fuel pellets comprising a mixture of deuterium and helium-3 at a rate of 250 per second using electron beams; this process resembled ICF with electron beams substituting for laser beams. The $\mathcal{Q}$-values for the first and second stages were determined to be 64 and 33, respectively.

11. ITER is one of the largest and most expensive scientific experiments undertaken thus far, with a total mass of $\sim 2.3 \times 10^7$ kg and a predicted final cost upward of $20 billion (US).

- Apart from the problems arising from the realization of ICF, one of the other major issues with Project Daedalus was that it necessitated a very large supply of helium-3. The solution proposed was to mine Jupiter's atmosphere over ~ 20 yr, but this does not seem practical for humans in the near future.

Project Icarus was conceived as an extension of the Project Daedalus concept (Long et al. 2011). A number of its features, such as the use of D and ^{3}He ICF pellets, resemble its predecessor, but Project Icarus differs in a couple of noteworthy respects: it allows for three- and four-stage configurations, and it was optimized to permit a high payload mass.

Two other candidates that merit a mention are Project Longshot and Project VISTA. The former is akin to Project Daedalus in some respects, such as its choice of fuel, and posited a maximum speed of $0.05c$. Project VISTA (Vehicle for Interplanetary Space Transport Applications) was designed for interplanetary travel and can be envisioned as a scaled-down version of Project Daedalus, except that the fuel comprised deuterium and tritium. The estimated exhaust velocity was around 167 km/s, and the goal was to reach Mars in a span of months and all of the outer planets in a timescale of years.

The last method of fusion propulsion that we seek to highlight here is nuclear pulse propulsion. In 1946–1947, Stanislaw Ulam—one of the pioneers of nuclear physics—suggested that nuclear detonations could be used to propel spacecraft, although the antecedents of this idea (albeit not in the context of nuclear energy) date back to the nineteenth century. This notion was subsequently developed by many others and culminated in Project Orion, which was pursued between 1958 and 1965; an excellent historical account can be found in A. R. Martin and Bond (1979). The underlying premise of nuclear pulse propulsion is that the energy source is situated at some distance outside the spacecraft at the time of explosion (via ejection). The expanding debris would impinge on a gigantic *pusher plate* attached to the spacecraft that imparts momentum to it. The impulse absorbed by the spacecraft would cause an acceleration of a few gees, but the crew would be protected by pneumatic shock absorbers situated behind the pusher plate.

The interstellar version of Project Orion, as delineated by Freeman Dyson (1968), relied on deuterium burning (fusion) instead of nuclear fission. The effective exhaust velocity has an upper bound of $v_{\rm ex}/4$, where

$v_{ex}/c \approx 0.09$ for the deuterium-tritium reaction. The factor of 4 is included because the pusher plate intercepts a fraction of all the exploded debris and therefore acquires only a fraction of the total momentum. Let us examine the best-case scenario wherein the exhaust velocity is $2.2 \times 10^{-2}c$ and we specify $\Delta v = 10^4$ km/s. By substituting this expression in (10.63), which amounts to presuming that the rocket equation is valid, we obtain $M_i/M_f \approx 4.4$. Next, we wish to know the number of bombs required for this spacecraft. Dyson argued that the velocity imparted to the ship per bomb ought not exceed $v_b \sim 30$ m/s in order to ensure that the resultant stresses are lower than the tensile strength of the shock absorbers.

Therefore, the number of bombs required is $\Delta v/v_b \approx 3.3 \times 10^5$. If each bomb has a mass of ~ 1 ton, we see that the total mass of the fuel must be $\sim 3.3 \times 10^8$ kg. Using $M_i/M_f \approx 4.4$ (obtained earlier), the mass of the spacecraft sans fuel is on the order of $\sim 10^8$ kg. This estimate is more than two orders of magnitude higher than that of the ISS, implying that the spacecraft could, in principle, transport a small number of humans or other putative biological and post-biological ETIs. The desired time interval between two successive explosions (Δt_e) is found by demanding that a steady acceleration of ~ 1 gee is maintained. This constraint yields $\Delta t_e = v_b/g \approx 3$ s. Dyson (1968) calculated the economic cost of this mission concept and arrived at $\sim$ \$60 to \$600 billion (US); note that the estimate is not up-to-date because it was derived in the 1960s. Although the economic cost seems very high, it is still within the realm of human possibility given that the GDP of the United States is $\sim$ \$20 trillion; another natural point of comparison is the ISS, whose total cost thus far has been $\sim$ \$150 billion (US).

Needless to say, while the nuclear pulse propulsion approach remains technically viable, genuine political, sociological, biological, and ethical obstacles should be taken into account when it comes to utilizing nuclear detonations. In his arresting autobiography, *Disturbing the Universe,* Freeman Dyson offered a strong critique of Project Orion, as the following lines clearly reveal (1979a, pp. 114–115):

> Sometimes I am asked by friends who shared the joys and sorrows of Orion whether I would revise the project if by some miracle the necessary funds were suddenly to become available. My answer is an emphatic no. . . . By its very nature, the Orion ship is a filthy creature and leaves its radioactive mess behind it wherever it goes. . . . Many things that were acceptable in 1958 are no longer acceptable today. My own standards have changed too.

10.4.4.2 Nuclear fission

As we have noted, Project Orion was originally conceived in the context of using *fission* bombs for propulsion. We will not delve into fission in much detail because it appears to have two major disadvantages with respect to fusion. First, as shown below, the maximum exhaust velocities attainable are typically lower. Second, even if the fission products are quickly expelled or situated outside the spacecraft, the danger of radioactive fallout adversely affecting onboard biota cannot be easily dismissed. On the other hand, one of the major pros linked with nuclear fission propulsion is that humans have made greater advances in harnessing fission compared to fusion.

Let us consider the classic example of a fission reaction involving uranium-235. The reaction is expressible as

$$^{235}\mathrm{U} + \mathrm{n} \rightarrow {}^{141}\mathrm{Ba} + {}^{92}\mathrm{Kr} + 3\mathrm{n}. \tag{10.82}$$

As before, we are free to compute the fractional mass deficit, which is equal to $\epsilon_{\rm kin}$ in the idealized limit. After simplification, we end up with $\epsilon_{\rm kin} \approx 7.9 \times 10^{-4}$. It is, therefore, approximately five times smaller than the corresponding values of $\epsilon_{\rm kin}$ calculated for fusion reactions earlier. From (10.68), the maximum exhaust velocity for the above reaction is determined to be $v_{\rm ex}/c \approx 0.04$. In other words, the ideal exhaust velocity for ^{235}U fission is about two times smaller compared to its fusion counterparts. Although a factor of 2 may not seem significant, recall that the mass ratio M_i/M_f has an exponential dependence on $c/v_{\rm ex}$. Hence, even this modest decrease in $v_{\rm ex}$ as one moves from fusion to fission translates to a mass ratio that could be orders of magnitude higher.

Before wrapping up our analysis, a famous fictitious example of fission propulsion is worth highlighting. In Arthur Clarke's seminal *2001: A Space Odyssey* (1968), the spacecraft, aptly named *Discovery,* appears to have relied on nuclear fission. Given the spacecraft design, it has been suggested that its exhaust velocity was conceivably as high as ~ 60 km/s (Long 2012).

10.4.5 Antimatter propulsion

We briefly examined the prospect of using electron-positron reactions to power rockets in Section 9.5.2. Instead, we will focus on rockets powered by proton-antiproton reactions, as they have been investigated in more

detail. One of the first publications that touched on antimatter propulsion for interstellar travel was the wide-ranging study by Shepherd (1952). An early account of this field is presented in Forward (1982), whereas Semyonov (2017) provides a state-of-the-art overview of the breakthroughs and inherent hurdles.

The proton-antiproton reaction yields the following products:

$$\mathrm{p} + \bar{\mathrm{p}} \rightarrow n\pi^+ + n\pi^- + m\pi^0, \tag{10.83}$$

where p and $\bar{\mathrm{p}}$ denote the proton and antiproton, π^+ and π^- are positively and negatively charged pions, and π^0 represents neutral pions. On average, it has been determined that $2n + m \sim 5$. The charged pions decay to yield charged muons ($\mu^\pm$) and neutrinos (ν) via

$$\pi^\pm \rightarrow \mu^\pm + \nu, \tag{10.84}$$

and the muons themselves decay to yield electrons, positrons, and neutrinos as per the reaction

$$\mu^\pm \rightarrow e^\pm + \nu + \bar{\nu}, \tag{10.85}$$

where $\bar{\nu}$ represents antineutrinos. The electrons and positrons, if spatially localized, further annihilate to yield gamma rays.

At relativistic speeds, one can use (10.63) to compute the required mass ratio. For antimatter rockets, a careful analysis by Westmoreland (2010) suggests that $v_{\mathrm{ex}} \approx 0.58c$ is achievable, and even higher values are theoretically possible if the energy carried by gamma rays can be harnessed. Naturally, the upper bound that is realizable is $v_{\mathrm{ex}} = c$, which corresponds to specifying $\epsilon_{\mathrm{kin}} = 1$ in (10.68). Henceforth, we will suppose that the spacecraft is moving at weakly relativistic speeds such that the nonrelativistic rocket equation is reasonably accurate. Our analysis will mirror the treatment in Crawford (1990).

Let us consider the simplified system in which $M_i \approx M_v + M_r$ and $M_f \approx M_v$, where M_v is the mass of the vehicle and M_r is the mass of reaction fluid. The basic idea behind introducing a reaction fluid is that the kinetic energy of the charged pions is used to heat the former; the reaction fluid is subsequently expelled to generate the desired thrust. We can rewrite (10.65) to obtain

$$M_r \approx M_v \left[\exp\left(\frac{\Delta v}{v_{\mathrm{ex}}} \right) - 1 \right]. \tag{10.86}$$

Similarly, if the mass of fuel annihilated is denoted by M_ℓ and $\epsilon_{\rm kin}$ is converted into exhaust kinetic energy, then we have

$$\epsilon_{\rm kin} M_\ell c^2 = \frac{1}{2} M_r v_{\rm ex}^2, \tag{10.87}$$

which is, broadly speaking, the nonrelativistic analog of (10.67). By defining $x_e = \Delta v / v_{\rm ex}$ and substituting (10.86) in (10.87), we arrive at

$$M_\ell \approx \frac{M_v}{2\epsilon_{\rm kin}} \left(\frac{\Delta v}{c} \right)^2 \frac{e^{x_e} - 1}{x_e^2}. \tag{10.88}$$

It is straightforward to demonstrate that M_ℓ attains a minimum when $x_e \approx 1.59$ (von Hoerner 1962); in other words, we find $\Delta v \approx 1.59 v_{\rm ex}$. Note that we have effectively held Δv fixed, implying that $v_{\rm ex}$ is calculated, for a particular choice of Δv, by using the preceding relation. Next, from (10.86), we obtain $M_r \approx 3.9 M_v$ (i.e., the optimum ratio of the reaction fluid and vehicle masses is around 4). Finally, we observe that the mass of antimatter (M_{am}) is one-half of the total annihilated mass M_l. Thus, using this fact along with (10.88), we have

$$M_a \approx \frac{0.39}{\epsilon_{\rm kin}} \left(\frac{\Delta v}{c} \right)^2 M_v \sim \left(\frac{\Delta v}{c} \right)^2 M_v, \tag{10.89}$$

where the last equality follows from noting that $\epsilon_{\rm kin} \sim 0.4$ for antiproton-proton reactions. Hence, in order to accelerate a 10^3 kg vehicle to a speed of $0.1c$, the mass of antimatter required is ~ 10 kg, whereas the mass of reaction fluid must be 3.9×10^3 kg. A more accurate calculation can be undertaken along the above lines that explicitly breaks down the total vehicle mass into the engine ($\sim 2 \times 10^4$ kg), structural ($\sim 0.05\, M_r$), and payload masses. To accelerate a one-ton payload to $0.1c$ and then subsequently decelerate it to rest, $\sim 7.2 \times 10^5$ kg of reaction fluid and $\sim 1.9 \times 10^3$ kg of antimatter are required. The mass of reaction fluid required is quite modest, as it is comparable to the propellant mass of Solid Rocket Boosters used in the Space Shuttle program.

On the other hand, the mass of antimatter required is many orders of magnitude higher than what is currently manufactured per year on Earth. Moreover, current estimates for the economic cost incurred in synthesizing 1 g of antimatter range from billions to trillions of dollars. Lastly, even

if the production of antimatter could be elevated by orders of magnitude, long-term storage of antimatter presents its own issues. Hence, when these limitations are viewed collectively, it seems improbable that humans can harness antimatter propulsion in the next few decades. With that being said, there are no compelling reasons to dismiss the prospects for antimatter propulsion altogether on *physical* grounds.

10.4.6 Laser power

A. A. Jackson and Whitmire (1978) suggest that inert reaction fuel carried by a spacecraft could be ignited via a laser beam transmitted from the vehicle's point of origin.[12] We estimate the beam energy E_b that is required to this effect by positing that the propulsion system is weakly relativistic. If a fraction $\epsilon_{\rm kin}$ of the beam energy is converted into kinetic energy, we have

$$\epsilon_{\rm kin} E_b = \frac{1}{2} M_r v_{\rm ex}^2. \tag{10.90}$$

After substituting (10.86) into the above equation and using the previous definition of x_e, we find that

$$E_b = \frac{M_v \, (\Delta v)^2}{2\epsilon_{\rm kin}} \frac{e^{x_e} - 1}{x_e^2}. \tag{10.91}$$

As before, the optimal value occurs at $x_e \approx 1.59$, implying that the corresponding mass ratio is given by $M_r \approx 3.9 M_v$. For this idealized choice, we see that (10.91) is transformed into

$$E_b \approx 0.772 \, \frac{M_v \, (\Delta v)^2}{\epsilon_{\rm kin}} \approx 1.4 \times 10^{23} \, {\rm J} \left(\frac{M_v}{10^6 \, {\rm kg}} \right) \left(\frac{\Delta v}{c} \right)^2 \left(\frac{\epsilon_{\rm kin}}{0.5} \right)^{-1}. \tag{10.92}$$

If we choose $\Delta v = 0.1c$, the beamed energy required is $E_b \approx 1.4 \times 10^{21}$ J. The requisite energy is clearly very high, as it is close to the annual energy consumption of humanity. Moreover, not all of the energy beamed will

12. On a related note, intense laser beams can cause ablation of material, thereby generating plasma/vapor jets and thrust akin to conventional chemical rockets; the speeds achieved may approach ~ 100 km/s in the near future (Phipps et al. 2010).

actually reach the spacecraft, as we shall see afterward in our discussion of light sails.

It is straightforward to generalize the above result to the relativistic case. If all of the beamed power is transferred to the exhaust jet ($\epsilon_{\rm kin} = 1$), A. A. Jackson and Whitmire (1978) show that the relativistic rocket equation is transformed into

$$\ln\left(\frac{M_i}{M_f}\right) = \frac{[\gamma_{\rm ex}\,(1 + v_{\rm ex}/c) - 1]^{-1}}{2} \ln\left(\frac{1 + \Delta v/c}{1 - \Delta v/c}\right). \tag{10.93}$$

In the limit $v_{\rm ex}/c \ll 1$, it is easy to verify that the first factor on the right-hand side of (10.93) reduces to its counterpart in (10.63).

10.5 INTERSTELLAR TRAVEL WITHOUT ONBOARD FUEL

As we have seen thus far, one of the major concerns with rockets is that they need to carry their own fuel. A number of innovative solutions have been proposed to bypass this issue (Matloff 2005). We will delve into some of the more prominent candidates below. As before, we shall not tackle the engineering aspects for the most part, despite their indubitable relevance.

10.5.1 Interstellar ramjets

The central issue with rockets is that achieving high Δv calls for either an exceptionally high mass ratio, which is cumbersome, or technologies such as fusion and antimatter propulsion. In 1960, Robert Bussard proposed that an "interstellar ramjet" could, instead, opt to scoop up protons from the ISM and utilize them to carry out nuclear fusion and produce thrust; in the latter respect, it resembles fusion rockets (discussed previously). As Bussard's original analysis is somewhat intricate, we will describe a simplified model that dovetails with the treatment presented in Blatter and Greber (2017).

The mass of the ramjet is denoted by M_j and it captures a mass dm. We work in the idealized limit wherein (1) the energy-momentum of the acquired mass is not directly deposited into the spacecraft and (2) the entire energy resulting from nuclear fusion of the captured mass is used for propulsion. The energy harvested for propulsion is given by $\epsilon_{\rm kin}\, dm\, c^2$ (i.e., the fraction of rest mass energy utilizable toward propulsion is $\epsilon_{\rm kin}$). As per these assumptions, momentum conservation yields

$$p_f = p_i + (1 - \epsilon_{\rm kin})\, \gamma_m v_m dm, \tag{10.94}$$

where γ_m and v_m are the Lorentz factor and coordinate velocity of the infinitesimal mass, while p_f and p_i are the final and initial momenta of the ramjet. The second term on the right-hand side is the exhaust momentum imparted to the ramjet. Note that an additional deceleration term associated with stopping the captured mass has been neglected.

By applying energy conservation, we obtain

$$E_f = E_i + \epsilon_{\rm kin} dm\, c^2 - (1 - \epsilon_{\rm kin})\, (\gamma_m - 1)\, dm\, c^2, \tag{10.95}$$

where E_f and E_i are the final and initial energies of the ramjet, while the last term on the right-hand side signifies the energy carried away by the exhaust. Next, we note that $E_f - E_i = \left(\gamma_f - \gamma_i\right) M_j c^2 \approx d\gamma_j M_j c^2$ in the infinitesimal limit; here, γ_j represents the Lorentz factor of the ramjet. Using this relation, the above expression simplifies to

$$M_j d\gamma_j + \gamma_m\, (1 - \epsilon_{\rm kin})\, dm = dm. \tag{10.96}$$

Next, we employ $dm = A_j m_p n_{\rm ISM} v_j\, dt$, where A_j is the collecting area of the ramjet's scoop, m_p is the proton mass, and $n_{\rm ISM}$ is the proton number density in the ISM; in addition, v_j denotes the coordinate velocity of the ramjet, and dt is the differential time step. It is more advantageous to instead express our result in terms of the proper acceleration $\tilde{a}_j = du_j/dt$ and the proper velocity $u_j = \gamma_j v_j$. The physical intuition behind introducing the proper acceleration is that it encapsulates the acceleration experienced by a moving observer in the ramjet. Hence, after making use of $d\gamma_j/dt = v_j \tilde{a}_j/c^2$, we find

$$M_j \tilde{a}_j + \gamma_m\, (1 - \epsilon_{\rm kin})\, A_j m_p n_{\rm ISM} c^2 = A_j m_p n_{\rm ISM} c^2. \tag{10.97}$$

Lastly, we can solve for γ_m from (10.94) in the infinitesimal limit of $p_f - p_i = dp_j$ and substitute it into (10.97), thereby obtaining a quadratic equation for $\tilde{a}$ given by

$$M_j^2 \tilde{a}_j^2 + 2 M_j A_j m_p n_{\rm ISM} u_j^2 \tilde{a}_j - \left(A_j m_p n_{\rm ISM} c u_j\right)^2 \epsilon_{\rm kin}\, (2 - \epsilon_{\rm kin}) = 0. \tag{10.98}$$

By solving this quadratic and selecting the positive root that yields the acceleration, we arrive at

$$\tilde{a}_j = \frac{A_j m_p n_{\mathrm{ISM}}}{M_j} \times u_j^2 \left(\sqrt{1 + \frac{c^2 \epsilon_{\mathrm{kin}} \, (2 - \epsilon_{\mathrm{kin}})}{u_j^2}} - 1 \right). \tag{10.99}$$

It is instructive to contrast the two limiting regimes. Upon evaluating the low-speed limit $u_j \to 0$, we end up with

$$\tilde{a}_j \to \left(\frac{A_j m_p n_{\mathrm{ISM}} c^2}{M_j} \right) \left(\frac{u_j}{c} \right) \sqrt{\epsilon_{\mathrm{kin}} \, (2 - \epsilon_{\mathrm{kin}})}. \tag{10.100}$$

Hence, at very low speeds, we see that the proper acceleration has a linear dependence on the proper velocity to leading order. Next, we are allowed to take the limit $u_j \to \infty$ because u_j does not have an upper bound of c. In this instance, we find

$$\tilde{a}_j \to \epsilon_{\mathrm{kin}} \left(1 - \frac{\epsilon_{\mathrm{kin}}}{2} \right) \left(\frac{A_j m_p n_{\mathrm{ISM}} c^2}{M_j} \right), \tag{10.101}$$

implying that the proper acceleration approaches a constant asymptotic value. It is found that (10.100) and (10.101) are consistent with the more detailed calculations presented in Bussard (1960).

It is left as an exercise for the reader to calculate the time elapsed in the frame of the moving observer (proper time), denoted by τ_f, to reach a proper velocity of u_f via

$$\tau_f = \int_{u_0}^{u_f} \frac{du_j'}{\gamma_j(u_j') \, \tilde{a}_j(u_j')}, \tag{10.102}$$

where u_0 is the initial proper velocity. The distance traversed d_f as measured in the space frame is

$$d_f = \int_0^{\tau_f} u_j(\tau_j') \, d\tau_j', \tag{10.103}$$

where it should be recalled that τ is the proper time. In principle, if the area density M_j/A_j is sufficiently low and the capture process is highly efficient, an asymptotic proper acceleration of ~ 1 gee might be achievable by an interstellar ramjet.

Thus far, our calculations were purely theoretical in nature. In practice, a number of serious issues confront the functioning of an interstellar ramjet. The first pivotal drawback is that p-p fusion reactions are extremely slow,

implying that extremely high densities are required in order to meet the desired reaction rates. One possibility is to scoop up deuterium instead of protons, but this poses another issue: the abundance of deuterium is much lower compared to hydrogen. Instead, it may be feasible to bypass the difficulties stemming from the p-p reaction by using the captured interstellar protons to initiate a catalytic cycle of nuclear reactions (Whitmire 1975). The best-known example is the CNO cycle that occurs naturally in stars and is $\sim 10^{18}$ times faster than the p-p reaction. As the catalytic fuel is not depleted, it can be carried on board the spacecraft.

Second, in order for the ramjet to have a large collection area, most proposals invoke magnetic fields to capture protons in the ISM. It has, however, been determined that the majority of particles would undergo reflection by the magnetic field because of the so-called *magnetic mirror* effect and that the requisite magnetic field would increase by orders of magnitude as the ramjet approaches relativistic speeds (A. R. Martin 1973). If electrostatic fields are employed in place of magnetic fields, some of these problems might be bypassed. Lastly, if the ramjet aims to sustain a terminal acceleration of ~ 1 gee, it is susceptible to material stresses that could come into play and prevent the realization of this goal.

In order to bypass the manifold problems that plague the conventional interstellar ramjet, a number of improvements have been delineated. We will not address them here, but the interested reader may consult Crawford (1990). For example, one proposal advocated the ramjet carrying its own source of energy but utilizing interstellar gas as a reaction fluid. For the same value of $\epsilon_{\rm kin}$, this hybrid vehicle requires a lower mass ratio than conventional rockets, albeit only up to semirelativistic speeds of 0.5–$0.7c$. Other alternatives include using beamed laser power as the ramjet's energy source to burn the fuel and laying down fuel pellets in advance along the ramjet's trajectory that are scooped up by the interstellar ramjet.

10.5.2 Light sails

> But as soon as somebody demonstrates the art of flying, settlers from our species of man will not be lacking. Who would once have thought that the crossing of the wide ocean was calmer and safer than of the narrow Adriatic Sea, Baltic Sea, or English Channel? Given ships or sails adapted to the breezes of heaven, there will be those who will not shrink from even that vast expanse. Therefore, for the sake of those who, as it were, will presently

> be on hand to attempt this voyage, let us establish the astronomy, Galileo, you of Jupiter, and me of the moon.
>
> —Johannes Kepler, "Conversation with the Sidereal Messenger" (1610)

Light sails refer to spacecraft that are propelled by radiation pressure—that is, they harness the momentum carried by photons. The intricate and convoluted history of light sails makes it difficult to definitively identify one individual as the unequivocal inventor of the light sail. As Kepler sagaciously illustrated in his letter to Galileo, he was remarkably close to the right track, but his propulsion method, which is manifestly short on details, was reliant on "breezes" (*aurae* in the original Latin). Toward the end of the nineteenth century, the French science-fiction authors Georges Le Faure and Henry de Graffigny mused about using mirrors for propulsion. In the 1920s, notable scientists such as Konstantin Tsiolkovsky and J. D. Bernal wrote cogently about this issue.

Perhaps the first indisputable technical exposition of using photon propulsion appears in the seminal, but sadly forgotten, paper by F. A. Tsander (1924) nearly a hundred years ago, as the following lines reveal (1964, p. 222):

> When we wish to fly to other planets, velocities of 11–18 km/sec must be attained. In this case it is possible to use a rocket but it will be probably more advantageous to fly with the aid of mirrors or screens made of thin sheets. . . . The mirrors do not require fuel and in case of need they may be used as fuel in the rocket. These are the two advantages of the mirrors; in addition they do not produce large stresses in the material of the ship and have a smaller weight than that of a rocket together with the propellant. However, the mirrors can be damaged by meteors more easily than a rocket.

In this excerpt, as well as in subsequent publications during the 1920s, Tsander correctly identified (and explicitly worked out) not only the propulsion scheme and the benefits stemming from light sails but also some of their potential drawbacks.

The next major set of breakthroughs are traceable to the 1950s by the likes of Carl Wiley and Richard Garwin. It was, however, only in the 1960s that bona fide in-depth and rigorous calculations concerning the deployment of laser beams, as opposed to the Sun, for powering light sails were delineated. The most noteworthy contributions in this period were from

Robert Forward, György Marx, and J. L. Redding. Even so, there are compelling grounds for arguing that the first modern and technically robust formulation of light sails propelled by laser beams appeared in Forward (1984). The reader is referred to Gilster (2004) for an engaging and thorough historical account of light sails commencing with Kepler.

In our subsequent analysis, we will expound some of the salient characteristics of laser-pushed light sails. Alternatively, one may take advantage of high-energy astrophysical sources to accelerate light sails to relativistic speeds while incurring minimal energy expenditure, but we shall not examine this strategy herein; the relevant details are presented in Lingam and Loeb (2020f). We will primarily operate in the non- or weakly relativistic regime (i.e., corresponding to $\lesssim 0.2c$), as the results are physically more transparent, but a few results pertaining to relativistic light sails are also outlined. We will mostly adopt the methodology and notation employed in McInnes (1999) and Lubin (2016a).

10.5.2.1 Nonrelativistic photon propulsion

Although we briefly delved into light sails in Section 9.5.2.1, it is instructive to carry out a more detailed treatment here. In order to simplify our discussion, we will suppose that the beams are diffraction-limited and that we are operating in the far-field (Fraunhofer) regime, where the angular resolution is $\theta_A \approx 1.22\lambda/D_t$, in which D_t is the effective diameter of the beam lens.

Let us begin by calculating the acceleration that results from a laser beam with power output of W_t. We will suppose that the reflectivity of the light sail is denoted by $\mathcal{R}_s$, and it is necessary to ensure that $\mathcal{R}_s \approx 1$ across the wavelengths under consideration. If photons carrying momentum dp impinge on the light sail, we see that $\mathcal{R}_s\, dp$ are reflected. If this occurs over a time dt, the total change in momentum is $(1+\mathcal{R}_s)\, dp$. Next, we use the energy-momentum relationship $p = E/c$ for photons and the fact that the power is given by dE/dt. Putting these relations together, we obtain

$$a_s = \frac{(1+\mathcal{R}_s)\, W_t}{m_{\text{tot}} c}, \tag{10.104}$$

where a_s is the acceleration experienced by the light sail, and $m_{\text{tot}} = m_{pl} + \pi D_s^2 h_s \rho_s/4$ denotes the total mass of the light sail. In the latter formula, m_{pl} is the mass of the payload, whereas D_s, h_s, and ρ_s are the diameter, thickness, and density of the light sail, respectively. Note, however, that in

this formulation a small fraction of the incident energy will be absorbed by the light sail. If we model the light sail as a blackbody endowed with an emissivity η_s on *both* sides, we can invoke the Stefan-Boltzmann law to arrive at

$$(1-\mathcal{R}_s)\, W_t = \frac{\pi \sigma \eta_s D_s^2 T_s^4}{2}, \tag{10.105}$$

where T_s is the temperature of the light sail. By combining (10.104) and (10.105), it is easy to eliminate W_t and express the acceleration solely in terms of the sail parameters. Hence, depending on the materials utilized to construct the light sail, the maximum feasible acceleration can be determined accordingly.

There is, however, a subtle feature worth highlighting in connection with (10.104). This result is applicable only when *all* of the beamed energy is incident on the light sail. Beyond a certain distance ℓ_c, the beam divergence becomes such that only a fraction of the transmitted power is captured by the light sail, and the rest is electromagnetic leakage along the lines delineated in Section 9.5.2.1. We can estimate this distance by demanding that the angular resolution of the light sail at this distance should equal the beam diffraction angle θ_A:

$$\frac{D_s}{2\ell_c} \approx \frac{1.22\lambda}{D_t} \quad \Rightarrow \quad \ell_c \approx \frac{D_s D_t}{2.44\lambda} \tag{10.106}$$

Hence, at distances $\ell_s > \ell_c$, the acceleration experienced by the light sail (a_s) is given by

$$a_s \approx \frac{(1+\mathcal{R}_s)\, W_t}{m_{\rm tot} c} \left(\frac{\ell_c}{\ell_s}\right)^2. \tag{10.107}$$

If we hold the right-hand side of (10.104) constant and assume that the light sail starts from the origin and has an initial velocity of zero, its velocity is found by using $a_s = v_s dv_s/d\ell_s$. Therefore, we obtain

$$v_s^2 = \left(\frac{2\,(1+\mathcal{R}_s)\, W_t}{m_{\rm tot} c}\right) \ell_s \qquad\qquad \ell_s < \ell_c \tag{10.108a}$$

and

$$v_s^2 = v_c^2 + \left(\frac{2\,(1+\mathcal{R}_s)\, W_t \ell_c^2}{m_{\rm tot} c}\right)\left(\frac{1}{\ell_c} - \frac{1}{\ell_s}\right) \qquad \ell_s > \ell_c, \tag{10.108b}$$

where v_c is found by substituting $\ell_s = \ell_c$ in (10.108a) and is thus equal to

$$v_c = \sqrt{\frac{(1+\mathcal{R}_s)\, W_t D_s D_t}{1.22\left(m_{pl} + \pi D_s^2 h_s \rho_s/4\right)\lambda c}}. \tag{10.109}$$

The time t_c at which $\ell_s = \ell_c$ is easily determined via (10.108a) and is given by

$$t_c = \sqrt{\frac{cD_s D_t\left(m_{pl} + \pi D_s^2 h_s \rho_s/4\right)}{1.22\lambda\,(1+\mathcal{R}_s)\, W_t}}. \tag{10.110}$$

At this juncture, an interesting corollary is derivable. Let us suppose that $\ell_s \to \infty$ and let us denote the corresponding velocity by v_∞. By invoking (10.108a) and using the above relations, we end up with $v_\infty = \sqrt{2} v_c$. In other words, even by expending a huge amount of power in the regime $\ell_s > \ell_c$, we will only gain a modest increase in the sail velocity. Hence, from an energetic standpoint, it would be optimal to turn off the laser after a duration that is a few times higher than (10.110), because most of the extra gain of $\sqrt{2}$ would be achieved by then.

We turn our attention to a simple, but important, optimization problem. How do we maximize the terminal velocity v_∞? This question is important because it reduces the transit time to the destination. We need to calculate the maximum of v_c, given that $v_\infty = \sqrt{2} v_c$. By inspecting (10.109), we see that v_c has a monotonic dependence on all quantities except the sail diameter D_s. Remarkably, we find that the v_c is maximized when the sail mass equals the payload mass. Thus, the sail diameter in this scenario is computed via

$$D_s = \sqrt{\frac{4 m_{pl}}{\pi \rho_s h_s}}. \tag{10.111}$$

The maximum terminal velocity achievable ($v_{\rm opt}$) is given by

$$v_{\rm opt} = \sqrt{\frac{4\,(1+\mathcal{R}_s)\, W_t D_t}{1.22\pi D_s h_s \rho_s \lambda c}}. \tag{10.112}$$

Hence, if all the external parameters are held fixed, we see that $v_{\rm opt}$ is highest when the sail diameter and thickness are as small as engineering constraints

allow. Naturally, lowering the sail diameter is tantamount to having a smaller payload, as seen from (10.111). The corresponding value of (10.110), after making use of (10.111), is

$$t_{\rm opt} = \sqrt{\frac{\pi c D_s^3 D_t h_s \rho_s}{2.44\lambda\,(1+\mathcal{R}_s)\,W_t}}. \tag{10.113}$$

Finally, a comment on the energetics is in order. The kinetic energy of the spacecraft (KE_s) at the time t_c is

$$\mathrm{KE}_s = \frac{1}{2} m_{\rm tot} v_c^2 = \frac{(1+\mathcal{R}_s)\,W_t D_s D_t}{2.44 c\lambda}. \tag{10.114}$$

The photon energy incident on the light sail (E_γ) up to time t_c is estimated as

$$E_\gamma = W_t t_c = \sqrt{\frac{c D_s D_t \left(m_{pl} + \pi D_s^2 h_s \rho_s/4\right) W_t}{1.22\lambda\,(1+\mathcal{R}_s)}}. \tag{10.115}$$

Therefore, we may define an energy efficiency (δ_s) for the light sail as follows:

$$\delta_s = \frac{\mathrm{KE}_s}{E_\gamma} = \left(\frac{1+\mathcal{R}_s}{2}\right)\frac{v_c}{c} \tag{10.116}$$

As we have shown previously, the maximum value of v_c is attained when (10.111) is satisfied. Seeing as $\delta_s \propto v_c$, we find that the maximum efficiency is attained when the same criterion is valid—that is, when the payload and sail masses are equal to one another.

Hitherto, we have dealt with a laser system propelling the light sail. The photons that are reflected from the light sail were implicitly assumed to be lost. Instead, we can study a system wherein the photons are recycled between the spacecraft and the laser array. This method, known as photon recycling, is advantageous because it mitigates energy requirements. It was recently investigated by Kulkarni et al. (2018); we will mirror their analysis here. In the regime $\ell_s < \ell_c$, none of the power is lost due to leakage, because the spot sizes are sufficiently small, but some of it will be unavailable since a small fraction is absorbed. Let us denote the reflectivity of the beamer by $\mathcal{R}_b$. In this simple model, the reflected beam from the light sail is also taken to be diffraction-limited.

In essence, the photon recycling system can be treated as having an effective power ($W_{\rm eff}$) expressible as

$$W_{\rm eff} = W_t + W_t \mathcal{R}_s \mathcal{R}_b + W_t (\mathcal{R}_s \mathcal{R}_b)^2 + \cdots = \frac{W_t}{1 - \mathcal{R}_s \mathcal{R}_b}. \quad (10.117)$$

Thus, in the plausible limit where the reflectivities of the light sail and the beamer are close to unity, we see that the net gain resulting from photon recycling is high, as it has the form $(1 - \mathcal{R}_s \mathcal{R}_b)^{-1}$. Note that this treatment is highly idealized because we have modeled the reflectivities as being constant, whereas they actually depend on the wavelength; over each photon cycle, the frequencies are Doppler-shifted and the reflectivities will vary accordingly.

Next, consider the regime wherein $\ell_s > \ell_c$, implying that not all of the emitted power is incident upon the sail. In fact, only a fraction $(\ell_c/\ell_s)^2$ is incident upon it. Now, on account of symmetry, the same fraction will be intercepted by the beamer after reflection from the light sail; this result also follows naturally from an explicit calculation. Thus, the total loss due to leakage in one cycle is $(\ell_c/\ell_s)^4$. Hence, the net effective power of the system is given by

$$\begin{aligned} W_{\rm eff} &= W_t \left[1 + \mathcal{R}_s \mathcal{R}_b \left(\frac{\ell_c}{\ell_s} \right)^4 + (\mathcal{R}_s \mathcal{R}_b)^2 \left(\frac{\ell_c}{\ell_s} \right)^8 + \cdots \right] \\ &= W_t \left[1 - \mathcal{R}_s \mathcal{R}_b \left(\frac{\ell_c}{\ell_s} \right)^4 \right]^{-1}. \end{aligned} \quad (10.118)$$

In actuality, the above formula is an upper bound on $W_{\rm eff}$ for a given initial choice of ℓ_s, on account of ℓ_s being held fixed across all reflections. Even for reflectivities close to unity, the gain drops steeply with ℓ_s, implying that the photon recycling process is truly effective only when $\ell_s \lesssim \ell_c$ holds true.

Before bringing our nonrelativistic analysis to a close, let us gauge the relative efficiency of light sails and rockets. Our discussion will closely follow the approach outlined in McInnes (1999). The energy required for the rocket (E_R) to attain a velocity of Δv is estimated from the conservation of energy as

$$E_R = \int_0^t \frac{\dot{M}}{2} v_{\text{ex}}^2 \, dt = \frac{1}{2} M_f v_{\text{ex}}^2 \left[\exp\left(\frac{\Delta v}{v_{\text{ex}}} \right) - 1 \right] \tag{10.119}$$

after making use of (10.65). Now, let us suppose that we are operating in the regime wherein $\ell_s < \ell_c$ such that all of the laser power is being harnessed by a light sail endowed with the same *final* mass as the rocket; recall that the *final* mass is roughly synonymous with the *vehicle* mass. We have stated that the infinitesimal change in momentum imparted to the light sail is $(1 + \mathcal{R}_s)\ dp'$, where $dp' = M_f\ dv'$ in the nonrelativistic regime. Combining these expressions with the energy-momentum relationship for photons gives us

$$M_f\ dv' = \frac{1 + \mathcal{R}_s}{c}\ dE'. \tag{10.120}$$

By integrating this differential equation from $v' = 0$ to $v' = \Delta v$ and $E' = 0$ to $E' = E_L$, we find

$$E_L = \frac{M_f c \Delta v}{1 + \mathcal{R}_s}, \tag{10.121}$$

where E_L represents the energy expended by the laser. By taking the ratio of the rocket and laser energies, we have

$$\frac{E_R}{E_L} \approx \frac{v_{\text{ex}}^2}{c \Delta v} \left[\exp\left(\frac{\Delta v}{v_{\text{ex}}} \right) - 1 \right] \tag{10.122}$$

after invoking the relation $\mathcal{R}_s \approx 1$. Let us consider the regime $\Delta v / v_{\text{ex}} \ll 1$. We see that (10.122) simplifies to

$$\frac{E_R}{E_L} \approx \frac{v_{\text{ex}}}{c}. \tag{10.123}$$

In most cases, since $v_{\text{ex}}/c \ll 1$, the above equation suggests that rockets are more advantageous. However, an important point worth noting is that $\Delta v / v_{\text{ex}} \ll 1$ and $v_{\text{ex}}/c \ll 1$ collectively imply that the velocities attained are typically very low. Hence, even though rockets are relatively more advantageous in this regard, the transit time for interstellar travel will be very high. From the perspective of interstellar travel, it makes more sense to study the case where $\Delta v / v_{\text{ex}} > 1$, in which (10.122) is given by

$$\frac{E_R}{E_L} \sim \frac{v_{\text{ex}}^2}{c \Delta v} \exp\left(\frac{\Delta v}{v_{\text{ex}}} \right). \tag{10.124}$$

Owing to the exponential factor, we see that light sails are highly favored when fast speeds are desired. To illustrate this point, let us consider rockets with chemical propulsion that are characterized by $v_{\rm ex} \sim 10$ km/s. If we wish to attain $\Delta v \sim 0.1c$, we end up with the ridiculously high factor of $\sim 10^{1294}$. Instead, let us suppose that we have fission or fusion rockets with $v_{\rm ex} \sim 0.01c$. Even in this idealized setup, we obtain $E_R/E_L \sim 22$, which is much higher than unity.

10.5.2.2 Relativistic photon propulsion

In the relativistic analysis, we opt to work with a perfectly reflecting light sail for the sake of simplicity. Instead of adopting a full-fledged rigorous approach (see Parkin 2018; Füzfa et al. 2020), we will draw on a heuristic framework that nevertheless yields the correct result. Our calculations are undertaken in the rest frame of the laser situated on the home world.

Let us suppose that the rate of photons impinging on the moving light sail per unit time is denoted by $\dot{N}'$; the reason for introducing the prime symbol will become apparent shortly hereafter. The frequency of the incident photon is ν, whereas that of the reflected photon is ν_r. The process of incidence and reflection occurs over an infinitesimal time interval dt (measured in the laser frame). From the conservation of (relativistic) linear momentum, we have

$$m_{\rm tot}\, d\left(\gamma_s v_s\right) = \frac{\dot{N}' h\left(\nu + \nu_r\right)}{c} dt, \tag{10.125}$$

where the right-hand side follows from the energy-momentum relationship for photons. Note that $\gamma_s = \left(1 - v_s^2/c^2\right)^{-1/2}$ is the Lorentz factor of the moving spacecraft. By the same token, from energy conservation, we obtain

$$m_{\rm tot} c^2\, d\gamma_s = \dot{N}' h\left(\nu - \nu_r\right)\, dt. \tag{10.126}$$

Next, we use the identity $v\, d\left(\gamma v\right) = c^2\, d\gamma$ in (10.126). After dividing (10.125) and (10.126) and simplifying, we arrive at

$$\nu_r = \nu \left(\frac{1 - v_s/c}{1 + v_s/c}\right). \tag{10.127}$$

Now, let us suppose that the laser is emitting $\dot{N}$ photons of energy $h\nu$ per second. At first glimpse, it is tempting to directly equate $\dot{N}$ with $\dot{N}'$, but

the relative motion of the target and the source must be taken into account (Kulkarni et al. 2018). If dN is the infinitesimal number of photons emitted by the laser that are subsequently incident upon the light sail, this quantity must be conserved. Hence, by imposing this requirement, we have

$$dN = \dot{N}'\, dt = \dot{N}\, dt', \tag{10.128}$$

where dt' is the retarded time. Using $t' = t - x/c$ and $dx = v_s\, dt$, we see that $dt' = dt\,(1 - v_s/c)$. The introduction of the retarded time is apropos because the light sail situated at (x, t) experiences causal effects only from photons emitted by the laser before t'. From these relations, we find $\dot{N}' = \dot{N}\,(1 - v_s/c)$. Finally, with the laser power determined via $W_t = \dot{N}h\nu$ and drawing on (10.125), the equation of motion is given by

$$\frac{d\,(\gamma_s v_s)}{dt} = \frac{2W_t}{m_{\rm tot}c}\left(\frac{1 - v_s/c}{1 + v_s/c}\right). \tag{10.129}$$

After further rearrangement, we end up with

$$\frac{dv_s}{dt} = \frac{2W_t}{m_{\rm tot}c}\frac{1}{\gamma_s^3}\left(\frac{1 - v_s/c}{1 + v_s/c}\right), \tag{10.130}$$

which is identical to the governing dynamical equation prescribed in McInnes (1999) and Macchi et al. (2013). We caution, however, that the above expression is strictly valid only for rectilinear (i.e., one-dimensional) motion of perfectly reflecting light sails.

We will employ the notation $\beta_s = v_s/c$ henceforth for the sake of compactness. Next, we make use of $d\beta_s/dt = c\beta_s d\beta_s/d\ell_s$, thus allowing us to simplify (10.130) as follows:

$$d\ell_s = \frac{m_{\rm tot}c^3\gamma_s^3\beta_s}{2W_t}\left(\frac{1 + \beta_s}{1 - \beta_s}\right)\, d\beta_s. \tag{10.131}$$

The equation can be integrated to obtain ℓ_s as a function of v_s. However, in deriving this expression, we have supposed that there is no leakage of the radiation emitted by the laser. Naturally, this assumption is valid only until $\ell_s = \ell_c$. Now, let us integrate the above expression until ℓ_c and recall

that the latter is given by (10.106). Therefore, after some rearrangement, we find

$$W_t D_t = 1.22\lambda c^3 \left(\frac{m_{pl} + \pi D_s^2 h_s \rho_s / 4}{D_s} \right) \int_0^{\beta_c} \frac{\gamma_s \beta_s}{(1-\beta_s)^2}\, d\beta_s. \qquad (10.132)$$

This expression is useful because all variables involving the beamer are on the left-hand side, whereas the sail parameters are exclusively on the right-hand side. Now, suppose that we wish to either minimize the power required by the laser array or its effective size. It is found that the minimum value is achieved when the sail and payload masses are equal to one another. Hence, we see that the same criterion is valid regardless of whether the treatment is relativistic or nonrelativistic.

10.5.2.3 Light sails: Progress and challenges

We will briefly describe some of the major obstacles faced by light sail propulsion and the headway that has been made in this realm. To begin with, many of the issues that confront relativistic propulsion systems are also applicable to light sails. One drawback that we have already encountered is the role of the ISM. However, as noted in Section 9.3.1, collisions with dust grains, gas molecules, and cosmic rays do not appear to pose an insurmountable threat as they call for a modest amount of shielding. Two more potential issues stand out in this context: (1) the drag exerted by particles in the ISM that lead to the slowdown of the spacecraft and (2) the spacecraft acquiring a net charge due to collisions with ISM particles and UV photons.

Detailed analytical and numerical models indicate that (1) is not likely to pose a significant threat for spacecraft traveling at $\sim 0.1c$ across distances on the order of parsecs (Hoang 2017). Next, let us turn our attention to (2). If the spacecraft acquires a charge of Q_s, its gyroradius ($R_{\rm gyro}$) is given by

$$R_{\rm gyro} = \frac{m_{\rm tot} v_s}{Q_s B_{\rm ISM}}, \qquad (10.133)$$

where $B_{\rm ISM}$ is the average magnetic field in the ISM. The deflection impact parameter b_0 is estimated from $b_0/\mathrm{d}_\star \sim \mathrm{d}_\star / R_{\rm gyro}$, where $\mathrm{d}_\star$ is the distance from the target system. Therefore, we find that the deflection due to the interstellar magnetic fields is (Hoang & Loeb 2017)

$$b_0 \sim 1.7 \times 10^{-2}\,\mathrm{AU}\left(\frac{\mathrm{d}_\star}{1\,\mathrm{pc}}\right)^2\left(\frac{B_{\mathrm{ISM}}}{10^{-10}\,\mathrm{T}}\right)\left(\frac{\mathcal{Z}_s}{10^{10}}\right)$$
$$\times\left(\frac{m_{\mathrm{tot}}}{10^{-3}\,\mathrm{kg}}\right)^{-1}\left(\frac{v_s}{0.2c}\right)^{-1}, \tag{10.134}$$

where we have introduced the notation $Q_s \equiv \mathcal{Z}_s e$, and the normalization has been chosen for Breakthrough Starshot, a project that we will touch on a couple of paragraphs hereafter. A general estimation of $\mathcal{Z}_s$ is difficult, but it appears likely that the charging and subsequent deflection engendered are both significant for interstellar light sails. Hence, it will be necessary to develop suitable technologies for either getting rid of the excess charge or deploying onboard thrusters to make adjustments to the trajectory.

Thrusters also become important from a different perspective. As it stands, light sails are capable of acceleration but not deceleration. In case they wish to slow down at their destination, they may require onboard thrusters. Alternatively, if the transmitters are very large, the beamed laser power might be used for the purposes of slowing down by splitting the sail into two concentric rings and carrying out complex engineering maneuvers (Forward 1984). A third possibility is to utilize the combination of stellar gravitational and radiation pressure forces to decelerate. However, this method is not viable for large spacecraft because the net velocity decrement (v_{red}) is expressible as (Heller & Hippke 2017)

$$v_{\mathrm{red}} \sim \sqrt{\frac{L_\star D_s^2}{4cm_{\mathrm{tot}}s_{\mathrm{min}}}} - \sqrt{\frac{2GM_\star}{s_{\mathrm{min}}}}, \tag{10.135}$$

where s_{min} represents the minimum distance from the star. As the above formula illustrates, the maximal reduction of the velocity will decrease when the total mass of the light sail is increased.

Apart from these constraints, myriad social, engineering, and economic challenges confront humanity's endeavor to build relativistic light sails. For starters, it is absolutely essential to ensure that a light sail riding a laser beam is dynamically stable. A number of proposals have been put forth to address this aspect, ranging from the use of novel sail designs (Manchester & Loeb 2017; Srivastava & Swartzlander 2020) to the embedding and integration of self-stabilizing nanoscale structures and dielectric metamaterials in the sail surface (Ilic & Atwater 2019; Siegel et al. 2019; Myilswamy et al. 2020; Jin

et al. 2020). It is also necessary to ensure that the beam is accurately pointed toward the light sail at all times, thereby continually adjusting for the relative motion of the beamer and sail; this procedure is especially difficult when the payload and sail diameter are small.

If the light sail is meant to maintain communication with its origination point, issues such as data rates and transmission from the spacecraft become deeply important and demanding, but not impossible to surmount (Messerschmitt et al. 2020). There is also the matter of the laser array itself. If it is spread out across a large area, it will be necessary to ensure that the various elements are in perfect synchronicity when beaming radiation to the light sail and to avoid overheating. Lastly, a number of challenges revolve around material science. More precisely, the ideal material(s) should possess a combination of high reflectance, low mass density, and minimal absorptivity (to avoid excessive heating). While it is indisputable that there are major hurdles to overcome on this front, preliminary research has already provided encouraging results (Atwater et al. 2018).

In view of the manifold advantages stemming from light sails and their relative feasibility, it is natural to wonder whether any real-world technical studies have been pursued. Fortunately, this subject has witnessed substantive theoretical and empirical progress over the past couple of decades. The reader should refer to Macdonald and McInnes (2011) and Gong and Macdonald (2019) for reviews of the salient advances through the 2010s. The canonical, and arguably the first bona fide, example of a deep-space light sail propelled by solar radiation pressure is the Interplanetary Kite-craft Accelerated by Radiation Of the Sun (IKAROS), which was launched by the Japan Aerospace Exploration Agency in 2010 (Tsuda et al. 2011). This sail was characterized by a size and thickness of 14 m and 7.5 μm, respectively, and was fabricated from a polymer known as polyimide. It achieved a speed of $\sim$0.4 km/s in August 2013.[13] The LightSail 2 mission debuted in 2019 by The Planetary Society—of 5.6 m size and 4.6 μm thickness—successfully demonstrated that controlled light sail propulsion in low Earth orbit (modulated by sunlight) is realizable (Spencer et al. 2020).

When it comes to building prototypes, it may be argued that all light sails constructed thus far are rather limited in scope. In 2016, the Breakthrough Starshot initiative was inaugurated with a proposed funding of

13. IKAROS (2013, August 29), IKAROS Blog, Japan Aerospace Exploration Agency (JAXA), http://www.isas.jaxa.jp/home/IKAROS-blog/?itemid=1017

$100 million (US).[14] The aim of Breakthrough Starshot is to send a probe to the Alpha Centauri system at a speed of $0.2c$ so that it will reach the target in a couple of decades; in comparison, our fastest rockets have achieved peak speeds that are three orders of magnitude smaller. One of the unique characteristics of Breakthrough Starshot is that it has a very small size and payload; it has a sail diameter of ~ 4 m and its payload is on the order of a gram. The engineering design and rationale for such ultralight interstellar probes date back to at least Robert Forward's seminal Starwisp proposal (Forward 1985)[15] and have witnessed a revival over the past few years by a number of authors; see Lubin (2016a) for more details.

It goes without saying that an enterprise as ambitious as Breakthrough Starshot will face manifold physical and technical hazards, and perhaps socioeconomic and political ones as well. A detailed analysis by Parkin (2018) suggests that the total capital expenditure incurred would be $\sim$ $8 billion (US), but the expense for each subsequent launch is relatively lower at $\sim$ $6 million (US). While the cost of this project is indubitably high, it is comparable to other major scientific and engineering projects undertaken by humanity. Further, irrespective of whether it succeeds in beaming images of the Alpha Centauri system, it is predicted to yield significant benefits insofar as solar system exploration is concerned, to say nothing of revolutionizing current technologies in the best-case scenario. Realistic interstellar precursor mission designs for exploring the inner and outer Solar system over fairly rapid timescales by means of solar- and laser-propelled light sails have already been propounded (Heller et al. 2020; Turyshev et al. 2020).

10.5.3 Magnetic and electric sails

Although magnetic and electric sails are reliant on different physical principles, they share some commonalities, owing to which we have opted to classify them together. In the context of interstellar travel, the chief advantage offered by magnetic and electric sails is that they can decelerate the spacecraft without the necessity of carrying additional onboard fuel.

---

14. See https://breakthroughinitiatives.org/initiative/3

15. As opposed to utilizing lasers at optical or near-infrared wavelengths, Starwisp was reliant on microwaves for beamed propulsion.

10.5.3.1 Magnetic sails

The concepts underpinning the magnetic sail were developed by Dana Andrews and Robert Zubrin in 1988. We refer the reader to Andrews and Zubrin (1990) and Zubrin and Andrews (1991) for classic expositions of the basic physical principles of magnetic sails; a more up-to-date review can be found in Djojodihardjo (2018). The magnetic sail works by utilizing an onboard magnetic field to deflect plasma situated in the ISM or the interplanetary medium (IPM), thereby facilitating the acceleration or deceleration of the spacecraft depending on the design parameters. To elaborate, when the speed of the spacecraft is higher than that of the ambient plasma, the magnetic sail enables deceleration, and vice versa.

The magnetic sail relies on a loop of superconducting cable, conceivably tens of kilometers in diameter, to generate the magnetic field by initiating an electric current; in principle, the current would be maintained almost indefinitely in the superconductor. As fast charged particles undergo deflection, by momentum conservation, they impart momentum to the loop. In the presence of sustained plasma flow (e.g., stellar winds), the acceleration experienced by the relatively slower spacecraft will align with the direction of the flow. On the other hand, in order to effectuate deceleration, the spacecraft must be moving faster than the background plasma. In this case, the spacecraft's magnetic field deflects the charged particles in the ISM and loses momentum to the medium, thus giving rise to an effective drag that slows down the spacecraft. Magnetic sails have been more widely researched in the context of deceleration, as they do not require onboard fuel, unlike conventional rockets. In the context of providing acceleration, they are well suited for interplanetary operations but not for swift interstellar missions, as the resultant acceleration is merely ~ 0.01 m/s^2 at ~ 1 AU from Solar-type stars (Zubrin & Andrews 1991).

Therefore, we shall focus on deceleration in the ISM via magnetic sails as this application has been subjected to extensive investigation. Since the derivation of the drag force is slightly involved, we will instead work through a pared-down nonrelativistic version that captures the salient details. Let us suppose that the plasma density is denoted by ρ_p. For the sake of simplicity, we may presume that the spacecraft is moving at a higher speed (v_s) than the ambient plasma medium. In the spacecraft frame, the particles will apparently approach it with a speed of v_s; an implicit assumption employed here is that the plasma flow and spacecraft motion are aligned. Along the lines of

Section 4.2.1, we derive the magnetopause distance (R_{mp}) for the spacecraft as follows:

$$\rho_p v_s^2 \approx \frac{B_s^2}{2\mu_0}, \tag{10.136}$$

where B_s is the magnetic field of the spacecraft. If it can be modeled as a dipole, its radial dependence is given by

$$B_s = \frac{\mu_0 \mathcal{M}_s}{4\pi R_{mp}^3}, \tag{10.137}$$

where $\mathcal{M}_s$ constitutes the magnetic moment of the spacecraft. After solving for R_{mp} using the above two equations, we find

$$R_{mp} \sim \left(\frac{\mu_0 \mathcal{M}_s^2}{32\pi^2 \rho_p v_s^2}\right)^{1/6}. \tag{10.138}$$

Next, we observe that the classical drag force (F_D) experienced by a moving object in the fluid is

$$F_D = \frac{C_D}{2} \rho_p v_s^2 \langle A \rangle, \tag{10.139}$$

where C_D is the drag coefficient and $\langle A \rangle$ denotes the effective cross-sectional area of the object. In our simplified model, we specify $\langle A \rangle = \pi R_{mp}^2$ and $C_D \approx 1$, which transforms (10.139) into

$$F_D \approx \frac{(2\pi)^{1/3}}{8} \left(\sqrt{\mu_0} \mathcal{M}_s \rho_p v_s^2\right)^{2/3}. \tag{10.140}$$

Lastly, if the current in the superconducting loop of radius R_{sc} is I_{sc}, we have $\mathcal{M}_s = \pi I_{sc} R_{sc}^2$. Therefore, the final expression for the drag force is expressible as

$$F_D \approx 0.16\pi \left(\sqrt{\mu_0} I_{sc} R_{sc}^2 \rho_p v_s^2\right)^{2/3}. \tag{10.141}$$

In terms of its functional form, this equation is identical to the one obtained through a more accurate calculation, although the numerical coefficient must be raised by a factor of ~ 2 (Freeland 2015). Based on numerical simulations, Gros (2017) arrived at a different expression for the drag force, which is given by

$$F_D \approx 0.016\pi\rho_p v_s^2 R_{sc}^2 \left[\log\left(\frac{I_{sc}c}{v_s I_m}\right)\right]^3, \tag{10.142}$$

where $I_m = 1.55 \times 10^6$ A. Using the fact that $F_D = -m_{\rm tot} a_s$, it is possible to solve for the velocity v_s and displacement ℓ_s as a function of time by invoking either (10.141) or (10.142).

There are, of course, a number of challenges that magnetic sails must overcome. Most of them pertain to the deployment of the superconducting loop. Although the cosmic microwave background blackbody temperature of 2.7 K is lower than the critical temperature of high-temperature superconductors, it must be recognized that the ISM itself does *not* possess a homogeneous temperature of 2.7 K; it exceeds 100 K in many patches. It should also be noted that the temperature will rise as the spacecraft approaches the destination system. Yttrium barium copper oxide ($YBa_2Cu_3O_7$), one of the best-known high-temperature superconductors, has a critical temperature of ~ 90 K. Hence, if the target lies within the HZ of a star, the spacecraft would reach temperatures of ~ 200–300 K, thus calling for separate refrigeration systems, which increase the mass of the vehicle.

The magnetic sail, as we have seen, relies on naturally occurring plasma in the ISM and IPM. However, by analogy with laser-pushed light sails, one may envision magnetic sail propulsion by ion beams. This proposal was put forth by Robert Winglee in 2004,[16] but this technology is probably not well suited for interstellar missions, owing to the need for directing a collimated ion beam across large distances.

10.5.3.2 Electric sails

Instead of using magnetic fields to extract momentum from the ambient plasma, it is possible to invoke electric fields. A formal description of this propulsion system, dubbed the electric sail, was provided for the first time by Pekka Janhunen in 2004. The electric sail is suitable for interplanetary travel, as the acceleration experienced at ~ 1 AU from the Sun was estimated to be ~ 0.01 m/s^2 (Janhunen 2004). When launched in the vicinity of most stars, electric sails have the capacity to reach final velocities on the order

16. Magnetized beamed plasma propulsion (MagBeam), University of Washington Earth and Space Sciences, https://earthweb.ess.washington.edu/space/magbeam/

of 100 km/s (Lingam & Loeb 2020c). Instead, if high-energy astrophysical objects are deployed, electric sails are capable of attaining genuinely relativistic terminal speeds (Lingam & Loeb 2020f). Furthermore, just as with magnetic sails, they may play a key role in decelerating spacecraft moving at high speeds in the ISM.

To obtain an estimate for the force acting on the electric sail, we will mirror the approach delineated in Janhunen and Sandroos (2007); a more accurate and rigorous analysis can be found in Janhunen et al. (2010). Let us suppose that there exists a long charged wire and the plasma flow is perpendicular to the wire. By applying Gauss's law, the electric field (in vacuum) of a wire $E(x)$ measured at a distance x is given by

$$E(x) \cdot (2\pi x L_w) = \frac{\Lambda_w L_w}{\epsilon_0} \quad \Rightarrow \quad E(x) = \frac{\Lambda_w}{2\pi \epsilon_0 x}, \tag{10.143}$$

where L_w and Λ_w are the length and charge density (charge per unit length) of the wire; strictly speaking, the preceding calculation is accurate in the limit $L_w \gg x$ as the derivation presupposes an infinitely long wire. The electric potential $V(x)$ is obtained via $E(x) = -dV/dx$, thereby yielding

$$V(x) = \frac{\Lambda_w}{2\pi \epsilon_0} \ln\left(\frac{x_0}{x}\right), \tag{10.144}$$

where x_0 is the coordinate at which the potential vanishes. In practice, x_0 is chosen to be twice the electron Debye length of the plasma; in theory, it remains indeterminate. Instead of working with Λ_w, we may utilize the potential (V_0) at the surface of the wire. If the radius of the wire is denoted by x_w, we find

$$V(x) = V_0 \frac{\ln(x_0/x)}{\ln(x_0/x_w)}. \tag{10.145}$$

However, observe that this result is the vacuum solution. In actuality, the plasma electrons act as a shield, thus altering the potential relative to the vacuum solution. On the basis of numerical simulations, Janhunen and Sandroos (2007) found that the effective potential is well described by

$$V(x) = \frac{V_0}{2} \frac{\ln\left[1 + (x_0/x)^2\right]}{\ln(x_0/x_w)}. \tag{10.146}$$

In the limit where $x/x_0 \ll 1$, we see that (10.146) reduces to (10.145).

The force per unit length (dF/dz) experienced by the wire is roughly the product of the plasma dynamical pressure and the effective width (x_s) associated with the electrostatic potential:

$$\frac{dF}{dz} \approx 3.09 \rho_p v_s^2 x_s, \tag{10.147}$$

where we have supposed that the velocity of the spacecraft is much higher than the ambient plasma; in the opposite regime characterized by fast plasma flows, we must replace v_s with the plasma velocity (v_p). The factor of 3.09 has been introduced to ensure an accurate fit with numerical simulations. The effective width is determined by demanding that it equals the proton stopping distance, i.e., we have

$$eV(x_s) = \frac{1}{2} m_p v_s^2, \tag{10.148}$$

where the potential is given by (10.146). After solving for x_s, substituting it into (10.147), and integrating the resulting equation (for constant v_s), we find that the force experienced by a single wire (F_w) is

$$F_w \approx \frac{3.09 \rho_p v_s^2 x_0 L_w}{\sqrt{\exp\left[m_p v_s^2 \left(eV_0\right)^{-1} \ln\left(x_0/x_w\right)\right] - 1}}. \tag{10.149}$$

The total force experienced by the spacecraft is the product of F_w and the number of wires in the mesh that comprises the electric sail. We will not examine engineering challenges relating to the design of the mesh and its capacity to sustain an electric field across interstellar distances without being subject to major damage.

Although we have derived the salient aspects of electric and magnetic sails in isolation, it is feasible for a spacecraft to incorporate both features. In the case of large spacecraft with total masses of $\sim 10^4$ kg, using electric and magnetic sails in tandem could reduce the deceleration time by ~ 20 percent (Perakis & Hein 2016).

10.6 CONCLUSION

> We need the stars, Bankole. We need purpose! . . . If we're to be anything other than smooth dinosaurs who evolve, specialize and die, we need the stars. That's why the Destiny of Earthseed is to take root among the stars.
>
> —Octavia Butler, *Parable of the Talents*

What determines the distribution and future of life in the Universe? If panspermia is feasible across interstellar distances, it enhances the likelihood of a Universe inhabited by multifarious lifeforms to some degree, and vice versa. And even if panspermia was effectively confined to the boundaries of a single planetary system, it would nevertheless have important implications insofar as our ongoing quest for life on the likes of Mars, Venus, Europa, Titan, and Enceladus is concerned.

For much of the twentieth century, panspermia was sparsely studied, and therefore its probability of occurrence was poorly constrained. In the past two decades, we have witnessed numerous experiments conducted in space and harsh environmental conditions on Earth, thus establishing the wondrous resilience of microbes. Over the same period, we have beheld the advent of numerical simulations, thereby enabling us to predict the fraction of ejecta that could depart from the planets and moons of our Solar system and safely land on another. Looking beyond, state-of-the-art numerical and analytical models seemingly indicate that certain local environments are conducive to the transfer of life across interstellar scales. Yet, it must be cautioned that none of these publications provide concrete evidence for panspermia, although they may perhaps boost its likelihood of actualization.

In order to truly test panspermia on interstellar scales, we will need to await results from exoplanet surveys in the next few decades that might permit us to confirm or disprove the panspermia hypothesis, albeit only in the fortuitous event that we detect $\mathcal{O}(10)$ worlds with confirmed biosignatures. Closer to home, we will need to identify objects of interstellar origin and initiate missions to investigate them as well as conventional targets of astrobiological interest (Eubanks et al. 2020; Turyshev et al. 2020). If we are lucky enough to find actual samples of fossils or extant life, this will provide us with another channel to test panspermia. However, in order to pursue the second line of inquiry (in situ studies) within our Solar system, the necessity of nurturing a strong space exploration program becomes apparent.

The stars have been a perennial source of fascination for humans for thousands of years. Yet, it was only in 1957, with the momentous launch of Sputnik, that we took our first genuine steps into outer space. The next major landmark was unquestionably the celebrated landing of humans on the Moon in 1969. After more than five decades later, it would appear as though space exploration has stalled to some degree, as the Moon still represents the farthest that humans have ventured from Earth. The funding for spaceflight programs has diminished ever since its heyday in the 1960s, although the desire to engage in this ambitious enterprise is witnessing a rekindling of sorts. The stirring excerpt from Octavia Butler is a timely reminder that eclectic positives (both pragmatic and philosophical) arguably ensue from pursuing this goal (Dyson 1979a). One of them, as we have previously remarked, is the prospect of discovering life on other worlds in our Solar system.

Apart from complex and pressing socioeconomic and political issues, pivotal technological factors need to be taken into consideration. For much of our history, we have relied on chemical rockets. Yet, as we came to appreciate in this chapter, they require a huge amount of fuel even for achieving modest velocities. Fortunately, a number of intriguing alternatives have sprung up, many of which are being currently implemented and tested today. For example, if relativistic light sails are realized one day, they will usher in a new age of space exploration, whereby we can launch fleets of small spacecraft to all the potentially inhabited worlds in our Solar system with a travel time of weeks (at most) and investigate them in detail. The same logic could be extended to survey exoplanets in nearby stellar systems, although the total time lag in retrieving data would run into decades.

Lastly, if we look to the distant future, scores of existential threats will confront putative technological intelligence on Earth; it is conceivable that some of these perils may also beset ETIs, should they exist. In case humans successfully address the global catastrophic risks posed by anthropogenic climate change, biotechnology, and nuclear warfare, *inter alia,* it is likely that the Earth would experience a natural greenhouse effect $\sim$1-2 Gyr in the future because of the gradual increase in stellar luminosity. Even supposing that this obstacle is overcome through herculean feats of geoengineering, in $\sim$5 Gyr the Sun will morph into a red giant and render the Earth uninhabitable. Shortly thereafter, the Sun will pass through the planetary nebula phase and become a white dwarf. Thus, in the very long term, should any

technological entities be "alive" on Earth, they will probably need to relocate elsewhere. When viewed in this light, interstellar travel is rendered not only desirable but also necessary for long-lived technological intelligences.

Circling back to our original theme, we will bring our discussion to a close on a philosophical note. As we have argued herein, panspermia and interstellar travel are potential avenues for the Universe to be populated with biota. Of these multitudinous offshoots, perhaps some of their descendants could evolve consciousness and intelligence and seed other worlds in turn. To put it differently, the two processes might enhance the prevalence of biological species as well as technological entities over time. By doing so, these mechanisms may act in concordance to increase the prospects "for the cosmos to know itself," to borrow from the TV series *Cosmos,* narrated by Carl Sagan (Sagan et al. 1980):

> Some part of our being knows this is where we came from. We long to return. And we can. Because the cosmos is also within us. We're made of star-stuff. We are a way for the cosmos to know itself.

Epilogue

SIC ITUR AD ASTRA

> But it seems reasonable to believe—and I do believe—that the more clearly we can focus our attention on the wonders and realities of the Universe about us the less taste we shall have for the destruction of our race. . . . In one way or another all of us have been touched by an awareness of the world of nature. . . . If we have ever regarded our interest in natural history as an escape from the realities of our modern world, let us reverse this attitude. For the mysteries of living things, and the birth and death of continents and seas, are among the great realities.
>
> —Rachel Carson, "Design for Nature Writing"

The ceaseless motion of the stars and planets across the night sky has proven to be a perennial source of fascination for human beings over aeons. As our ancestors gazed upon the night skies, they were, in all likelihood, transfixed by the multitude of celestial objects populating them and may have found themselves musing about whether these worlds harbored life, both like and unlike that on Earth. Modern archaeology has revealed that the fascination with astronomy was not restricted to the civilizations and peoples of ancient Egypt, Mesopotamia, India, China, Mesoamerica, Elam, and Crete, to name a few, but might have extended even further beyond in the mists of time. It has been hypothesized, albeit not removed from controversy, that striking examples included the erstwhile inhabitants of Göbekli Tepe and Çatalhöyük in Turkey as well as those dwelling in the cave complexes of Altamira (Spain), Lascaux and Chauvet (France), and other places. It is therefore no wonder that astronomy, in one form or another, has occupied a place of prominence and reverence in human societies over millennia.

This decade carries with it the hopes, dreams, and promises of novel scientific developments in astrobiology via the medium of space exploration.

In our Solar system, the *Mars 2020 Perseverance Rover*[1] and the ExoMars mission comprising the *Rosalind Franklin* rover[2] have been outfitted to seek biological signatures of life on the Martian (sub)surface. Jupiter's moon, Europa, has a subsurface ocean of liquid water, whose volume is possibly higher than that of Earth's oceans. The primary purpose of the *Europa Clipper* mission,[3] with the Jupiter Icy Moons Explorer (JUICE) also meriting mention,[4] is to carry out flybys of Europa and search for biosignatures of putative lifeforms.[5] Aside from these flagship robotic projects, a number of national and private (e.g., SpaceX) agencies have expressed interest in landing humans on Mars, although it is vital to recognize there are substantive ethical, technical, and logistical challenges that need to be overcome first.

Looking beyond the confines of our Solar system, the discovery and characterization of exoplanets is already proceeding apace. The launch of the oft-delayed James Webb Space Telescope (JWST) will enable us to study a few temperate exoplanets around M-dwarfs and search for atmospheric biosignatures such as molecular oxygen.[6] In parallel, Extremely Large Telescopes (ELTs), with diameters of ~ 30 m, are under construction. These ground-based telescopes will, under optimal circumstances, facilitate the hunt for biosignatures on exoplanets orbiting low-mass stars. The long-term benefits of studying potentially habitable exoplanets arise from the fact they outnumber astrobiological targets in our Solar system by orders of magnitude and may therefore patently yield humanity a statistically significant number of samples in the quest for extraterrestrial life.

The organisms we alluded to heretofore are implicitly nontechnological, and thus most probably microbial, in nature. However, for reasons delineated throughout this book, there are numerous advantages to be gained from undertaking searches for signatures of extraterrestrial technology—that is, technosignatures. Historically, the pursuit of technosignatures witnessed a boom in the 1960s but subsequently underwent a decline and attained

---

1. See https://mars.nasa.gov/mars2020/
2. See http://exploration.esa.int/mars/
3. See https://www.jpl.nasa.gov/missions/europa-clipper/
4. See http://sci.esa.int/juice/
5. On a related note, the *Dragonfly* spacecraft, which is predicted to reach Titan in 2034, will carry out detailed in situ exploration of Titan's surface to gauge its habitability.
6. See https://www.jwst.nasa.gov/

its nadir in the 1990s. There are, however, encouraging signs that the status quo is starting to change for the better. A combination of budding private enterprises, like the Breakthrough Listen initiative,[7] and renewed interest shown by NASA bolsters the expectation that technosignatures constitute yet another compelling and viable avenue for finding signatures of life beyond our planet.

While seeking life in its "endless forms most beautiful and most wonderful," to borrow Darwin's poetic phrase from the closing lines of *On the Origin of Species* (1859, p. 490), our horizons of discovery are inevitably curbed by the bounds of our knowledge and imagination. We will undoubtedly enhance the efficacy of our search strategies by striving to responsibly create synthetic life in laboratory experiments or computer simulations under geochemical conditions that are akin, as well as alien, to those found on Earth. Likewise, new pathways open up in discerning technosignatures generated by human-level species through the visualization and actualization of novel technologies. In other words, scientific and technical advances in other realms could not only exert a direct impact on surveys for alien life via the design and construction of innovative instruments and data mining techniques but also inform us about how and where we should conduct searches for biological and technological signatures.

Hitherto, we have deliberately chosen to paint an optimistic picture of monotonic human progress, a weltanschauung that is naive and rosy hued at first glimpse. Needless to say, the dynamics of the human condition in totality will be unfathomably more complex and chaotic in nature. By the same token, any candidate biosignatures or technosignatures that we may detect will most likely evince ambiguity, thus warranting careful analysis in order to determine whether these signals are truly indicative of extraterrestrial life. Far more worrisome is the fact that we are currently living in a turbulent epoch beset by swiftly accelerating anthropogenic climate change and social, political, and economic schisms—perhaps, quite ironically, in the very same era when the first signatures of life beyond Earth might be perceptible by our instruments. The manifold challenges confronting humans in this day and age entail deepening cultural and socioeconomic inequities, global warming, ocean acidification, novel pandemics, massive biodiversity losses, artificial intelligence, potential biological and chemical weapons,

---

7. See https://breakthroughinitiatives.org/initiative/1

solar superflares, asteroid impacts, and countless others. Hence, it is perfectly natural to wonder whether, in this unsettling and crisis-ridden "Age of Extremes"—à la Eric Hobsbawm's magnum opus (1995)—the starry-eyed quest for life in the Cosmos offers us any (in)tangible benefits. Taking our cue from the inspiring words of Rachel Carson at the inception of this epilogue, we are moved to posit that the answer is provisionally in the affirmative.

To begin with, it is worth contemplating the impact of extraterrestrial life on subjects that have unmistakable practical applications for humanity. The identification and extraction of microbial life samples in our Solar system has the potential to revolutionize medicine and biotechnology. In the best-case scenario, where we establish contact with extraterrestrial technological intelligences (ETIs) much more advanced than our own, we might be able to acquire technological, social, and cultural knowledge that aids in the improvement of human society, as averred by several sanguine advocates of the Search for Extraterrestrial Technological Intelligence (SETI). If we "merely" stumble across relics of extinct ETIs, we may still have the capability to reverse-engineer those artifacts and garner valuable information from them. At the same time, one must acknowledge that there are possible downsides to encountering either microbial or technological life out there, with some of the most extreme examples arguably being contagion and invasion, respectively. Even in such worst-case scenarios, however, what cannot be denied is that finding life on other worlds would have a major tangible impact on human society.

In certain quarters, although it remains rarely admitted in the open, there appears to be a genuine anxiety that the discovery of extraterrestrial life, especially of the technological kind, could instigate an existential crisis by dethroning humankind from their perceived special position in the Universe. However, we opine that the putative detection of alien life ought not be construed as evidence either for or against the perception that we are exceptional—to wit, the smartest kid on the block. It seems more constructive, instead, to accept the humbling yet uplifting notion that each and every species is inherently special. Indeed, one may even contend that the human capacity to recognize and come to terms with our ostensibly nonprivileged position in the cosmic order renders us unique, *inter alia,* as eloquently expressed by the Italian poet and philosopher Giacomo Leopardi in his monumental tome *Zibaldone* (2013, p. 1302):

> When, in considering the multiplicity of worlds, he feels himself to be an infinitesimal part of a globe which itself is a negligible part of one of the infinite number of systems that go to make up the world . . . with this single act of thought he gives the greatest possible proof of the nobility and immense capability of his own mind, which, enclosed in such a small and negligible being, has nonetheless managed to know and understand things so superior to his own nature, and to embrace and contain this same intensity of existence and things in his thought.

We commenced our discussion with examining the relevance of the search for extraterrestrial life in these embattled and tumultuous times. It is thus natural to circle back to this theme even as this book draws to an end. Let us suppose, for instance, that we find evidence of life beyond Earth, irrespective of whether it is "alive" or "dead" at the moment of discovery. The epochal knowledge that we are not alone could be ideally harnessed, in principle (although not necessarily in reality), to bring human beings together under a common umbrella. In particular, if we were to detect signatures of ETIs that went extinct through self-destruction (e.g., nuclear warfare), it might prompt us to gain a deeper appreciation of the ephemeral nature of life and intelligence, consequently stimulating us to take better care of our planet and its billions and billions of biota.

In closing, however, we must confront the possibility that we will discern no evidence after decades more of searching.[8] Does this scenario imply that the search for life elsewhere in the Cosmos was a fruitless endeavor, a quixotic quest, a monumental folly that culminated in the waste of billions of taxpayer dollars? On the basis of what we have chronicled, the answer is surely an unequivocal *no*. At the minimum, it permits us to set rough constraints on the frequency of life in the Universe. In addition, the technologies and models developed toward this end could be transferable to other domains, thereby enriching the corpus of human knowledge. Last, but not least, if we fail to detect extraterrestrial life after a prolonged interval of time, it might suggest that Earthlike complex biospheres are rare in the Universe. This purported violation of the Copernican Principle should compel us to serve as mindful stewards of our planet and safeguard the wondrous

8. It goes without saying that the absence of evidence is not automatically equivalent to evidence of absence.

biological diversity we often take for granted, in order to ensure that subsequent generations are accorded the opportunity to experience and treasure Earth's myriad biomes firsthand.

Hence, broadly construed, we can identify a number of ineffable and profound connections that entwine astrobiology—specifically, the search for life in the boundless reaches of space—on the one hand with the current and future fates of human beings and societies on the other. If we perceive astrobiology through the prism of exploring our origins, evolution, and future, this voyage of perseverance and discovery—one that will be indubitably laden with innumerable cul-de-sacs, fruitless byways, and erroneous conclusions—might, for the first time, enable us to know ourselves and our place in the Cosmos at the end of all our wanderings. In the same vein, it behooves us to recall the perspicacious musings of Carl Sagan in his book *Broca's Brain: Reflections on the Romance of Science* (1980, pp. 314-315):

> Through all of our history we have pondered the stars and mused whether mankind is unique or if, somewhere else in the dark of the night sky, there are other beings who contemplate and wonder as we do, fellow thinkers in the cosmos. . . . Somewhere else there might be very exotic biologies and technologies and societies. In a cosmic setting vast and old beyond ordinary human understanding, we are a little lonely; and we ponder the ultimate significance, if any, of our tiny but exquisite blue planet. . . . In the deepest sense, the search for extraterrestrial intelligence is a search for ourselves.

Thus, the quest for life's beginnings, its current distribution in the Cosmos, and its enigmatic future is imbued with the promise of uniting humanity, revolutionizing numerous spheres of human knowledge, and navigating the uncharted and hazardous waters of the Anthropocene. Furthermore, we find ourselves poised at a unique moment in space and time to carry out this endeavor. Let us, therefore, muster our courage, initiate the search in earnest and strive to commune with the stars in the spirit of Virgil's famous phrase from the *Aeneid,* "*Macte nova virtute, puer; sic itur ad astra*" (1875, p. 210), which may be loosely translated into English as "Blessings on your fresh courage, child; thus one journeys to the stars."

REFERENCES

Aarnio, A. N., Matt, S. P., and Stassun, K. G. (2012). Mass loss in pre-main-sequence stars via coronal mass ejections and implications for angular momentum loss. *Astrophys. J.,* 760(1):9.

Abbot, D. S. and Switzer, E. R. (2011). The Steppenwolf: A Proposal for a Habitable Planet in Interstellar Space. *Astrophys. J. Lett.,* 735(2):L27.

Abe, K. and Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nat. Neurosci.,* 14(8):1067–1074.

Abe, Y., Abe-Ouchi, A., Sleep, N. H., and Zahnle, K. J. (2011). Habitable Zone Limits for Dry Planets. *Astrobiology,* 11(5):443–460.

Abeysekara, A. U., Archambault, S., Archer, A., Benbow, W., Bird, R., Buchovecky, M., Buckley, J. H., Byrum, K., Cardenzana, J. V., Cerruti, M., Chen, X., Christiansen, J. L., Ciupik, L., Cui, W., Dickinson, H. J., Eisch, J. D., Errando, M., Falcone, A., Fegan, . . . Zitzer, B. (2016). A Search for Brief Optical Flashes Associated with the SETI Target KIC 8462852. *Astrophys. J. Lett.,* 818(2):L33.

Abramov, O., Kring, D. A., and Mojzsis, S. J. (2013). The impact environment of the Hadean Earth. *Chem. Erde. Geochem.,* 73(3):227–248.

Abramowicz, M., Bejger, M., Gourgoulhon, E., and Straub, O. (2019). The Messenger: a galactic centre gravitational-wave beacon. *Sci. Rep.,* 10:7054.

Abrevaya, X. C., Leitzinger, M., Oppezzo, O. J., Odert, P., Patel, M. R., Luna, G. J. M., Forte-Giacobone, A. F., and Hanslmeier, A. (2020). The UV surface habitability of Proxima b: first experiments revealing probable life survival to stellar flares. *Mon. Not. R. Astron. Soc. Lett.,* 494(1):L69–L74.

Ackeret, J. (1946). Zur Theorie der Raketen. *Helv. Phys. Acta,* 19(2):103–112.

Adam, Z. A., Hongo, Y., Cleaves, H. J., Yi, R., Fahrenbach, A. C., Yoda, I., and Aono, M. (2018). Estimating the capacity for production of formamide by radioactive minerals on the prebiotic Earth. *Sci. Rep.,* 8:265.

Adamala, K. and Szostak, J. W. (2013). Nonenzymatic template-directed RNA synthesis inside model protocells. *Science,* 342(6162):1098–1100.

Adamala, K. P., Engelhart, A. E., and Szostak, J. W. (2016). Collaboration between primitive cell membranes and soluble catalysts. *Nat. Commun.,* 7:11041.

Adami, C. and LaBar, T. (2017). From entropy to information: Biased typewriters and the origin of life. In Walker, S. I., Davies, P. C. W., and Ellis, G. F. R., editors, *From Matter to Life: Information and Causality,* pages 130–154. Cambridge University Press, Cambridge, UK.

Adams, F. C. and Spergel, D. N. (2005). Lithopanspermia in Star-Forming Clusters. *Astrobiology,* 5(4):497–514.

Adams, F. C. (2010). The Birth Environment of the Solar System. *Annu. Rev. Astron. Astrophys.,* 48:47–85.

Adamski, P., Eleveld, M., Sood, A., Kun, Á., Szilágyi, A., Czárán, T., Szathmáry, E., and Otto, S. (2020). From self-replication to replicator systems en route to de novo life. *Nat. Rev. Chem.,* 4:386–403.

Adcock, C. T., Hausrath, E. M., and Forster, P. M. (2013). Readily available phosphate from minerals in early aqueous environments on Mars. *Nat. Geosci.,* 6(10):824–827.

Adriaense, J. E. C., Koski, S. E., Huber, L., and Lamm, C. (2020). Challenges in the comparative study of empathy and related phenomena in animals. *Neurosci. Biobehav. Rev.,* 112:62–82.

Agol, E., Steffen, J., Sari, R., and Clarkson, W. (2005). On detecting terrestrial planets with timing of giant planet transits. *Mon. Not. R. Astron. Soc.,* 359(2):567–579.

Agol, E. and Fabrycky, D. C. (2018). Transit-Timing and Duration Variations for the Discovery and Characterization of Exoplanets. In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets,* pages 797–816. Springer, Cham, Switzerland.

Airapetian, V. S., Glocer, A., Gronoff, G., Hébrard, E., and Danchi, W. (2016). Prebiotic chemistry and atmospheric warming of early Earth by an active young Sun. *Nat. Geosci.,* 9(6):452–455.

Airapetian, V. S., Jackman, C. H., Mlynczak, M., Danchi, W., and Hunt, L. (2017). Atmospheric Beacons of Life from Exoplanets Around G and K Stars. *Sci. Rep.,* 7:14141.

Airapetian, V. S., Barnes, R., Cohen, O., Collinson, G. A., Danchi, W. C., Dong, C. F., Del Genio, A. D., France, K., Garcia-Sage, K., Glocer, A., Gopalswamy,

N., Grenfell, J. L., Gronoff, G., Güdel, M., Herbst, K., Henning, W. G., Jackman, C. H., Jin, M., Johnstone, C. P., . . . Yamashiki, Y. (2020). Impact of space weather on climate and habitability of terrestrial-type exoplanets. *Int. J. Astrobiol.,* 19(2):136–194.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.,* 74:47–97.

Alcott, L. J., Mills, B. J. W., and Poulton, S. W. (2019). Stepwise Earth oxygenation is an inherent property of global biogeochemical cycling. *Science,* 366(6471):1333–1337.

Alexander, R. M. (1996). *Optima for Animals*. Princeton University Press, Princeton, NJ.

Allwood, A. C., Rosing, M. T., Flannery, D. T., Hurowitz, J. A., and Heirwegh, C. M. (2018). Reassessing evidence of life in 3,700-million-year-old rocks of Greenland. *Nature,* 563(7730):241–244.

Alvarado-Gómez, J. D., Drake, J. J., Cohen, O., Moschou, S. P., and Garraffo, C. (2018). Suppression of Coronal Mass Ejections in Active Stars by an Overlying Large-scale Magnetic Field: A Numerical Study. *Astrophys. J.,* 862(2):93.

Ambrose, S. H. (2001). Paleolithic technology and human evolution. *Science,* 291(5509):1748–1753.

Amend, J. P., LaRowe, D. E., McCollom, T. M., and Shock, E. L. (2013). The energetics of organic synthesis inside and outside the cell. *Phil. Trans. R. Soc. B.,* 368(1622):20120255.

Amend, J. P. and LaRowe, D. E. (2019). Minireview: Demystifying microbial reaction energetics. *Environ. Microbiol.,* 21(10):3539–3547.

Amodio, P., Boeckle, M., Schnell, A. K., Ostojić, L., Fiorito, G., and Clayton, N. S. (2019). Grow Smart and Die Young: Why Did Cephalopods Evolve Intelligence? *Trends Ecol. Evol.,* 34(1):45–56.

Anbar, A. D. (2008). Elements and Evolution. *Science,* 322(5907):1481–1483.

Andrews, D. G. and Zubrin, R. M. (1990). Magnetic Sails and Interstellar Travel. *J. Br. Interplanet. Soc.,* 43:265–272.

Andrews, K. (2020). *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. Routledge, New York, NY.

Angilletta, M. J. (2009). *Thermal Adaptation: A Theoretical and Empirical Synthesis*. Oxford University Press, Oxford, UK.

Anglada-Escudé, G., Amado, P. J., Barnes, J., Berdiñas, Z. M., Butler, R. P., Coleman, G. A. L., de La Cueva, I., Dreizler, S., Endl, M., Giesers, B., Jeffers, S. V., Jenkins, J. S., Jones, H. R. A., Kiraga, M., Kürster, M., López-González, M. J., Marvin, C. J., Morales, N., Morin, J., . . . Zechmeister, M. (2016). A terrestrial planet candidate in a temperate orbit around Proxima Centauri. *Nature,* 536(7617):437–440.

Annis, J. (1999a). An astrophysical explanation for the "great silence". *J. Br. Interplanet. Soc.,* 52(1):19–22.

Annis, J. (1999b). Placing a limit on star-fed Kardashev type III civilisations. *J. Br. Interplanet. Soc.,* 52(1):33–36.

Ansbro, E. (2001). New OSETI observatory to search for interstellar probes. *Proc. SPIE,* 4273:246–253.

Antón, S. C., Potts, R., and Aiello, L. C. (2014). Evolution of early *Homo*: An integrated biological perspective. *Science,* 345(6192):1236828.

Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., and Sheldon, B. C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature,* 518(7540):538–541.

Aplin, L. M. (2019). Culture and cultural evolution in birds: A review of the evidence. *Anim. Behav.,* 147:179–187.

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror System Hypothesis*. Oxford University Press, Oxford, UK.

Archibald, J. (2014). *One Plus One Equals One: Symbiosis and the Evolution of Complex Life*. Oxford University Press, Oxford, UK.

Archibald, J. M. (2015). Endosymbiosis and eukaryotic cell evolution. *Curr. Biol.,* 25(19):R911–R921.

Argiroffi, C., Reale, F., Drake, J. J., Ciaravella, A., Testa, P., Bonito, R., Miceli, M., Orlando, S., and Peres, G. (2019). A stellar flare-coronal mass ejection event revealed by X-ray plasma motions. *Nat. Astron.,* 3:742–748.

Aristotle. (1907). *De Anima.* Translated by R. D. Hicks. Cambridge University Press, Cambridge, UK.

Arkhipov, A. V. (1996). On the possibility of extraterrestrial-artefact finds on the Earth. *Observatory,* 116:175–176.

Arkhipov, A. V. (1998). Earth-Moon System as a Collector of Alien Artifacts. *J. Br. Interplanet. Soc.,* 51(5):181–184.

Armitage, J. (1977). The prospect of astro-palaeontology. *J. Br. Interplanet. Soc.,* 30:466–469.

Armitage, P. J. (2020). *Astrophysics of Planet Formation*. Cambridge University Press, Cambridge, UK (2nd edition).

Armstrong, J. C., Wells, L. E., and Gonzalez, G. (2002). Rummaging through Earth's Attic for Remains of Ancient Life. *Icarus,* 160(1):183–196.

Armstrong, S. and Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronaut.,* 89:1–13.

Arndt, N. and Nisbet, E. G. (2012). Processes on the young earth and the habitats of early life. *Annu. Rev. Earth Planet. Sci.,* 40:521–549.

Arney, G., Domagal-Goldman, S. D., Meadows, V. S., Wolf, E. T., Schwieterman, E., Charnay, B., Claire, M., Hébrard, E., and Trainer, M. G. (2016). The

Pale Orange Dot: The Spectrum and Habitability of Hazy Archean Earth. *Astrobiology,* 16(11):873–899.

Arney, G. N., Meadows, V. S., Domagal-Goldman, S. D., Deming, D., Robinson, T. D., Tovar, G., Wolf, E. T., and Schwieterman, E. (2017). Pale Orange Dots: The Impact of Organic Haze on the Habitability and Detectability of Earthlike Exoplanets. *Astrophys. J.,* 836(1):49.

Arney, G. N. (2019). The K Dwarf Advantage for Biosignatures on Directly Imaged Exoplanets. *Astrophys. J. Lett.,* 873(1):L7.

Arnold, L. F. A. (2005). Transit Light-Curve Signatures of Artificial Objects. *Astrophys. J.,* 627(1):534–539.

Arnold, L. (2013). Transmitting signals over interstellar distances: three approaches compared in the context of the Drake equation. *Int. J. Astrobiol.,* 12(3):212–217.

Arnscheidt, C. W., Wordsworth, R. D., and Ding, F. (2019). Atmospheric Evolution on Low-gravity Waterworlds. *Astrophys. J.,* 881(1):60.

Arrhenius, S. (1908). *Worlds in the Making: The Evolution of the Universe*. Harper & Brothers, New York, NY.

Artemieva, N. and Ivanov, B. (2004). Launch of martian meteorites in oblique impacts. *Icarus,* 171(1):84–101.

Astudillo-Defru, N., Delfosse, X., Bonfils, X., Forveille, T., Lovis, C., and Rameau, J. (2017). Magnetic activity in the HARPS M dwarf sample: The rotation-activity relationship for very low-mass stars through R'*HK*. *Astron. Astrophys.,* 600:A13.

Atri, D. (2017). Modelling stellar proton event-induced particle radiation dose on close-in exoplanets. *Mon. Not. R. Astron. Soc. Lett.,* 465(1):L34–L38.

Atri, D. (2020). Stellar Proton Event-induced surface radiation dose as a constraint on the habitability of terrestrial exoplanets. *Mon. Not. R. Astron. Soc. Lett.,* 492(1):L28–L33.

Attwater, J., Wochner, A., and Holliger, P. (2013). In-ice evolution of RNA polymerase ribozyme activity. *Nat. Chem.,* 5(12):1011–1018.

Attwater, J., Raguram, A., Morgunov, A. S., Gianni, E., and Holliger, P. (2018). Ribozyme-catalysed RNA synthesis using triplet building blocks. *eLife,* 7:e35255.

Atwater, H. A., Davoyan, A. R., Ilic, O., Jariwala, D., Sherrott, M. C., Went, C. M., Whitney, W. S., and Wong, J. (2018). Materials challenges for the Starshot lightsail. *Nat. Mater.,* 17:861–867.

Aubrey, A. D., Cleaves, H. J., and Bada, J. L. (2009). The Role of Submarine Hydrothermal Systems in the Synthesis of Amino Acids. *Orig. Life Evol. Biosph.,* 39(2):91–108.

Baaske, P., Weinert, F. M., Duhr, S., Lemke, K. H., Russell, M. J., and Braun, D. (2007). Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc. Natl. Acad. Sci. USA,* 104(22):9346–9351.

Bada, J. L. (2013). New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem. Soc. Rev.,* 42(5):2186–2196.

Bada, J. L. and Korenaga, J. (2018). Exposed areas above sea level on earth >3.5 Gyr ago: Implications for prebiotic and primitive biotic chemistry. *Life,* 8(4):55.

Badescu, V. and Cathcart, R. B. (2000). Stellar Engines for Kardashev's Type II Civilisations. *J. Br. Interplanet. Soc.,* 53:297–306.

Badescu, V., Cathcart, R. B., and Schuiling, R. D. (Eds.) (2006). *Macro-Engineering: A Challenge for the Future,* Water Science and Technology Library, Vol. 54. Springer, Dordrecht, Netherlands.

Bains, W. (2004). Many cemistries could be used to build living systems. *Astrobiology,* 4(2):137–167.

Bains, W. (2014). What do we think life is? A simple illustration and its consequences. *Int. J. Astrobiol.,* 13(2):101–111.

Bains, W., Seager, S., and Zsom, A. (2014). Photosynthesis in Hydrogen-Dominated Atmospheres. *Life,* 4(4):716–744.

Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., and Lloyd, K. G. (2020). Diversity, ecology and evolution of archaea. *Nat. Microbiol.,* 5:887–900.

Balbi, A. and Tombesi, F. (2017). The habitability of the Milky Way during the active phase of its central supermassive black hole. *Sci. Rep.,* 7:16626.

Balbi, A. (2018). The Impact of the Temporal Distribution of Communicating Civilizations on Their Detectability. *Astrobiology,* 18(1):54–58.

Balbus, S. A. (2014). Dynamical, biological and anthropic consequences of equal lunar and solar angular radii. *Proc. R. Soc. A,* 470(2168):20140263.

Ball, J. A. (1973). The Zoo Hypothesis. *Icarus,* 19(3):347–349.

Ball, P. (2017). Water is an active matrix of life for cell and molecular biology. *Proc. Natl. Acad. Sci. USA,* 114(51):13327–13335.

Ballesteros, F. J., Fernandez-Soto, A., and Martínez, V. J. (2019). Diving into exoplanets: Are water seas the most common? *Astrobiology,* 19(5):642–654.

Baluška, F. and Levin, M. (2016). On having no head: Cognition throughout biological systems. *Front. Psychol.,* 7:902.

Bambach, R. K. (2006). Phanerozoic biodiversity mass extinctions. *Annu. Rev. Earth Planet. Sci.,* 34:127–155.

Bannister, M. T., Bhandare, A., Dybczyński, P. A., Fitzsimmons, A., Guilbert-Lepoutre, A., Jedicke, R., Knight, M. M., Meech, K. J., McNeill, A., Pfalzner, S., Raymond, S. N., Snodgrass, C., Trilling, D. E., and Ye, Q. (2019). The natural history of 'Oumuamua. *Nat. Astron.,* 3:594–602.

Barclay, T., Quintana, E. V., Raymond, S. N., and Penny, M. T. (2017). The Demographics of Rocky Free-floating Planets and their Detectability by WFIRST. *Astrophys. J.,* 841(2):86.

Bardeen, C. G., Garcia, R. R., Toon, O. B., and Conley, A. J. (2017). On transient climate change at the Cretaceous-Paleogene boundary due to atmospheric soot injections. *Proc. Natl. Acad. Sci. USA,* 114(36):E7415–E7424.

Barge, L. M., Flores, E., Baum, M. M., VanderVelde, D. G., and Russell, M. J. (2019). Redox and pH gradients drive amino acid synthesis in iron oxyhydroxide mineral systems. *Proc. Natl. Acad. Sci. USA,* 116(11):4828–4833.

Barham, L. (2013). *From Hand to Handle: The First Industrial Revolution*. Oxford University Press, Oxford, UK.

Barkow, J. H., Cosmides, L., and Tooby, J. (Eds.) (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK.

Barnes, E. W. (1931). Contributions to a British Association Discussion on the Evolution of the Universe. *Nature,* 128(3234):719–722.

Barnes, R. (2017). Tidal locking of habitable exoplanets. *Celest. Mech. Dyn. Astron.,* 129(4):509–536.

Barney, B. L., Pratt, S. N., and Austin, D. E. (2016). Survivability of bare, individual *Bacillus subtilis* spores to high-velocity surface impact: Implications for microbial transfer through space. *Planet. Space Sci.,* 125:20–26.

Bar-On, Y. M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proc. Natl. Acad. Sci. USA,* 115(25):6506–6511.

Baross, J. A. and Hoffman, S. E. (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. Biosph.,* 15(4):327–345.

Baross, J. A., Benner, S. A., Cody, G. D., Copley, S. D., Pace, N. R., Scott, J. H., Shapiro, R., Sogin, M. L., Stein, J. L., Summons, R., and Szostak, J. W. (2007). *The Limits of Organic Life in Planetary Systems*. The National Academies Press, Washington, DC.

Bartlett, S. J. and Beckett, P. (2019). Probing complexity: Thermodynamics and computational mechanics approaches to origins studies. *Interface Focus,* 9(6):20190058.

Bartlett, S. and Wong, M. L. (2020). Defining lyfe in the Universe: From three privileged functions to four pillars. *Life,* 10(4):42.

Bar-Yosef, O. (2002). The Upper Paleolithic Revolution. *Annu. Rev. Anthropol.,* 31:363–393.

Batalha, N. E., Lewis, N. K., Line, M. R., Valenti, J., and Stevenson, K. (2018). Strategies for Constraining the Atmospheres of Temperate Terrestrial Planets with JWST. *Astrophys. J. Lett.,* 856(2):L34.

Baum, D. A. (2015). A comparison of autogenous theories for the origin of eukaryotic cells. *Am. J. Bot.,* 102(12):1954–1965.

Baumgartner, R. J., Van Kranendonk, M. J., Wacey, D., Fiorentini, M. L., Saunders, M., Caruso, S., Pages, A., Homann, M., and Guagliardo, P. (2019). Nanoporous pyrite and organic matter in 3.5-billion-year-old stromatolites record primordial life. *Geology,* 47(11):1039–1043.

Becker, S., Schneider, C., Okamura, H., Crisp, A., Amatov, T., Dejmek, M., and Carell, T. (2018). Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. *Nat. Commun.,* 9:163.

Becker, S., Feldmann, J., Wiedemann, S., Okamura, H., Schneider, C., Iwan, K., Crisp, A., Rossa, M., Amatov, T., and Carell, T. (2019). Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science,* 366(6461):76–82.

Beckstead, A. A., Zhang, Y., de Vries, M. S., and Kohler, B. (2016). Life in the light: Nucleic acid photoproperties as a legacy of chemical evolution. *Phys. Chem. Chem. Phys.,* 18(35):24228–24238.

Beech, M. (1990). Blue stragglers as indicators of extraterrestrial civilisations? *Earth Moon Planets,* 49:177–186.

Beerling, D. (2007). *The Emerald Planet: How Plants Changed Earth's History*. Oxford University Press, Oxford, UK.

Bejan, A. and Marden, J. H. (2006). Unifying constructal theory for scale effects in running, swimming and flying. *J. Exp. Biol.,* 209:238–248.

Bekenstein, J. D. (1981). Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D,* 23:287–298.

Bekoff, M. and Pierce, J. (2009). *Wild Justice: The Moral Lives of Animals*. The University of Chicago Press, Chicago, IL.

Belbruno, E., Moro-Martín, A., Malhotra, R., and Savransky, D. (2012). Chaotic Exchange of Solid Material Between Planetary Systems: Implications for Lithopanspermia. *Astrobiology,* 12(8):754–774.

Bell, E. A., Boehnke, P., Harrison, T. M., and Mao, W. L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. USA,* 112(47):14518–14521.

Belorizky, D. (1938). Le Soleil, Etoile Variable. *L'Astronomie,* 52:359–361.

Benford, J., Benford, G., and Benford, D. (2010). Messaging with Cost-Optimized Interstellar Beacons. *Astrobiology,* 10(5):475–490.

Benford, J. N. and Benford, D. J. (2016). Power Beaming Leakage Radiation as a SETI Observable. *Astrophys. J.,* 825(2):101.

Benford, J. (2019). Looking for Lurkers: Co-orbiters as SETI Observables. *Astron. J.,* 158(4):150.

Bengtson, S., Sallstedt, T., Belivanova, V., and Whitehouse, M. (2017). Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLoS Biol.,* 15(3):e2000735.

Benitez-Nelson, C. R. (2000). The biogeochemical cycling of phosphorus in marine systems. *Earth Sci. Rev.,* 51(1):109–135.

Benner, S. A., Ricardo, A., and Carrigan, M. A. (2004). Is there a common chemical model for life in the universe? *Curr. Opin. Chem. Biol.,* 8(6):672–689.

Benner, S. A. (2010). Defining life. *Astrobiology,* 10(10):1021–1030.

Benner, S. A., Kim, H.-J., and Carrigan, M. A. (2012). Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA. *Acc. Chem. Res.,* 45(12):2025–2034.

Benner, S. A. (2014). Paradoxes in the origin of life. *Orig. Life Evol. Biosph.,* 44(4):339–343.

Benner, S. A. and Kim, H.-J. (2015). The case for a Martian origin for Earth life. In *Instruments, Methods, and Missions for Astrobiology XVII,* volume 9606 of *Proc. SPIE,* page 96060C.

Benner, S. A., Kim, H.-J., and Biondi, E. (2019). Prebiotic chemistry that could not *not* have happened. *Life,* 9(4):84.

Benner, S. A., Bell, E. A., Biondi, E., Brasser, R., Carell, T., Kim, H.-J., Mojzsis, S. J., Omran, A., Pasek, M. A., and Trail, D. (2020). When did life likely emerge on Earth in an RNA-first process? *ChemSystemsChem,* 2(2):e1900035.

Bennett, C. H., Hanson, R., and Riedel, C. J. (2019). Comment on 'The Aestivation Hypothesis for Resolving Fermi's Paradox'. *Found. Phys.,* 49(8):820–829.

Benson-Amram, S., Dantzer, B., Stricker, G., Swanson, E. M., and Holekamp, K. E. (2016). Brain size predicts problem-solving ability in mammalian carnivores. *Proc. Natl. Acad. Sci. USA,* 113(9):2532–2537.

Benton, M. J. (2015). *Vertebrate Palaeontology*. Wiley-Blackwell, Chichester, UK (4th edition).

Berdyugina, S. V., Kuhn, J. R., Harrington, D. M., Šantl-Temkiv, T., and Messersmith, E. J. (2016). Remote sensing of life: polarimetric signatures of photosynthetic pigments as sensitive biomarkers. *Int. J. Astrobiol.,* 15(1):45–56.

Berdyugina, S. V., Kuhn, J. R., Langlois, M., Moretto, G., Krissansen-Totton, J., Catling, D., Grenfell, J. L., Santl-Temkiv, T., Finster, K., Tarter, J., Marchis, F., Hargitai, H., and Apai, D. (2018). The Exo-Life Finder (ELF) telescope: New strategies for direct detection of exoplanet biosignatures and technosignatures. In *Ground-based and Airborne Telescopes VII,* volume 10700 of *Proc. SPIE,* page 107004I.

Berera, A. (2017). Space Dust Collisions as a Planetary Escape Mechanism. *Astrobiology,* 17(12):1274–1282.

Berg, I. A. (2011). Ecological aspects of the distribution of different autotrophic CO_2 fixation pathways. *Appl. Environ. Microbiol.,* 77(6):1925–1936.

Berliner, A. J., Mochizuki, T., and Stedman, K. M. (2018). Astrovirology: Viruses at large in the Universe. *Astrobiology,* 18(2):207–223.

Bernal, J. D. (1929). *The World, the Flesh and the Devil: An Enquiry into the Future of the Three Enemies of the Rational Soul*. Kegan Paul, Trench, Trübner & Co., London, UK.

Bernal, J. D. (1951). *The Physical Basis of Life*. Routledge & Kegan Paul, London, UK.

Bernhard, J. M., Morrison, C. R., Pape, E., Beaudoin, D. J., Todaro, M. A., Pachiadaki, M. G., Ar. Kormas, K., and Edgcomb, V. P. (2015). Metazoans of redoxcline sediments in Mediterranean deep-sea hypersaline anoxic basins. *BMC Biol.,* 13:105.

Bernhardt, H. S. (2012). The RNA world hypothesis: The worst theory of the early evolution of life (except for all the others). *Biol. Direct,* 7:23.

Berwick, R. C. and Chomsky, N. (2016). *Why Only Us: Language and Evolution*. The MIT Press, Cambridge, MA.

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., and Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.,* 3(2):286–292.

Bétrémieux, Y. and Kaltenegger, L. (2013). Transmission Spectrum of Earth as a Transiting Exoplanet from the Ultraviolet to the Near-infrared. *Astrophys. J. Lett.,* 772(2):L31.

Betts, H. C., Puttick, M. N., Clark, J. W., Williams, T. A., Donoghue, P. C. J., and Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.,* 2:1556–1562.

Bialy, S., Sternberg, A., and Loeb, A. (2015). Water Formation During the Epoch of First Metal Enrichment. *Astrophys. J. Lett.,* 804(2):L29.

Bialy, S. and Loeb, A. (2018). Could Solar Radiation Pressure Explain 'Oumuamua's Peculiar Acceleration? *Astrophys. J. Lett.,* 868(1):L1.

Bich, L. and Green, S. (2018). Is defining life pointless? Operational definitions at the frontiers of biology. *Synthese,* 195(9):3919–3946.

Bickerton, D. (2014). *More than Nature Needs: Language, Mind, and Evolution*. Harvard University Press, Cambridge, MA.

Bierson, C. J., Nimmo, F., and Stern, S. A. (2020). Evidence for a hot start and early ocean formation on Pluto. *Nat. Geosci.,* 13(7):468–472.

Biller, B. A. and Bonnefoy, M. (2018). Exoplanet Atmosphere Measurements from Direct Imaging. In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets,* pages 2107–2135. Springer, Cham, Switzerland.

Birch, P. (1991). Supramundane Planets. *J. Br. Interplanet. Soc.,* 44:169–182.

Birkby, J. L. (2018). Spectroscopic Direct Detection of Exoplanets. In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets,* pages 1485–1508. Springer, Cham, Switzerland.

Björn, L. O. (1976). Why are plants green? Relationships between pigment absorption and photosynthetic efficiency. *Photosynthetica,* 10(2):121–129.

Björn, L. O. (Ed.) (2015). *Photobiology: The Science of Light and Life*. Springer-Verlag, New York, NY (3rd edition).

Blackman, E. G. and Tarduno, J. A. (2018). Mass, energy, and momentum capture from stellar winds by magnetized and unmagnetized planets: Implications for atmospheric erosion and habitability. *Mon. Not. R. Astron. Soc.,* 481(4):5146–5155.

Blackmond, D. G. (2019). The origin of biological homochirality. *Cold Spring Harb. Perspect. Biol.,* 11(3):a032540.

Blain, J. C. and Szostak, J. W. (2014). Progress toward synthetic cells. *Annu. Rev. Biochem.,* 83:615–640.

Blankenship, R. E. (2014). *Molecular Mechanisms of Photosynthesis*. Wiley-Blackwell, Chichester, UK (2nd edition).

Blatter, H. and Greber, T. (2017). Tau Zero: In the cockpit of a Bussard ramjet. *Am. J. Phys,* 85(12):915–920.

Blount, Z. D., Lenski, R. E., and Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science,* 362(6415):eaam5979.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. -U. (2006). Complex networks: Structure and dynamics. *Phys. Rep.,* 424:175–308.

Bogonovich, M. (2011). Intelligence's likelihood and evolutionary time frame. *Int. J. Astrobiol.,* 10(2):113–122.

Böhm-Vitense, E. (1992). *Introduction to Stellar Astrophysics,* vol. 3. Cambridge University Press, Cambridge, UK.

Boltzmann, L. (1974). The second law of thermodynamics. In McGinness, B., editor, *Theoretical Physics and Philosophical Problems: Select Writings,* pages 13–32. Reidel Publishing Co., Dordrecht, Netherlands.

Bond, A. and Martin, A. R. (1986). Project Daedalus reviewed. *J. Br. Interplanet. Soc.,* 39:385–390.

Bond, D. P. G. and Grasby, S. E. (2017). On the causes of mass extinctions. *Palaeogeogr. Palaeoclimatol. Palaeoecol,* 478:3–29.

Bonfils, X., Astudillo-Defru, N., Díaz, R., Almenara, J.-M., Forveille, T., Bouchy, F., Delfosse, X., Lovis, C., Mayor, M., Murgas, F., Pepe, F., Santos, N. C., Ségransan, D., Udry, S., and Wünsche, A. (2018). A temperate exo-Earth around a quiet M dwarf at 3.4 parsec. *Astron. Astrophys.,* 613:A25.

Bonfio, C., Valer, L., Scintilla, S., Shah, S., Evans, D. J., Jin, L., Szostak, J. W., Sasselov, D. D., Sutherland, J. D., and Mansy, S. S. (2017). UV-light-driven prebiotic synthesis of iron-sulfur clusters. *Nat. Chem.,* 9(12):1229–1234.

Booth, A. and Doolittle, W. F. (2015). Eukaryogenesis, how special really? *Proc. Natl. Acad. Sci. USA,* 112(33):10278–10285.

Borgue, O. and Hein, A. M. (2020). Near-Term Self-replicating Probes – A Concept Design. *arXiv e-prints,* arXiv:2005.12303.

Bornmann, L. and Marx, W. (2012). The Anna Karenina principle: A way of thinking about success in science. *J. Am. Soc. Inf. Sci. Tech.,* 63(10):2037–2051.

Bosak, T., Knoll, A. H., and Petroff, A. P. (2013). The meaning of stromatolites. *Annu. Rev. Earth Planet. Sci.,* 41:21–44.

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *J. Evol. Technol.,* 9:1–111.

Bostrom, N. and Ćirković, M. M. (Eds.) (2008). *Global Catastrophic Risks*. Oxford University Press, Oxford, UK.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK.

Bottke, W. F. and Norman, M. D. (2017). The Late Heavy Bombardment. *Annu. Rev. Earth Planet. Sci.,* 45:619–647.

Bouquet, A., Mousis, O., Glein, C. R., Danger, G., and Waite, J. H. (2019). The Role of Clathrate Formation in Europa's Ocean Composition. *Astrophys. J.,* 885(1):14.

Bourrier, V., de Wit, J., Bolmont, E., Stamenković, V., Wheatley, P. J., Burgasser, A. J., Delrez, L., Demory, B.-O., Ehrenreich, D., Gillon, M., Jehin, E., Leconte, J., Lederer, S. M., Lewis, N., Triaud, A. H. M. J., and Van Grootel, V. (2017). Temporal evolution of the high-energy irradiation and water content of TRAPPIST-1 exoplanets. *Astron. J.,* 154(3):121.

Bouvier, A. and Wadhwa, M. (2010). The age of the Solar system redefined by the oldest Pb-Pb age of a meteoritic inclusion. *Nat. Geosci.,* 3(9):637–641.

Bouwens, R. (2018). Distant galaxy formed stars only 250 million years after the Big Bang. *Nature,* 557(7705):312–313.

Bowler, B. P. (2016). Imaging Extrasolar Giant Planets. *Publ. Astron. Soc. Pac.,* 128(968):102001.

Bowles, S. and Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press, Princeton, NJ.

Bowman, J. C., Hud, N. V., and Williams, L. D. (2015). The ribosome challenge to the RNA world. *J. Mol. Evol.,* 80:143–161.

Braakman, R. and Smith, E. (2012). The emergence and early evolution of biological carbon-fixation. *PLOS Comput. Biol.,* 8(4): e1002455.

Bracewell, R. (1960). Communications from Superior Galactic Communities. *Nature,* 186(4726):670–671.

Bracewell, R. N. (1960). Communications from Superior Galactic Communities. *Nature,* 186(4726):670–671.

Bradbury, R. J., Cirkovic, M. M., and Dvorsky, G. (2011). Dysonian Approach to SETI: A Fruitful Middle Ground? *J. Br. Interplanet. Soc.,* 64:156–165.

Brain, D. A., Bagenal, F., Ma, Y.-J., Nilsson, H., and Stenberg Wieser, G. (2016). Atmospheric escape from unmagnetized bodies. *J. Geophys. Res. Planets,* 121(12):2364–2385.

Brandt, T. D. and Spiegel, D. S. (2014). Prospects for detecting oxygen, water, and chlorophyll on an exo-Earth. *Proc. Natl. Acad. Sci. USA,* 111(37):13278–13283.

Brandt, A., Posthoff, E., de Vera, J.-P., Onofri, S., and Ott, S. (2016). Characterisation of Growth and Ultrastructural Effects of the Xanthoria elegans Photobiont After 1.5 Years of Space Exposure on the International Space Station. *Orig. Life Evol. Biosph.,* 46:311–321.

Branscomb, E., Biancalani, T., Goldenfeld, N., and Russell, M. (2017). Escapement mechanisms and the conversion of disequilibria: The engines of creation. *Phys. Rep.,* 677:1–60.

Brasier, M. D., Matthewman, R., McMahon, S., and Wacey, D. (2011). Pumice as a remarkable substrate for the origin of life. *Astrobiology,* 11(7):725–735.

Brasier, M. D., Antcliffe, J., Saunders, M., and Wacey, D. (2015). Changing the picture of Earth's earliest fossils (3.5–1.9 Ga) with new approaches and new discoveries. *Proc. Natl. Acad. Sci. USA,* 112(16):4859–4864.

Breuer, D., Labrosse, S., and Spohn, T. (2010). Thermal Evolution and Magnetic Field Generation in Terrestrial Planets and Satellites. *Space Sci. Rev.,* 152(1–4): 449–500.

Brin, G. D. (1983). The Great Silence—the Controversy Concerning Extraterrestrial Intelligent Life, *Q. Jl. R. Astr. Soc.,* 24(3):283–309.

Brini, E., Fennell, C. J., Fernandez-Serra, M., Hribar-Lee, B., Lukšič, M., and Dill, K. A. (2017). How water's properties are encoded in its molecular structure and energies. *Chem. Rev.,* 117(19):12385–12414.

Brock, T. D. (1985). Life at High Temperatures. *Science,* 230(4722):132–138.

Brocks, J. J., Logan, G. A., Buick, R., and Summons, R. E. (1999). Archean molecular fossils and the early rise of eukaryotes. *Science,* 285(5430):1033–1036.

Brocks, J. J., Jarrett, A. J. M., Sirantoine, E., Hallmann, C., Hoshino, Y., and Liyanage, T. (2017). The rise of algae in Cryogenian oceans and the emergence of animals. *Nature,* 548(7669):578–581.

Brown, C. (2015). Fish intelligence, sentience and ethics. *Animal Cogn.,* 18(1): 1–17.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., and West, G. B. (2004). Toward a metabolic theory of ecology. *Ecology,* 85(7):1771–1789.

Brown, M., Johnson, T., and Gardiner, N. J. (2020). Plate Tectonics and theArchean Earth. *Annu. Rev. Earth Planet. Sci.,* 48:12.

Brunet, T. and King, N. (2017). The origin of animal multicellularity and cell differentiation. *Dev. Cell,* 43(2):124–140.

Brzycki, B., Siemion, A., Croft, S., Czech, D., DeBoer, D., DeMarines, J., Drew, J., Enriquez, J. E., Gajjar, V., Gizani, N., Isaacson, H., Lacki, B., Lebofsky, M., MacMahon, D. H. E., de Pater, I., Price, D. C., Sheikh, S., Webb, C., and Worden, S. P. (2019). Breakthrough Listen Follow-up of the Random Transiter (EPIC 249706694/HD 139139) with the Green Bank Telescope. *Res. Notes AAS,* 3(10):147.

Buckel, W. and Thauer, R. K. (2018). Flavin-based electron bifurcation, ferredoxin, flavodoxin, and anaerobic respiration with protons (Ech) or NAD^+ (Rnf) as electron acceptors: A historical review. *Front. Microbiol.,* 9:401.

Budin, I. and Szostak, J. W. (2010). Expanding Roles for Diverse Physical Phenomena During the Origin of Life. *Annu. Rev. Biophys.,* 39:245–263.

Bulzu, P.-A., Andrei, A.-Ş., Salcher, M. M., Mehrshad, M., Inoue, K., Kandori, H., Beja, O., Ghai, R., and Banciu, H. L. (2019). Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat. Microbiol.,* 4(7):1129–1137.

Burcar, B. T., Barge, L. M., Trail, D., Watson, E. B., Russell, M. J., and McGown, L. B. (2015). RNA Oligomerization in laboratory analogues of alkaline hydrothermal vent systems. *Astrobiology,* 15(7):509–522.

Burcar, B., Castañeda, A., Lago, J., Daniel, M., Pasek, M. A., Hud, N. V., Orlando, T. M., and Menor-Salván, C. (2019). A stark contrast to modern earth: Phosphate mineral transformation and nucleoside phosphorylation in an iron- and cyanide-rich early Earth scenario. *Angew. Chem. Int. Ed.,* 58(47):16981–16987.

Burchell, M. J., Galloway, J. A., Bunch, A. W., and Brandão, P. F. B. (2003). Survivability of Bacteria Ejected from Icy Surfaces after Hypervelocity Impact. *Orig. Life Evol. Biosph.,* 33(1):53–74.

Burchell, M. J. (2004). Panspermia today. *Int. J. Astrobiol.,* 3(2):73–80.

Burchell, M. J., Mann, J. R., and Bunch, A. W. (2004). Survival of bacteria and spores under extreme shock pressures. *Mon. Not. R. Astron. Soc.,* 352(4):1273–1278.

Burgess, S. (2019). Deciphering mass extinction triggers. *Science,* 363(6429):815–816.

Burkart, J. M., Schubiger, M. N., and van Schaik, C. P. (2017). The evolution of general intelligence. *Behav. Brain Sci.,* 40:e195.

Burkhart, B. and Loeb, A. (2017). The Detectability of Radio Auroral Emission from Proxima b. *Astrophys. J. Lett.,* 849(1):L10.

Burrows, A. and Liebert, J. (1993). The science of brown dwarfs. *Rev. Mod. Phys.,* 65(2):301–336.

Burrows, A. S. (2014). Spectra as windows into exoplanet atmospheres. *Proc. Natl. Acad. Sci. USA,* 111(35):12601–12609.

Burton, A. S., Stern, J. C., Elsila, J. E., Glavin, D. P., and Dworkin, J. P. (2012). Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev.,* 41(16):5459–5472.

Buss, D. M. (2019). *Evolutionary Psychology: The New Science of the Mind*. Routledge, New York, NY.

Bussard, R. W. (1960). Galactic Matter and Interstellar Flight. *Astronautica Acta.,* 6(4):179–194.

Butterfield, N. J. (2011). Animals and the invention of the Phanerozoic Earth system. *Trends Ecol. Evol.,* 26(2):81–87.

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., and Tanner, J. (2017). Great ape gestures: Intentional communication with a rich set of innate signals. *Animal Cogn.,* 20(4):755–769.

Cabrol, N. A. (2016). Alien Mindscapes–A Perspective on the Search for Extraterrestrial Intelligence. *Astrobiology,* 16(9):661–676.

Cafferty, B. J., Fialho, D. M., Khanam, J., Krishnamurthy, R., and Hud, N. V. (2016). Spontaneous formation and base pairing of plausible prebiotic nucleotides in water. *Nat. Commun.,* 7:11328.

Cairns-Smith, A. G. (1982). *Genetic takeover and the mineral origins of life*. Cambridge University Press, Cambridge, UK.

Callahan, M. P., Smith, K. E., Cleaves, H. J., Ruzicka, J., Stern, J. C., Glavin, D. P., House, C. H., and Dworkin, J. P. (2011). Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc. Natl. Acad. Sci. USA,* 108(34):13995–13998.

Calvo, P., Gagliano, M., Souza, G. M., and Trewavas, A. (2020). Plants are intelligent, here's how. *Ann. Bot.,* 125(1):11–28.

Campbell, J. B. (2006). Archaeology and direct imaging of exoplanets. In Aime, C. and Vakili, F., editors, *IAU Colloq. 200: Direct Imaging of Exoplanets—Science & Techniques,* pages 247–250. Cambridge University Press, Cambridge, UK.

Campbell, T. D., Febrian, R., McCarthy, J. T., Kleinschmidt, H. E., Forsythe, J. G., and Bracher, P. J. (2019). Prebiotic condensation through wet-dry cycling regulated by deliquescence. *Nat. Commun.,* 10:4508.

Camprubi, E., Jordan, S. F., Vasiliadou, R., and Lane, N. (2017). Iron catalysis at the origin of life. *IUBMB Life,* 69(6):373–381.

Camprubí, E., de Leeuw, J. W., House, C. H., Raulin, F., Russell, M. J., Spang, A., Tirumalai, M. R., and Westall, F. (2019). The emergence of life. *Space Sci. Rev.,* 215:56.

Canavelli, P., Islam, S., and Powner, M. W. (2019). Peptide ligation by chemoselective aminonitrile coupling in water. *Nature,* 571(7766):546–549.

Candau, Y. (2003). On the exergy of radiation. *Sol. Energy,* 75(3):241–247.

Canfield, D. E., Rosing, M. T., and Bjerrum, C. (2006). Early anaerobic metabolisms. *Phil. Trans. R. Soc. B,* 361(1474):1819–1836.

Cano, R. J. and Borucki, M. K. (1995). Revival and Identification of Bacterial Spores in 25- to 40-Million-Year-Old Dominican Amber. *Science,* 268(5213):1060–1064.

Cao, C. (2014). *Behold the Dark Green Sea*. Translated by D. Bowles. https://davidbowles.us/poetry/behold-the-dark-green-sea/.

Caplan, M. E. (2019). Stellar engines: Design considerations for maximizing acceleration. *Acta Astronaut.,* 165:96–104.

Cardona, T., Sánchez-Baracaldo, P., Rutherford, A. W., and Larkum, A. W. (2019). Early Archean origin of photosystem II. *Geobiology,* 17(2):127–150.

Carmody, R. N. and Wrangham, R. W. (2009). The energetic significance of cooking. *J. Hum. Evol.,* 57(4):379–391.

Carrier, B. L., Beaty, D. W., Meyer, M. A., Blank, J. G., Chou, L., DasSarma, S., Des Marais, D. J., Eigenbrode, J. L., Grefenstette, N., Lanza, N. L., Schuerger, A. C., Schwendner, P., Smith, H. D., Stoker, C. R., Tarnas, J. D., Webster, K. D., Bakermans, C., Baxter, B. K., Bell, M. S., . . . Xu, J. (2020). Mars Extant Life: What's Next? Conference Report. *Astrobiology,* 20(6):785–814.

Carrigan, R. A. (2009). IRAS-Based Whole-Sky Upper Limit on Dyson Spheres. *Astrophys. J.,* 698(2):2075–2086.

Carrigan, R. A. (2012). Is interstellar archeology possible? *Acta Astronaut.,* 78:121–126.

Carroll-Nellenback, J., Frank, A., Wright, J., and Scharf, C. (2019). The Fermi Paradox and the Aurora Effect: Exo-civilization Settlement, Expansion, and Steady States. *Astron. J.,* 158(3):117.

Carter, B. (1974). Large number coincidences and the anthropic principle in cosmology. In Longair, M. S., editor, *Confrontation of Cosmological Theories with Observational Data,* IAU Symposium, Vol. 63, pages 291–298. D. Reidel Publishing Co., Dordrecht, Netherlands.

Carter, B. (1983). The anthropic principle and its implications for biological evolution. *Phil. Trans. R. Soc. A,* 310(1512):347–363.

Carter, B. (2008). Five- or six-step scenario for evolution? *Int. J. Astrobiol.,* 7(2):177–182.

Carter, C. W. (2015). What RNA world? Why a peptide/RNA partnership merits renewed experimental attention. *Life,* 5(1):294–320.

Cassibry, J., Cortez, R., Stanic, M., Watts, A., Seidler, W., Adams, R., Statham, G., and Fabisinski, L. (2015). Case and Development Path for Fusion Propulsion. *J. Spacecraft Rockets,* 52(2):595–612.

Castelle, C. J. and Banfield, J. F. (2018). Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell,* 172(6):1181–1197.

Castillo-Rogez, J. C., Neveu, M., Scully, J. E. C., House, C. H., Quick, L. C., Bouquet, A., Miller, K., Bland, M., De Sanctis, M. C., Ermakov, A., Hendrix, A. R., Prettyman, T. H., Raymond, C. A., Russell, C. T., Sherwood, B. E., and Young, E. (2020). Ceres: Astrobiological Target and Possible Ocean World. *Astrobiology,* 20(2):269–291.

Catling, D. C., Zahnle, K. J., and McKay, C. P. (2001). Biogenic Methane, Hydrogen Escape, and the Irreversible Oxidation of Early Earth. *Science,* 293(5531):839–843.

Catling, D. C., Glein, C. R., Zahnle, K. J., and McKay, C. P. (2005). Why O_2 is required by complex life on habitable planets and the concept of planetary "oxygenation time." *Astrobiology,* 5(3):415–438.

Catling, D. C. and Kasting, J. F. (2017). *Atmospheric Evolution on Inhabited and Lifeless Worlds*. Cambridge University Press, Cambridge, UK.

Catling, D. C., Krissansen-Totton, J., Kiang, N. Y., Crisp, D., Robinson, T. D., DasSarma, S., Rushby, A. J., Del Genio, A., Bains, W., and Domagal-Goldman, S. (2018). Exoplanet Biosignatures: A Framework for Their Assessment. *Astrobiology,* 18(6):709–738.

Catling, D. C. and Zahnle, K. J. (2020). The Archean atmosphere. *Sci. Adv.,* 6(9):eaax1420.

Cavalier-Smith, T. (2009). Predation and eukaryote cell origins: A coevolutionary perspective. *Int. J. Biochem. Cell Biol.,* 41(2):307–322.

Cavalier-Smith, T. (2010). Origin of the cell nucleus, mitosis and sex: Roles of intracellular coevolution. *Biol. Direct,* 5:7.

Cavalier-Smith, T. (2017). Origin of animal multicellularity: precursors, causes, consequences–the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Phil. Trans. R. Soc. B,* 372(1713):20150476.

Ceplecha, Z., Borovička, J., Elford, W. G., Revelle, D. O., Hawkes, R. L., Porubčan, V., and Šimek, M. (1998). Meteor Phenomena and Bodies. *Space Sci. Rev.,* 84:327–471.

Chabrier, G. and Baraffe, I. (2000). Theory of low-mass stars and substellar objects. *Annu. Rev. Astron. Astrophys.,* 38:337–377.

Chan, M. A., Hinman, N. W., Potter-McIntyre, S. L., Schubert, K. E., Gillams, R. J., Awramik, S. M., Boston, P. J., Bower, D. M., Des Marais, D. J., Farmer, J. D., Jia, T. Z., King, P. L., Hazen, R. M., Léveillé, R. J., Papineau, D., Rempfert, K. R., Sánchez-Román, M., Spear, J. R., Southam, . . . Cleaves, H. J. (2019). Deciphering Biosignatures in Planetary Contexts. *Astrobiology,* 19(9):1075–1102.

Chandru, K., Gilbert, A., Butch, C., Aono, M., and Cleaves, H. J. (2016). The abiotic chemistry of thiolated acetate derivatives and the origin of life. *Sci. Rep.,* 6:29883.

Charnay, B., Wolf, E. T., Marty, B., and Forget, F. (2020). Is the faint young sun problem for Earth solved? *Space Sci. Rev.,* 216(5):90.

Chen, H., Forbes, J. C., and Loeb, A. (2018a). Habitable Evaporated Cores and the Occurrence of Panspermia Near the Galactic Center. *Astrophys. J. Lett.,* 855(1):L1.

Chen, H., Wolf, E. T., Kopparapu, R., Domagal-Goldman, S., and Horton, D. E. (2018b). Biosignature Anisotropy Modeled on Temperate Tidally Locked M-dwarf Planets. *Astrophys. J. Lett.,* 868(1):L6.

Chen, I. A., Roberts, R. W., and Szostak, J. W. (2004). The emergence of competition between model protocells. *Science,* 305(5689):1474–1476.

Chen, J. and Kipping, D. (2018). On the rate of abiogenesis from a bayesian informatics perspective. *Astrobiology,* 18(12):1574–1584.

Chen, X., Ling, H.-F., Vance, D., Shields-Zhou, G. A., Zhu, M., Poulton, S. W., Och, L. M., Jiang, S.-Y., Li, D., Cremonese, L., and Archer, C. (2015). Rise to modern levels of ocean oxygenation coincided with the Cambrian radiation of animals. *Nat. Commun.,* 6:7142.

Cheney, D. L. and Seyfarth, R. M. (2007). *Baboon Metaphysics: The Evolution of a Social Mind*. The University of Chicago Press, Chicago, IL.

Chennamangalam, J., Siemion, A. P. V., Lorimer, D. R., and Werthimer, D. (2015). Jumping the energetics queue: Modulation of pulsar signals by extraterrestrial civilizations. *New. Astron.,* 34:245–249.

Chintalapati, M. and Moorjani, P. (2020). Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.,* 62:58–64.

Chittka, L. and Niven, J. (2009). Are bigger brains better? *Curr. Biol.,* 19(21):R995–R1008.

Chittka, L., Giurfa, M., and Riffell, J. A. (2019). The mechanisms of insect cognition. *Front. Psychol.,* 10:2751.

Choblet, G., Tobie, G., Sotin, C., Běhounková, M., Čadek, O., Postberg, F., and Souček, O. (2017a). Powering prolonged hydrothermal activity inside Enceladus. *Nat. Astron.,* 1:841–847.

Choblet, G., Tobie, G., Sotin, C., Kalousová, K., and Grasset, O. (2017b). Heat transport in the high-pressure ice mantle of large icy moons. *Icarus,* 285:252–262.

Chopra, A. and Lineweaver, C. H. (2016). The Case for a Gaian Bottleneck: The Biology of Habitability. *Astrobiology,* 16(1):7–22.

Choueiri, E. Y. (2004). A Critical History of Electric Propulsion: The First 50 Years (1906–1956). *J. Propuls. Power,* 20(2):193–203.

Christensen, U. R. (2010). Dynamo scaling laws and applications to the planets. *Space Sci. Rev.,* 152:565–590.

Church, G. M., Gao, Y., and Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. *Science,* 337(6102):1628.

Churkina, G. and Running, S. W. (1998). Contrasting Climatic Controls on the Estimated Productivity of Global Terrestrial Biomes. *Ecosystems,* 1(2):206–215.

Chyba, C. and Sagan, C. (1992). Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: An inventory for the origins of life. *Nature,* 355(6356):125–132.

Chyba, C. F. and Hand, K. P. (2001). Life Without Photosynthesis. *Science,* 292(5524):2026–2027.

Chyba, C. F. and Phillips, C. B. (2001). Possible ecosystems and the search for life on Europa. *Proc. Natl. Acad. Sci. USA,* 98(3):801–804.

Ćirković, M. M. (2004). The Temporal Aspect of the Drake Equation and SETI. *Astrobiology,* 4(2):225–231.

Ćirković, M. M., Vukotić, B. and Dragićević, I. (2009). Galactic Punctuated Equilibrium: How to Undermine Carter's Anthropic Argument in Astrobiology. *Astrobiology,* 9(5):491–501.

Ćirković, M. M. (2015). Kardashev's Classification at 50+: A Fine Vehicle With Room for Improvement. *Serb. Astron. J.,* 191:1–15.

Ćirković, M. M. and Vukotić, B. (2016). Long-term prospects: Mitigation of supernova and gamma-ray burst threat to intelligent beings. *Acta Astronaut.,* 129:438–446.

Ćirković, M. M. (2018a). *The Great Silence: Science and Philosophy of Fermi's Paradox*. Cambridge University Press, Cambridge, UK.

Ćirković, M. M. (2018b). Woodpeckers and diamonds: Some aspects of evolutionary convergence in astrobiology. *Astrobiology,* 18(5):491–502.

Clack, J. A. (2012). *Gaining Ground: The Origin and Evolution of Tetrapods*. Indiana University Press, Bloomington, IN (2nd edition).

Claessen, D., Rozen, D. E., Kuipers, O. P., Søgaard-Andersen, L., and van Wezel, G. P. (2014). Bacterial solutions to multicellularity: A tale of biofilms, filaments and fruiting bodies. *Nat. Rev. Microbiol.,* 12:115–124.

Clapham, M. E. and Renne, P. R. (2019). Flood basalts and mass extinctions. *Annu. Rev. Earth Planet. Sci.,* 47:275–303.

Clark, B. C. (2001). Planetary Interchange of Bioactive Material: Probability Factors and Implications. *Orig. Life Evol. Biosph.,* 31:185–197.

Clark, G. and Henneberg, M. (2017). *Ardipithecus ramidus* and the evolution of language and singing: An early origin for hominin vocal capability. *HOMO,* 68(2):101–121.

Clark, J. R. and Cahoy, K. (2018). Optical Detection of Lasers with Near-term Technology at Interstellar Distances. *Astrophys. J.,* 867(2):97.

Clarke, A. C. (1968). *2001: A Space Odyssey*. Hutchinson, London, UK.

Clarke, A. (2014). The thermal limits to life on Earth. *Int. J. Astrobiol.,* 13(2):141–154.

Clarke, A. (2017). *Principles of Thermal Ecology: Temperature, Energy and Life*. Cambridge University Press, Cambridge, UK.

Clayton, N. S. and Emery, N. J. (2015). Avian models for human cognitive neuroscience: A proposal. *Neuron,* 86(6):1330–1342.

Cleaves, H. J. and Miller, S. L. (1998). Oceanic Protection of Prebiotic Organic Compounds from UV Radiation. *Proc. Natl. Acad. Sci. USA,* 95(13):7260–7263.

Cleaves, H. J., Scott, A. M., Hill, F. C., Leszczynski, J., Sahai, N., and Hazen, R. (2012). Mineral-organic interfacial processes: Potential roles in the origins of life. *Chem. Soc. Rev.,* 41(16):5502–5525.

Cleaves, H. J., Meringer, M., and Goodwin, J. (2015). 227 views of RNA: Is RNA unique in its chemical isomer space? *Astrobiology,* 15(7):538–558.

Cleaves, H. J., Butch, C., Burger, P. B., Goodwin, J., and Meringer, M. (2019). One among millions: The chemical space of nucleic acid-like molecules. *J. Chem. Inf. Model.,* 59(10):4266–4277.

Cleland, C. E. (2019). *The Quest for a Universal Theory of Life: Searching for Life As We Don't Know It*. Cambridge University Press, Cambridge, UK.

Cleland, C. E. and Chyba, C. F. (2002). Defining "life." *Orig. Life Evol. Biosph.,* 32(4):387–393.

Clements, D. L. (2018). Life Before Fermi—Back to the Solar System. *J. Br. Interplanet. Soc.,* 71:222–224.

Cocconi, G. and Morrison, P. (1959). Searching for Interstellar Communications. *Nature,* 184(4690):844–846.

Cockell, C. S. (1999). Life on venus. *Planet. Space Sci.,* 47(12):1487–1501.

Cockell, C. S. and Knowland, J. (1999). Ultraviolet radiation screening compounds. *Biol. Rev.,* 74(3):311–345.

Cockell, C. S. (2014). Habitable worlds with no signs of life. *Phil. Trans. R. Soc. A,* 372(2014):20130082.

Cockell, C. S., Bush, T., Bryce, C., Direito, S., Fox-Powell, M., Harrison, J. P., Lammer, H., Landenmark, H., Martin-Torres, J., Nicholson, N., Noack, L., O'Malley-James, J., Payler, S. J., Rushby, A., Samuels, T., Schwendner, P., Wadsworth, J., and Zorzano, M. P. (2016). Habitability: A Review. *Astrobiology,* 16(1):89–117.

Cody, G. D. (2004). Transition metal sulfides and the origins of metabolism. *Annu. Rev. Earth Planet. Sci.,* 32:569–599.

Cohen, O., Glocer, A., Garraffo, C., Drake, J. J., and Bell, J. M. (2018). Energy dissipation in the upper atmospheres of TRAPPIST-1 planets. *Astrophys. J. Lett.,* 856(1):L11.

Cole, D. B., Mills, D. B., Erwin, D. H., Sperling, E. A., Porter, S. M., Reinhard, C. T., and Planavsky, N. J. (2020). On the co-evolution of surface oxygen levels and animals. *Geobiology,* 18(3):260–281.

Colomer, I., Borissov, A., and Fletcher, S. P. (2020). Selection from a pool of self-assembling lipid replicators. *Nat. Commun.,* 11:176.

Comisso, L., Lingam, M., Huang, Y.-M., and Bhattacharjee, A. (2016). General theory of the plasmoid instability. *Phys. Plasmas,* 23(10):100702.

Comisso, L., Lingam, M., Huang, Y.-M., and Bhattacharjee, A. (2017). Plasmoid Instability in Forming Current Sheets. *Astrophys. J.,* 850(2):142.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., and Friston, K. J. (2018). A variational approach to niche construction. *J. R. Soc. Interface,* 15(141):20170685.

Cook, P. and Wilson, M. (2010). Do young chimpanzees have extraordinary working memory? *Psychon. Bull. Rev.,* 17(4):599–600.

Coolidge, F. L. and Wynn, T. (2018). *The Rise of Homo Sapiens: The Evolution of Modern Thinking*. Oxford University Press, Oxford, UK (2nd edition).

Cooper, J. F., Johnson, R. E., Mauk, B. H., Garrett, H. B., and Gehrels, N. (2001). Energetic Ion and Electron Irradiation of the Icy Galilean Satellites. *Icarus,* 149(1):133–159.

Copley, S. D., Smith, E., and Morowitz, H. J. (2007). The origin of the RNA world: Co-evolution of genes and metabolism. *Bioorg. Chem.,* 35(6): 430–443.

Corballis, M. C. (2017). *The Truth about Language: What It Is and Where It Came From*. The University of Chicago Press, Chicago, IL.

Corbet, R. H. D. (1997). SETI at X-Ray Energies: Parasitic Searches from Astrophysical Observations. *J. Br. Interplanet. Soc.,* 50(7):253–257.

Corbet, R. H. D. (1999). The Use of Gamma-Ray Bursts as Direction and Time Markers in SETI Strategies. *Publ. Astron. Soc. Pac.,* 111(761):881–885.

Corbet, R. H. D. (2003). Synchronized SETI–The Case for "Opposition". *Astrobiology,* 3(2):305–315.

Cordes, J. M. and Lazio, T. J. (1991). Interstellar scattering effects on the detection of narrow-band signals. *Astrophys. J.,* 376:123–133.

Cornell, C. E., Black, R. A., Xue, M., Litz, H. E., Ramsay, A., Gordon, M., Mileant, A., Cohen, Z. R., Williams, J. A., Lee, K. K., Drobny, G. P., and Keller, S. L. (2019). Prebiotic amino acids bind to and stabilize prebiotic fatty acid membranes. *Proc. Natl. Acad. Sci. USA,* 116(35):17239–17244.

Cornish-Bowden, A. and Cárdenas, M. L. (2020). Contrasting theories of life: Historical context, current theories. In search of an ideal theory. *Biosystems,* 188:104063.

Cossins, A. R. and Bowler, K. (1987). *Temperature Biology of Animals*. Chapman & Hall, London, UK.

Cottin, H., Kotler, J. M., Billi, D., Cockell, C., Demets, R., Ehrenfreund, P., Elsaesser, A., d'Hendecourt, L., van Loon, J. J. W. A., Martins, Z., Onofri, S., Quinn, R. C., Rabbow, E., Rettberg, P., Ricco, A. J., Slenzka, K., de la Torre, R., de Vera, J.-P., Westall, F., . . . Klamm, B. A. (2017). Space as a Tool for Astrobiology: Review and Recommendations for Experimentations in Earth Orbit and Beyond. *Space Sci. Rev.,* 209:83–181.

Coupé, C., Oh, Y., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.,* 5(9):eaaw2594.

Coveney, P. V., Swadling, J. B., Wattis, J. A. D., and Greenwell, H. C. (2012). Theory, modelling and simulation in origins of life studies. *Chem. Soc. Rev.,* 41(16):5430–5446.

Cowan, N. B., Abbot, D. S., and Voigt, A. (2012). A False Positive for Ocean Glint on Exoplanets: The Latitude-Albedo Effect. *Astrophys. J. Lett.,* 752(1):L3.

Cowan, N. B. and Abbot, D. S. (2014). Water Cycling between Ocean and Mantle: Super-Earths Need Not Be Waterworlds. *Astrophys. J.,* 781(1):27.

Cranmer, S. R. (2017). Mass-loss Rates from Coronal Mass Ejections: A Predictive Theoretical Model for Solar-type Stars. *Astrophys. J.,* 840(2):114.

Crapo, J. D. (1986). Morphologic Changes in Pulmonary Oxygen Toxicity. *Annu. Rev. Physiol.,* 48:721–731.

Crawford, I. A. (1990). Interstellar Travel: A Review for Astronomers. *Q. Jl. R. Astr. Soc.,* 31:377–400.

Crick, F. H. C. and Orgel, L. E. (1973). Directed panspermia. *Icarus,* 19(3):341–346.

Crutzen, P. J., Isaksen, I. S. A., and Reid, G. C. (1975). Solar Proton Events: Stratospheric Sources of Nitric Oxide. *Science,* 189(4201):457–459.

Cunningham, J. A., Liu, A. G., Bengtson, S., and Donoghue, P. C. J. (2017). The origin of animals: Can molecular clocks and the fossil record be reconciled? *BioEssays,* 39(1):1–12.

Dacks, J. B. and Field, M. C. (2018). Evolutionary origins and specialisation of membrane transport. *Curr. Opin. Cell Biol.,* 53:70–76.

Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.,* 13(6):e1006810.

Da Cunha, V., Gaia, M., Nasir, A., and Forterre, P. (2018). Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet.,* 14(3):e1007215.

Daeschler, E. B., Shubin, N. H., and Jenkins, F. A. (2006). A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature,* 440(7085):757–763.

Dai, X. and Guerras, E. (2018). Probing Extragalactic Planets Using Quasar Microlensing. *Astrophys. J. Lett.,* 853(2):L27.

Dalai, P. and Sahai, N. (2019). Mineral-lipid interactions in the origins of life. *Trends Biochem. Sci.,* 44(4):331–341.

Damasio, A. (2012). *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon Books, New York, NY.

Damer, B. and Deamer, D. (2020). The hot spring hypothesis for an origin of life. *Astrobiology,* 20(4):429–452.

Danger, G., Plasson, R., and Pascal, R. (2012). Pathways for the formation and evolution of peptides in prebiotic environments. *Chem. Soc. Rev.,* 41(16):5416–5429.

Danger, G., Le Sergeant d'Hendecourt, L., and Pascal, R. (2020). On the conditions for mimicking natural selection in chemical systems. *Nat. Rev. Chem.,* 4:102–109.

Danovaro, R., Dell'Anno, A., Pusceddu, A., Gambi, C., Heiner, I., and Kristensen, R. M. (2010). The first metazoa living in permanently anoxic conditions. *BMC Biol.,* 8:30.

Danovaro, R., Gambi, C., Dell'Anno, A., Corinaldesi, C., Pusceddu, A., Neves, R. C., and Kristensen, R. M. (2016). The challenge of proving the existence of metazoan life in permanently anoxic deep-sea sediments. *BMC Biol.,* 14:43.

Darroch, S. A. F., Smith, E. F., Laflamme, M., and Erwin, D. H. (2018). Ediacaran extinction and Cambrian explosion. *Trends Ecol. Evol.,* 33(9):653–663.

Dartnell, L. R. (2011). Ionizing Radiation and Life. *Astrobiology,* 11(6):551–582.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, London, UK.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. John Murray, London, UK.

DasSarma, S. and Schwieterman, E. W. (2018). Early evolution of purple retinal pigments on Earth and implications for exoplanet biosignatures. *Int. J. Astrobiol.,* DOI: 10.1017/S1473550418000423.

Davenport, J. R. A. (2016). The Kepler Catalog of Stellar Flares. *Astrophys. J.,* 829(1):23.

Davenport, J. R. A., Kipping, D. M., Sasselov, D., Matthews, J. M., and Cameron, C. (2016). MOST Observations of Our Nearest Neighbor: Flares on Proxima Centauri. *Astrophys. J. Lett.,* 829(2):L31.

Davies, P. (2010). *The Eerie Silence: Renewing Our Search for Alien Intelligence*. Houghton Mifflin Harcourt, Boston, MA.

Davies, P. C. W., Benner, S. A., Cleland, C. E., Lineweaver, C. H., McKay, C. P., and Wolfe-Simon, F. (2009). Signatures of a shadow biosphere. *Astrobiology,* 9(2):241–249.

Davies, P. C. W. and Wagner, R. V. (2013). Searching for alien artifacts on the moon. *Acta Astronaut.,* 89:261–265.

Dawkins, R. and Wong, Y. (2016). *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*. Mariner Books, Boston, MA (2nd edition).

Deamer, D. W. (1997). The first living systems: a bioenergetic perspective. *Microbiol. Mol. Biol. Rev.,* 61(2):239–261.

Deamer, D. and Weber, A. L. (2010). Bioenergetics and life's origins. *Cold Spring Harb. Perspect. Biol.,* 2(2):a004929.

Deamer, D. W. and Georgiou, C. D. (2015). Hydrothermal conditions and the origin of cellular life. *Astrobiology,* 15(12):1091–1095.

Deamer, D. (2017). The role of lipid membranes in life's origin. *Life,* 7(1):5.

Deamer, D. (2019). *Assembling Life: How Can Life Begin on Earth and Other Habitable Planets?* Oxford University Press, Oxford, UK.

Deardorff, J. W. (1987). Examination of the embargo hypothesis as an explanation for the Great Silence. *J. Br. Interplanet. Soc.,* 40:373–379.

de Duve, C. (1991). *Blueprint for a Cell: The Nature and Origin of Life*. Neil Patterson Publishers, Burlington, NC.

de Duve, C. and Miller, S. L. (1991). Two-dimensional life? *Proc. Natl. Acad. Sci. USA,* 88(22):10014–10017.

de Duve, C. (2005). *Singularities: Landmarks on the Pathways of Life*. Cambridge University Press, Cambridge, UK.

Deguchi, S., Shimoshige, H., Tsudome, M., Mukai, S.-a., Corkery, R. W., Ito, S., and Horikoshi, K. (2011). Microbial growth at hyperaccelerations up to $403,627 \times g$. *Proc. Natl. Acad. Sci. USA,* 108(19):7997–8002.

de la Torre, R., Sancho, L. G., Horneck, G., de los Ríos, A., Wierzchos, J., Olsson-Francis, K., Cockell, C. S., Rettberg, P., Berger, T., de Vera, J.-P. P., Ott, S., Frías, J. M., Melendi, P. G., Lucas, M. M., Reina, M., Pintado, A., and Demets, R. (2010). Survival of lichens and bacteria exposed to outer space conditions—Results of the Lithopanspermia experiments. *Icarus,* 208(2):735–748.

Delgado Mena, E., Tsantaki, M., Adibekyan, V. Zh., Sousa, S. G., Santos, N. C., González Hernández, J. I., and Israelian, G. (2017). Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program. II. Cu, Zn, Sr, Y, Zr, Ba, Ce, Nd, and Eu. *Astron. Astrophys.,* 606:A94.

de Saint-Exupéry, A. (1948). *Citadelle*. Gallimard, Paris, France.

de Sousa Mello, F. and Friaça, A. C. S. (2020). The end of life on Earth is not the end of the world: Converging to an estimate of life span of the biosphere? *Int. J. Astrobiol.,* 19(1):25–42.

de Souza, T. A. J. and Pereira, T. C. Caenorhabditis elegans Tolerates Hyperaccelerations up to 400, 000 × g. *Astrobiology,* 18(7):825–833.

de Vera, J.-P., Alawi, M., Backhaus, T., Baqué, M., Billi, D., Böttger, U., Berger, T., Bohmeier, M., Cockell, C., Demets, R., de la Torre Noetzel, R., Edwards, H., Elsaesser, A., Fagliarone, C., Fiedler, A., Foing, B., Foucher, F., Fritz, J., Hanke, F., . . . Zucconi, L. (2019). Limits of Life and the Habitability of Mars: The ESA Space Experiment BIOMEX on the ISS. *Astrobiology,* 19(2): 145–157.

de Waal, F. B. M. (2008). Putting the altruism Back into altruism: The evolution of empathy. *Annu. Rev. Psychol.,* 59:279–300.

de Waal, F. (2016). *Are We Smart Enough to Know How Smart Animals Are?* W. W. Norton & Co., New York, NY.

de Waal, F. (2019). *Mama's Last Hug: Animal Emotions and What They Tell Us about Ourselves*. W. W. Norton & Co., New York, NY.

Dell, A. I., Pawar, S., and Savage, V. M. (2011). Systematic variation in the temperature dependence of physiological and ecological traits. *Proc. Natl. Acad. Sci. USA,* 108(26):10591–10596.

DeLong, J. P., Okie, J. G., Moses, M. E., Sibly, R. M., and Brown, J. H. (2010). Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. *Proc. Natl. Acad. Sci. USA,* 107(29):12941–12945.

Demets, R. (2012). Darwin's Contribution to the Development of the Panspermia Theory. *Astrobiology,* 12(10):946–950.

Deming, L. D. and Seager, S. (2017). Illusion and reality in the atmospheres of exoplanets. *J. Geophys. Res. Planets,* 122(1):53–75.

Deming, D., Louie, D., and Sheets, H. (2019). How to Characterize the Atmosphere of a Transiting Exoplanet. *Publ. Astron. Soc. Pac.,* 131(995):013001.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton & Co., New York, NY.

Derex, M. and Mesoudi, A. (2020). Cumulative Cultural Evolution within Evolving Population Structures. *Trends Cogn. Sci.,* 24(8):654–667.

Des Marais, D. J., Harwit, M. O., Jucks, K. W., Kasting, J. F., Lin, D. N. C., Lunine, J. I., Schneider, J., Seager, S., Traub, W. A., and Woolf, N. J. (2002). Remote Sensing of Planetary Properties and Biosignatures on Extrasolar Terrestrial Planets. *Astrobiology,* 2(2):153–181.

Desiraju, G. D. and Steiner, T. (1999). *The Weak Hydrogen Bond: In Structural Chemistry and Biology*. Oxford University Press, Oxford, UK.

Diamond, J. M. (1975). The island dilemma: Lessons of modern biogeographic studies for the design of natural reserves. *Biol. Conserv.,* 7(2):129–146.

Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton & Co., New York, NY.

Dick, S. J. (1993). The Search for Extraterrestrial Intelligence and the NASA High Resolution Microwave Survey/HRMS—Historical Perspectives. *Space Sci. Rev.,* 64:93–139.

Dick, S. J. (1996). *The Biological Universe: The Twentieth Century Extraterrestrial Life Debate and the Limits of Science*. Cambridge University Press, Cambridge, UK.

Dick, S. J. (2003). Cultural evolution, the postbiological universe and SETI. *Int. J. Astrobiol.,* 2(1):65–74.

Dick, S. J. (2018). *Astrobiology, Discovery, and Societal Impact*. Cambridge University Press, Cambridge, UK.

Dicke, U. and Roth, G. (2016). Neuronal factors determining high intelligence. *Phil. Trans. R. Soc. B,* 371(1685):20150180.

Dieterich, S. B., Henry, T. J., Golimowski, D. A., Krist, J. E., and Tanner, A. M. (2012). The Solar Neighborhood. XXVIII. The Multiplicity Fraction of Nearby Stars from 5 to 70 AU and the Brown Dwarf Desert around M Dwarfs. *Astron. J.,* 144(2):64.

Dincer, I. and Rosen, M. A. (2013). *Exergy: Energy, Environment and Sustainable Development*. Elsevier, Amsterdam, Netherlands (2nd edition).

Di Stefano, R. and Ray, A. (2016). Globular Clusters as Cradles of Life and Advanced Civilizations. *Astrophys. J.,* 827(1):54.

Dittmann, J. A., Irwin, J. M., Charbonneau, D., Bonfils, X., Astudillo-Defru, N., Haywood, R. D., Berta-Thompson, Z. K., Newton, E. R., Rodriguez, J. E., Winters, J. G., Tan, T.-G., Almenara, J.-M., Bouchy, F., Delfosse, X., Forveille, T., Lovis, C., Murgas, F., Pepe, F., Santos, N. C., . . . Dressing, C. D. (2017). A temperate rocky super-Earth transiting a nearby cool star. *Nature,* 544(7650):333–336.

Djojodihardjo, H. (2018). Review of Solar Magnetic Sailing Configurations for Space Travel. *Adv. Astronaut. Sci. Technol.,* 1(2):207–219.

Djokic, T., van Kranendonk, M. J., Campbell, K. A., Walter, M. R., and Ward, C. R. (2017). Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits. *Nat. Commun,* 8:15263.

Do, A., Tucker, M. A., and Tonry, J. (2018). Interstellar Interlopers: Number Density and Origins of ‘Oumuamua-like Objects. *Astrophys. J. Lett.,* 855(1): L10.

Döbler, N. A. (2020). The concept of developmental relativity: Thoughts on the technological synchrony of interstellar civilizations. *Space Policy,* 54: 101391.

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., O'Neil, J., and Little, C. T. S. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature,* 543(7643):60–64.

Dohnanyi, J. S. (1969). Collisional Model of Asteroids and Their Debris. *J. Geophys. Res.,* 74(10):2531–2554.

Doig, A. J. (2017). Frozen, but no accident–why the 20 standard amino acids were selected. *FEBS J.,* 284:1296–1305.

Dokulil, M. T. (2019). Gross and Net Production in Different Environments. In Fath, B. D., editor, *Encyclopedia of Ecology,* vol. 2, pages 334–345. Elsevier, Amsterdam, Netherlands (2nd edition).

Dole, S. (1964). *Habitable planets for man*. Blaisdell Pub. Co., New York, NY.

Dole, S. H. (1964). *Habitable Planets for Man*. Blaisdell Publishing Company, New York, NY.

Domagal-Goldman, S. D., Meadows, V. S., Claire, M. W., and Kasting, J. F. (2011). Using Biogenic Sulfur Gases as Remotely Detectable Biosignatures on Anoxic Planets. *Astrobiology,* 11(5):419–441.

Donaldson, D. J., Tuck, A. F., and Vaida, V. (2001). Spontaneous fission of atmospheric aerosol particles. *Phys. Chem. Chem. Phys.,* 3(23):5270–5273.

Donau, C., Späth, F., Sosson, M., Kriebisch, B. A. K., Schnitter, F., Tena-Solsona, M., Kang, H.-S., Salibi, E., Sattler, M., Mutschler, H., and Boekhoven, J. (2020). Active coacervate droplets as a model for membraneless organelles and protocells. *Nat. Commun.,* 11:5167.

Dong, C., Huang, Z., Lingam, M., Tóth, G., Gombosi, T., and Bhattacharjee, A. (2017). The Dehydration of Water Worlds via Atmospheric Losses. *Astrophys. J. Lett.,* 847(1):L4.

Dong, C., Lingam, M., Ma, Y., and Cohen, O. (2017). Is Proxima Centauri b habitable? A study of atmospheric loss. *Astrophys. J. Lett.,* 837(2):L26.

Dong, C., Jin, M., Lingam, M., Airapetian, V. S., Ma, Y., and van der Holst, B. (2018). Atmospheric escape from the TRAPPIST-1 planets and implications for habitability. *Proc. Natl. Acad. Sci. USA,* 115(2):260–265.

Dong, C., Lee, Y., Ma, Y., Lingam, M., Bougher, S., Luhmann, J., Curry, S., Toth, G., Nagy, A., Tenishev, V., Fang, X., Mitchell, D., Brain, D., and Jakosky, B. (2018). Modeling Martian atmospheric losses over time: Implications for exoplanetary climate evolution and habitability. *Astrophys. J. Lett.,* 859(1):L14.

Dong, C., Huang, Z., and Lingam, M. (2019). Role of planetary obliquity in regulating atmospheric escape: G-dwarf versus M-dwarf Earth-like exoplanets. *Astrophys. J. Lett.,* 882(2):L16.

Dong, C., Jin, M., and Lingam, M. (2020). Atmospheric escape from TOI-700 d: Venus versus Earth analogs. *Astrophys. J. Lett.,* 896(2):L24.

Doughty, C. E. and Wolf, A. (2010). Detecting Tree-like Multicellular Life on Extrasolar Planets. *Astrobiology,* 10(9):869–879.

Doughty, C. E., Abraham, A., Windsor, J., Mommert, M., Gowenlock, M., Robinson, T., and Trilling, D. (2020). Distinguishing multicellular life on exoplanets by testing Earth as an exoplanet. *Int. J. Astrobiol.,* 19(6):492–499.

Drake, F. D. (1960). How Can We Detect Radio Transmissions from Distant Planetary Systems? *Sky Telesc.,* 39:140–143.

Drake, F. (1961). Project Ozma. *Physics Today,* 14(4):40–46.

Drake, F. D. (1965). The Radio Search for Intelligent Extraterrestrial Life. In Mamikunian, G. and Briggs, M. H., editors, *Current Aspects of Exobiology,* pages 323–345. Pergamon Press, Oxford, UK.

Drake, F. D. and Sagan, C. (1973). Interstellar Radio Communication and the Frequency Selection Problem. *Nature,* 245(5423):257–258.

Drake, J. J., Cohen, O., Yashiro, S., and Gopalswamy, N. (2013). Implications of Mass and Energy Loss due to Coronal Mass Ejections on Magnetically Active Stars. *Astrophys. J.,* 764(2):170.

Dressing, C. D. and Charbonneau, D. (2015). The occurrence of potentially habitable planets orbiting M dwarfs estimated from the full Kepler dataset and an empirical measurement of the detection sensitivity. *Astrophys. J.,* 807(1):45.

Drobot, B., Iglesias-Artola, J. M., Le Vay, K., Mayr, V., Kar, M., Kreysing, M., Mutschler, H., and Tang, T.-Y. D. (2018). Compartmentalised RNA catalysis in membrane-free coacervate protocells. *Nat. Commun.,* 9:3643.

Droser, M. L., Tarhan, L. G., and Gehling, J. G. (2017). The rise of animals in a changing environment: Global ecological innovation in the late ediacaran. *Annu. Rev. Earth Planet. Sci.,* 45:593–617.

Ducrot, E., Gillon, M., Delrez, L., Agol, E., Rimmer, P., Turbet, M., Günther, M. N., Demory, B-O., Triaud, A. H. M. J., Bolmont, E., Burgasser, A., Carey, S. J., Ingalls, J. G., Jehin, E., Leconte, J., Lederer, S. M., Queloz, D., Raymond, S. N., Selsis, F., Van Grootel, V., and de Wit, J. (2020). TRAPPIST-1: Global Results of the Spitzer Exploration Science Program *Red Worlds*. *Astron. Astrophys.,* 640:A112.

Dufour, D. L. and Piperata, B. A. (2018). Reflections on nutrition in biological anthropology. *Am. J. Phys. Anthropol.,* 165(4):855–864.

Dunbar, R. I. M. (2003). The social brain: Mind, language, and society in evolutionary perspective. *Annu. Rev. Anthropol.,* 32:163–181.

Dunbar, R. I. M. and Shultz, S. (2007). Evolution in the social brain. *Science,* 317(5843):1344–1347.

Dunbar, R. I. M. and Shultz, S. (2017). Why are there so many explanations for primate brain evolution? *Phil. Trans. R. Soc. B,* 372(1727):20160244.

Duval, S., Baymann, F., Schoepp-Cothenet, B., Trolard, F., Bourrié, G., Grauby, O., Branscomb, E., Russell, M. J., and Nitschke, W. (2019). Fougerite: The not so simple progenitor of the first cells. *Interface Focus,* 9(6):20190063.

Dyson, F. J. (1960). Search for Artificial Stellar Sources of Infrared Radiation. *Science,* 131(3414):1667–1668.

Dyson, F. J. (1968). Interstellar Transport. *Phys. Today,* 21(10):41–45.

Dyson, F. J. (1979a). *Disturbing the Universe*. Harper & Row, New York, NY.

Dyson, F. J. (1979b). Time without end: Physics and biology in an open universe. *Rev. Mod. Phys.,* 51(3):447–460.

Dyson, F. (1985). *Origins of life*. Cambridge University Press, Cambridge, UK.

Eastwood, J. P., Biffis, E., Hapgood, M. A., Green, L., Bisi, M. M., Bentley, R. D., Wicks, R., McKinnell, L. A., Gibbs, M., and Burnett, C. (2017). The Economic Impact of Space Weather: Where Do We Stand? *Risk Anal.,* 37(2):206–218.

Edmondson, W. H. and Stevens, I. R. (2003). The utilization of pulsars as SETI beacons. *Int. J. Astrobiol.,* 2(4):231–271.

Egan, H., Jarvinen, R., Ma, Y., and Brain, D. (2019). Planetary magnetic field control of ion escape from weakly magnetized planets. *Mon. Not. R. Astron. Soc.,* 488(2):2108–2120.

Eguchi, J., Seales, J., and Dasgupta, R. (2020). Great Oxidation and Lomagundi Events linked by deep cycling and enhanced degassing of carbon. *Nat. Geosci.,* 13(1):71–76.

Ehlmann, B. L. and Edwards, C. S. (2014). Mineralogy of the Martian Surface. *Annu. Rev. Earth Planet. Sci.,* 42:291–315.

Ehlmann, B. L., Anderson, F. S., Andrews-Hanna, J., Catling, D. C., Christensen, P. R., Cohen, B. A., Dressing, C. D., Edwards, C. S., Elkins-Tanton, L. T., Farley, K. A., Fassett, C. I., Fischer, W. W., Fraeman, A. A., Golombek, M. P., Hamilton, V. E., Hayes, A. G., Herd, C. D. K., Horgan, B., Hu, . . . Zahnle, K. J. (2016). The sustainability of habitability on terrestrial planets: Insights, questions, and needed measurements from Mars for understanding the evolution of Earth-like worlds. *J. Geophys. Res. Planets,* 121(10):1927–1961.

Ehrenreich, D., Lecavelier des Etangs, A., Beaulieu, J.-P., and Grasset, O. (2006). On the Possible Properties of Small and Cold Extrasolar Planets: Is OGLE 2005-BLG-390Lb Entirely Frozen? *Astrophys. J.,* 651(1):535–543.

Eigen, M. and Schuster, P. (1979). *The Hypercycle: A Principle of Natural Self-Organization*. Springer-Verlag, Berlin, Germany.

Eigenbrode, J. L., Summons, R. E., Steele, A., Freissinet, C., Millan, M., Navarro-González, R., Sutter, B., McAdam, A. C., Franz, H. B., Glavin, D. P., Archer, P. D., Mahaffy, P. R., Conrad, P. G., Hurowitz, J. A., Grotzinger, J. P., Gupta,

S., Ming, D. W., Sumner, D. Y., Szopa, C., . . . and Coll, P. (2018). Organic matter preserved in 3-billion-year-old mudstones at Gale crater, Mars. *Science,* 360(6393):1096–1101.

El Albani, A., Mangano, M. G., Buatois, L. A., Bengtson, S., Riboulleau, A., Bekker, A., Konhauser, K., Lyons, T., Rollion-Bard, C., Bankole, O., Baghekema, S. G. L., Meunier, A., Trentesaux, A., Mazurier, A., Aubineau, J., Laforest, C., Fontaine, C., Recourt, P., Chi Fru, E., . . . Canfield, D. E. (2019). Organism motility in an oxygenated shallow-marine environment 2.1 billion years ago. *Proc. Natl. Acad. Sci. USA,* 116(9):3431–3436.

Eldredge, N. and Gould, S. J. (1972). Punctuated equilibria: An alternative to phyletic gradualism. In Schopf, T. J. M., editor, *Models in Paleobiology,* pages 82–115. Freeman, Cooper & Co., San Francisco, CA.

Ellis, E. C. (2018). *Anthropocene: A Very Short Introduction*. Oxford University Press, Oxford, UK.

Elsila, J. E., Aponte, J. C., Blackmond, D. G., Burton, A. S., Dworkin, J. P., and Glavin, D. P. (2016). Meteoritic Amino Acids: Diversity in Compositions Reflects Parent Body Histories. *ACS Cent. Sci.,* 2(6):370–379.

Embley, T. M. and Martin, W. (2006). Eukaryotic evolution, changes and challenges. *Nature,* 440(7084):623–630.

Eme, L., Spang, A., Lombard, J., Stairs, C. W., and Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.,* 15:711–723.

Emery, N. J. and Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science,* 306(5703):1903–1907.

Encrenaz, T., DeWitt, C., Richter, M. J., Greathouse, T. K., Fouchet, T., Montmessin, F., Lefèvre, F., Bézard, B., Atreya, S. K., Aoki, S., and Sagawa, H. (2018). New measurements of D/H on Mars using EXES aboard SOFIA. *Astron. Astrophys.,* 612:A112.

Engelhart, A. E., Powner, M. W., and Szostak, J. W. (2013). Functional RNAs exhibit tolerance for non-heritable 2–5 versus 3–5 backbone heterogeneity. *Nat. Chem.,* 5(5):390–394.

Engels, F. (1934). *Dialectics of nature*. Translated by C. Dutt. Progress Publishers, Moscow, USSR.

England, J. L. (2013). Statistical physics of self-replication. *J. Chem. Phys.,* 139(12):121923.

Enriquez, J. E., Siemion, A., Foster, G., Gajjar, V., Hellbourg, G., Hickish, J., Isaacson, H., Price, D. C., Croft, S., DeBoer, D., Lebofsky, M., MacMahon, D. H. E., and Werthimer, D. (2017). The Breakthrough Listen Search for Intelligent Life: 1.1–1.9 GHz Observations of 692 Nearby Stars. *Astrophys. J.,* 849(2):104.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen,* 6:290–297.

Erkaev, N. V., Kulikov, Y. N., Lammer, H., Selsis, F., Langmayr, D., Jaritz, G. F., and Biernat, H. K. (2007). Roche lobe effects on the atmospheric loss from "Hot Jupiters." *Astron. Astrophys.,* 472(1):329–334.

Erwin, D. H., Laflamme, M., Tweedt, S. M., Sperling, E. A., Pisani, D., and Peterson, K. J. (2011). The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. *Science,* 334(6059):1091–1097.

Erwin, D. H. (2020). The origin of animal body plans: A view from fossil evidence and the regulatory genome. *Development,* 147(4):dev182899.

Eschenmoser, A. (2007). The search for the chemistry of life's origin. *Tetrahedron,* 63(52):12821–12844.

Eshleman, V. R. (1979). Gravitational lens of the sun—Its potential for observations and communications over interstellar distances. *Science,* 205(4411):1133–1135.

Estrela, S., Kerr, B., and Morris, J. J. (2016). Transitions in individuality through symbiosis. *Curr. Opin. Microbiol.,* 31:191–198.

Estrela, R. and Valio, A. (2018). Superflare Ultraviolet Impact on Kepler-96 System: A Glimpse of Habitability When the Ozone Layer First Formed on Earth. *Astrobiology,* 18(11):1414–1424.

Etiope, G. and Sherwood Lollar, B. (2013). Abiotic Methane on Earth. *Rev. Geophys.,* 51(2):276–299.

Etiope, G. and Whiticar, M. J. (2019). Abiotic methane in continental ultramafic rock systems: Towards a genetic model. *Appl. Geochem.,* 102:139–152.

Eubanks, T. M., Schneider, J., Hein, A. M., Hibberd, A., and Kennedy, R. (2020). Exobodies in Our Back Yard: Science from Missions to Nearby Interstellar Objects [White paper]. *arXiv e-prints,* arXiv:2007.12480.

Evans, K. L. and Gaston, K. J. (2005). Can the evolutionary-rates hypothesis explain species-energy relationships? *Funct. Ecol.,* 19(6):899–915.

Evans, V. (2014). *The Language Myth: Why Language Is Not an Instinct*. Cambridge University Press, Cambridge, UK.

Fabian, A. C. (1977). Signalling over stellar distances with X-rays. *J. Br. Interplanet. Soc.,* 30:112–113.

Fahrig, L. (2020). Why do several small patches hold more species than few large patches? *Glob. Ecol. Biogeogr.,* 29(4):615–628.

Fajardo-Cavazos, P., Link, L., Melosh, H. J., and Nicholson, W. L. (2005). Bacillus subtilis Spores on Artificial Meteorites Survive Hypervelocity Atmospheric Entry: Implications for Lithopanspermia. *Astrobiology,* 5(6):726–736.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science,* 320(5879):1034–1039.

Farmer, J. D., Kauffman, S. A., Packard, N. H. (1986). Autocatalytic replication of polymers. *Physica D,* 22(1):50–67.

Fauchez, T. J., Turbet, M., Villanueva, G. L., Wolf, E. T., Arney, G., Kopparapu, R. K., Lincowski, A., Mandell, A., de Wit, J., Pidhorodetska, D., Domagal-Goldman, S. D., and Stevenson, K. B. (2019). Impact of Clouds and Hazes on the Simulated *JWST* Transmission Spectra of Habitable Zone Planets in the TRAPPIST-1 System. *Astrophys. J.,* 887(2):194.

Fauchez, T. J., Villanueva, G. L., Schwieterman, E. W., Turbet, M., Arney, G., Pidhorodetska, D., Kopparapu, R. K., Mandell, A., and Domagal-Goldman, S. D. (2020). Sensitive probing of exoplanetary oxygen via mid-infrared collisional absorption. *Nat. Astron.,* 4:372–376.

Fei, H., Yamazaki, D., Sakurai, M., Miyajima, N., Ohfuji, H., Katsura, T., and Yamamoto, T. (2017). A nearly water-saturated mantle transition zone inferred from mineral viscosity. *Sci. Adv.,* 3(6):e1603024.

Feinberg, T. E. and Mallatt, J. M. (2018). *Consciousness Demystified*. The MIT Press, Cambridge, MA.

Feller, G. (2017). Cryosphere and psychrophiles: Insights into a cold origin of life? *Life,* 7(2):25.

Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., and Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Sci. Adv.,* 6(26):eaaz1002.

Ferris, J. P. (2006). Montmorillonite-catalysed formation of RNA oligomers: The possible role of catalysis in the origins of life. *Phil. Trans. R. Soc. B,* 361(1474): 1777–1786.

Ferus, M., Pietrucci, F., Saitta, A. M., Knížek, A., Kubelík, P., Ivanek, O., Shestivska, V., and Civiš, S. (2017). Formation of nucleobases in a Miller-Urey reducing atmosphere. *Proc. Natl. Acad. Sci. USA,* 114(17):4306–4311.

Fialho, D. M., Roche, T. P., and Hud, N. V. (2020). Prebiotic syntheses of noncanonical nucleosides and nucleotides. *Chem. Rev.,* 120(11):4806–4830.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science,* 281(5374):237–240.

Filippelli, G. M. (2008). The Global Phosphorus Cycle: Past, Present, and Future. *Elements,* 4(2):89–95.

Filippova, L. N. and Strelnitskij, V. S. (1988). Ecliptic as an Attractor for SETI. *Astronomicheskij Tsirkulyar,* 1531:31.

Fischer, H. (2010). *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, New York, NY.

Fischer, W. W., Hemp, J., and Johnson, J. E. (2016). Evolution of oxygenic photosynthesis. *Annu. Rev. Earth Planet. Sci.,* 44:647–683.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press, Cambridge, UK.

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.,* 24(1):3–33.

Fleischaker, G. R. (1990). Origins of life: An operational definition. *Orig. Life Evol. Biosph.,* 20(2):127–137.

Fogg, M. J. (1987). Temporal aspects of the interaction among the first galactic civilizations: The "interdict hypothesis". *Icarus,* 69(2):370–384.

Foley, B. J. and Driscoll, P. E. (2016). Whole planet coupling between climate, mantle, and core: Implications for rocky planet evolution. *Geochem. Geophys. Geosyst.,* 17(5):1885–1914.

Foley, B. J. and Smye, A. J. (2018). Carbon Cycling and Habitability of Earth-Sized Stagnant Lid Planets. *Astrobiology,* 18(7):873–896.

Foley, R. A. (2016). Mosaic evolution and the pattern of transitions in the hominin lineage. *Phil. Trans. R. Soc. B,* 371(1698):20150244.

Fontana, W. and Buss, L. W. (1994). What would be conserved if "the tape were played twice"? *Proc. Natl. Acad. Sci. USA,* 91(2):757–761.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., Gibbs, R. A., Hanson, M. B., Korneliussen, T. S., Martin, M. D., Robertson, K. M., Sousa, V. C., Vieira, F. G., Vinař, T., Wade, P., Worley, K. C., Excoffier, L., Morin, P. A., Gilbert, M. T. P., and Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.,* 7:11693.

Forbes, J. and Loeb, A. (2018). Evaporation of planetary atmospheres due to XUV illumination by quasars. *Mon. Not. R. Astron. Soc.,* 479(1):171–182.

Forbes, J. C. and Loeb, A. (2019). Turning up the Heat on 'Oumuamua. *Astrophys. J. Lett.,* 875(2):L23.

Ford, E. B., Seager, S., and Turner, E. L. (2001). Characterization of extrasolar terrestrial planets from diurnal photometric variability. *Nature,* 412(6850):885–887.

Forgan, D. H. and Elvis, M. (2011). Extrasolar asteroid mining as forensic evidence for extraterrestrial intelligence. *Int. J. Astrobiol.,* 10(4):307–313.

Forgan, D. H. (2013). On the Possibility of Detecting Class A Stellar Engines using Exoplanet Transit Curves. *J. Br. Interplanet. Soc.,* 66:144–154.

Forgan, D. H. (2014). Can Collimated Extraterrestrial Signals be Intercepted? *J. Br. Interplanet. Soc.,* 67:232–236.

Forgan, D. H. (2017). The Galactic Club or Galactic Cliques? Exploring the limits of interstellar hegemony and the Zoo Hypothesis. *Int. J. Astrobiol.,* 16(4):349–354.

Forgan, D. H. (2019). *Solving Fermi's Paradox*. Cambridge University Press, Cambridge, UK.

Forterre, P. (2016). To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. *Stud. Hist. Philos. Biol. Biomed. Sci.,* 59:100–108.

Forterre, P. and Gaïa, M. (2016). Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.,* 31:44–49.

Forward, R. L. (1982). Antimatter Propulsion. *J. Br. Interplanet. Soc.,* 35:391–395.

Forward, R. L. (1984). Roundtrip interstellar travel using laser-pushed lightsails. *J. Spacecraft Rockets,* 21(2):187–195.

Forward, R. L. (1985). Starwisp: An Ultra-Light Interstellar Probe. *J. Spacecraft Rockets,* 22(3):345–350.

Foster, G. V. (1972). Non-human artifacts in the solar system. *Spaceflight,* 14:447–453.

Foucher, F., Westall, F., Brandstätter, F., Demets, R., Parnell, J., Cockell, C. S., Edwards, H. G. M., Bény, J.-M., and Brack, A. (2010). Testing the survival of microfossils in artificial martian sedimentary meteorites during entry into Earth's atmosphere: The STONE 6 experiment. *Icarus,* 207(2):616–630.

France, K., Loyd, R. O. P., Youngblood, A., Brown, A., Schneider, P. C., Hawley, S. L., Froning, C. S., Linsky, J. L., Roberge, A., Buccino, A. P., Davenport, J. R. A., Fontenla, J. M., Kaltenegger, L., Kowalski, A. F., Mauas, P. J. D., Miguel, Y., Redfield, S., Rugheimer, S., Tian, . . . Weisenburger, K. L. (2016). The MUSCLES Treasury Survey. I. Motivation and overview. *Astrophys. J.,* 820(2):89.

Frank, A. and Sullivan, W. T. (2016). A New Empirical Constraint on the Prevalence of Technological Species in the Universe. *Astrobiology,* 16(5):359–362.

Frank, A., Kleidon, A., and Alberti, M. (2017). Earth as a Hybrid Planet: The Anthropocene in an Evolutionary Astrobiological Context. *Anthropocene,* 19:13–21.

Frank, F. C. (1953). On spontaneous asymmetric synthesis. *Biochim. Biophys. Acta,* 11:459–463.

Frebel, A. (2010). Stellar archaeology: Exploring the Universe with metal-poor stars. *Astron. Nachr.,* 331(5):474–488.

Freeland, R. M. (2015). Mathematics of Magsails. *J. Br. Interplanet. Soc.,* 68:306–323.

Freeman, J. and Lampton, M. (1975). Interstellar Archaeology and the Prevalence of Intelligence. *Icarus,* 25(2):368–369.

Freidberg, J. P. (2008). *Plasma Physics and Fusion Energy*. Cambridge University Press, Cambridge, UK.

Freitas, R. A. (1980). A self-reproducing interstellar probe. *J. Br. Interplanet. Soc.,* 33(7):251–264.

Freitas, R. A. and Valdes, F. (1980). A search for natural or artificial objects located at the earth-moon libration points. *Icarus,* 42:442–447.

Freitas, R. A. and Valdes, F. (1985). The search for extraterrestrial artifacts (SETA). *Acta Astronaut.,* 12(12):1027–1034.

French, K. L., Hallmann, C., Hope, J. M., Schoon, P. L., Zumberge, J. A., Hoshino, Y., Peters, C. A., George, S. C., Love, G. D., Brocks, J. J., Buick, R., and Summons, R. E. (2015). Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proc. Natl. Acad. Sci. USA,* 112(19):5915–5920.

Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., and Leman, L. J. (2020). Prebiotic peptides: Molecular hubs in the origin of life. *Chem. Rev.,* 120(11):4707–4765.

Frisbee, R. H. (2003). Advanced Space Propulsion for the 21st Century. *J. Propuls. Power,* 19(6):1129–1154.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.,* 11(2):127–138.

Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., and van Schaik, C. P. (2019). Multimodal communication and language origins: Integrating gestures and vocalizations. *Biol. Rev.,* 94(5):1809–1829.

Fry, I. (2000). *The Emergence of Life on Earth: A Historical and Scientific Overview*. Rutgers University Press, New Brunswick, NJ.

Fry, I. (2011). The role of natural selection in the origin of life. *Orig. Life Evol. Biosph.,* 41(1):3–16.

Fu, R., O'Connell, R. J., and Sasselov, D. D. (2010). The Interior Dynamics of Water Planets. *Astrophys. J.,* 708(2):1326–1334.

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: Insights into the early evolution of life? *Annu. Rev. Microbiol.,* 65:631–658.

Fujii, Y., Kawahara, H., Suto, Y., Taruya, A., Fukuda, S., Nakajima, T., and Turner, E. L. (2010). Colors of a Second Earth: Estimating the Fractional Areas of Ocean, Land, and Vegetation of Earth-like Exoplanets. *Astrophys. J.,* 715(2):866–880.

Fujii, Y., Angerhausen, D., Deitrick, R., Domagal-Goldman, S., Grenfell, J. L., Hori, Y., Kane, S. R., Pallé, E., Rauer, H., Siegler, N., Stapelfeldt, K., and Stevenson, K. B. (2018). Exoplanet Biosignatures: Observational Prospects. *Astrobiology,* 18(6):739–778.

Furukawa, Y., Chikaraishi, Y., Ohkouchi, N., Ogawa, N. O., Glavin, D. P., Dworkin, J. P., Abe, C., and Nakamura, T. (2019). Extraterrestrial ribose and other sugars in primitive meteorites. *Proc. Natl. Acad. Sci. USA,* 116(49): 24440–24445.

Futuyma, D. J. (2017). Evolutionary biology today and the call for an extended synthesis. *Interface Focus,* 7(5):20160145.

Füzfa, A., Dhelonga-Biarufu, W., and Welcomme, O. (2020). Sailing Towards the Stars Close to the Speed of Light. *Phys. Rev. Res.,* 2(4):043186.

Gaidos, E. J., Nealson, K. H., and Kirschvink, J. L. (1999). Life in Ice-Covered Oceans. *Science,* 284(5420):1631–1633.

Gaidos, E. and Williams, D. M. (2004). Seasonality on terrestrial extrasolar planets: inferring obliquity and surface conditions from infrared light curves. *New Astron.,* 10(1):67–77.

Gaidos, E. (2017). Transit detection of a 'starshade' at the inner lagrange point of an exoplanet. *Mon. Not. R. Astron. Soc.,* 469(4):4455–4464.

Galera, E., Galanti, G. R., and Kinouchi, O. (2019). Invasion percolation solves Fermi Paradox but challenges SETI projects. *Int. J. Astrobiol.,* 18(4):316–322.

Galperin, M. Y. (2013). Genome Diversity of Spore-Forming *Firmicutes*. *Microbiol. Spectr.,* 1(2):TBS-0015–2012.

Galway-Witham, J., Cole, J., and Stringer, C. (2019). Aspects of human physical and behavioural evolution during the last 1 million years. *J. Quat. Sci.,* 34(6):355–378.

Gánti, T. (2003). *The Principles of Life*. Oxford University Press, Oxford, UK.

Garcia-Escartin, J. C. and Chamorro-Posada, P. (2013). Scouting the spectrum for interstellar travellers. *Acta Astronaut.,* 85:12–18.

Garcia-Sage, K., Glocer, A., Drake, J. J., Gronoff, G., and Cohen, O. (2017). On the magnetic protection of the atmosphere of Proxima Centauri b. *Astrophys. J. Lett.,* 844(1):L13.

Garraffo, C., Drake, J. J., and Cohen, O. (2016). The space weather of Proxima Centauri b. *Astrophys. J. Lett.,* 833(1):L4.

Garrett, M. A. (2015). Application of the mid-IR radio correlation to the $\hat{G}$ sample and the search for advanced extraterrestrial civilisations. *Astron. Astrophys.,* 581:L5.

Garrison, W. M., Morrison, D. C., Hamilton, J. G., Benson, A. A., and Calvin, M. (1951). Reduction of carbon dioxide in aqueous solutions by ionizing radiation. *Science,* 114(2964):416–418.

Gaspard, P. (2016). Kinetics and thermodynamics of living copolymerization processes. *Phil. Trans. R. Soc. A,* 374(2080):20160147.

Gertz, J. (2016). ET Probes: Looking Here as Well as There. *J. Br. Interplanet. Soc.,* 69:88–91.

Gertz, J. (2019). There's No Place Like Home (in Our Own Solar System): Searching for ET Near White Dwarfs. *J. Br. Interplanet. Soc.,* 72(11):386–395.

Gibard, C., Bhowmik, S., Karki, M., Kim, E.-K., and Krishnamurthy, R. (2018). Phosphorylation, oligomerization and self-assembly in water under potential prebiotic conditions. *Nat. Chem.,* 10:212–217.

Gibson, T. M., Shih, P. M., Cumming, V. M., Fischer, W. W., Crockford, P. W., Hodgskiss, M. S. W., Wörndle, S., Creaser, R. A., Rainbird, R. H., Skulski, T. M., and Halverson, G. P. (2018). Precise age of Bangiomorpha pubescens dates the origin of eukaryotic photosynthesis. *Geology,* 46(2):135–138.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature,* 319(6055):618.

Gillon, M. (2014). A novel SETI strategy targeting the solar focal regions of the most nearby stars. *Acta Astronaut.,* 94(2):629–633.

Gillon, M., Triaud, A. H. M. J., Demory, B.-O., Jehin, E., Agol, E., Deck, K. M., Lederer, S. M., de Wit, J., Burdanov, A., Ingalls, J. G., Bolmont, E., Leconte, J., Raymond, S. N., Selsis, F., Turbet, M., Barkaoui, K., Burgasser, A., Burleigh, M. R., Carey, S. J., . . . Queloz, D. (2017). Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1. *Nature,* 542(7642):456–460.

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M. and Charnov, E. L. (2001). Effects of Size and Temperature on Metabolic Rate. *Science,* 293(5538):2248–2251.

Gillooly, J. F., Allen, A. P., West, G. B., and Brown, J. H. (2006). The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. USA,* 102(1):140–145.

Gilster, P. (2004). *Centauri Dreams: Imagining and Planning Interstellar Exploration*. Copernicus Books, New York, NY.

Gindilis, L. M. and Gurvits, L. I. (2019). SETI in Russia, USSR and the post-Soviet space: A century of research. *Acta Astronaut.,* 162:1–13.

Ginsburg, I., Loeb, A., and Wegner, G. A. (2012). Hypervelocity planets and transits around hypervelocity stars. *Mon. Not. R. Astron. Soc. Lett.,* 423(1):948–954.

Ginsburg, I., Lingam, M., and Loeb, A. (2018). Galactic Panspermia. *Astrophys. J. Lett.,* 868(1):L12.

Ginsburg, S. and Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. The MIT Press, Cambridge, MA.

Giovannetti, V., García-Patrón, R., Cerf, N. J., and Holevo, A. S. (2014). Ultimate classical communication rates of quantum optical channels. *Nat. Photonics,* 8:796–800.

Giuranna, M., Viscardy, S., Daerden, F., Neary, L., Etiope, G., Oehler, D., Formisano, V., Aronica, A., Wolkenberg, P., Aoki, S., Cardesín-Moinelo, A.,

de la Parra, J. M.-Y., Merritt, D., and Amoroso, M. (2019). Independent confirmation of a methane spike on Mars and a source region east of Gale Crater. *Nat. Geosci.,* 12:326–332.

Gladman, B. J., Burns, J. A., Duncan, M., Lee, P., and Levison, H. F. (1996). The Exchange of Impact Ejecta Between Terrestrial Planets. *Science,* 271(5254):1387–1392.

Gladman, B., Dones, L., Levison, H. F., and Burns, J. A. (2005). Impact Seeding and Reseeding in the Inner Solar System. *Astrobiology,* 5(4):483–496.

Glaser, D. M., Hartnett, H. E., Desch, S. J., Unterborn, C. T., Anbar, A., Buessecker, S., Fisher, T., Glaser, S., Kane, S. R., Lisse, C. M., Millsaps, C., Neuer, S., O'Rourke, J. G., Santos, N., Walker, S. I., and Zolotov, M. (2020). Detectability of Life Using Oxygen on Pelagic Planets and Water Worlds. *Astrophys. J.,* 893(2):163.

Glavin, D. P., Burton, A. S., Elsila, J. E., Aponte, J. C., and Dworkin, J. P. (2020). The search for chiral asymmetry as a potential biosignature in our Solar system. *Chem. Rev.,* 120(11):4660–4689.

Glazier, D. S. (2015). Is metabolic rate a universal 'pacemaker' for biological processes? *Biol. Rev.,* 90(2):377–407.

Glazier, A. L., Howard, W. S., Corbett, H., Law, N. M., Ratzloff, J. K., Fors, O., and del Ser, D. (2020). Evryscope and K2 Constraints on TRAPPIST-1 Superflare Occurrence and Planetary Habitability. *Astrophys. J.,* 900(1):27.

Glein, C. R., Baross, J. A., and Waite, J. H. (2015). The pH of Enceladus' ocean. *Geochim. Cosmochim. Acta,* 162:202–219.

Glein, C. R., Postberg, F., and Vance, S. D. (2018). The Geochemistry of Enceladus: Composition and Controls. In Schenk, P. M., Clark, R. N., Howett, C. J. A., Verbiscer, A. J., and Waite, J. H., editors, *Enceladus and the Icy Moons of Saturn,* pages 39–56. The University of Arizona Press, Tucson, AZ.

Glein, C. R. and Waite, J. H. (2020). The Carbonate Geochemistry of Enceladus' Ocean. *Geophys. Res. Lett.,* 47(3):e2019GL08588.

Gleiser, M. and Walker, S. I. (2012). Life's chirality from prebiotic environments. *Int. J. Astrobiol.,* 11(4):287–296.

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux, New York, NY.

Godolt, M., Tosi, N., Stracke, B., Grenfell, J. L., Ruedas, T., Spohn, T., and Rauer, H. (2019). The habitability of stagnant-lid Earths around dwarf stars. *Astron. Astrophys.,* 625:A12.

Goebel, D. M. and Katz, I. (2008). *Fundamentals of Electric Propulsion: Ion and Hall Thrusters*. John Wiley & Sons, Inc., Hoboken, NJ.

Goksøyr, J. (1967). Evolution of eucaryotic cells. *Nature,* 214(5093):1161.

Gold, T. (1992). The Deep, Hot Biosphere. *Proc. Natl. Acad. Sci. USA,* 89(13): 6045–6049.

Goldblatt, C. and Watson, A. J. (2012). The runaway greenhouse: implications for future climate change, geoengineering and planetary atmospheres. *Phil. Trans. R. Soc. A,* 370:4197–4216.

Goldford, J. E., Hartman, H., Smith, T. F., and Segrè, D. (2017). Remnants of an ancient metabolism without phosphate. *Cell,* 168(6):1126–1134.

Goldford, J. E., Hartman, H., Marsland, R., and Segrè, D. (2019). Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. *Nat. Ecol. Evol.,* 3(12):1715–1724.

Goldreich, O., Juba, B., and Sudan, M. (2014). A theory of goal-oriented communication. *J. ACM,* 59(2):8.

Gombosi, T. I. (1998). *Physics of the Space Environment*. Cambridge University Press, Cambridge, UK.

Gómez-Consarnau, L., Raven, J. A., Levine, N. M., Cutter, L. S., Wang, D., Seegers, B., Arístegui, J., Fuhrman, J. A., Gasol, J. M., and Sañudo-Wilhelmy, S. A. (2019). Microbial rhodopsins are major contributors to the solar energy captured in the sea. *Sci. Adv.,* 5(8):eaaw8855.

Gong, S. and Macdonald, M. (2019). Review on solar sail technology. *Astrodyn.,* 3:93–125.

Gonzalez, G. (2005). Habitable zones in the Universe. *Orig. Life Evol. Biosph.,* 35(6):555–606.

González-Forero, M. and Gardner, A. (2018). Inference of ecological and social drivers of human brain-size evolution. *Nature,* 557(7706):554–557.

Goodenough, U. and Heitman, J. (2014). Origins of eukaryotic sexual reproduction. *Cold Spring Harb. Perspect. Biol.,* 6(3):a016154.

Gould, A. and Loeb, A. (1992). Discovering planetary systems through gravitational microlenses. *Astrophys. J.,* 396(1):104–114.

Gould, S. B., Garg, S. G., and Martin, W. F. (2016). Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. *Trends Microbiol.,* 24(7):525–534.

Gould, S. J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton & Co., New York, NY.

Goulinski, N. and Ribak, E. N. (2018). Capture of free-floating planets by planetary systems. *Mon. Not. R. Astron. Soc.,* 473(2):1589–1595.

Gowlett, J. A. J. (2016). The discovery of fire by humans: A long and convoluted process. *Phil. Trans. R. Soc. B,* 371(1696):20150164.

Gray, J. S. (1997). Marine biodiversity: patterns, threats and conservation needs. *Biodivers. Conserv.,* 6(1):153–175.

Gray, R. H. and Ellingsen, S. (2002). A Search for Periodic Emissions at the Wow Locale. *Astrophys. J.,* 578(2):967–971.

Gray, R. H. (2015). The Fermi Paradox Is Neither Fermi's Nor a Paradox. *Astrobiology,* 15(3):195–199.

Gray, R. H. and Mooley, K. (2017). A VLA Search for Radio Signals from M31 and M33. *Astron. J.,* 153(3):110.

Gray, R. H. (2020). The Extended Kardashev Scale. *Astron. J.,* 159(5):228.

Grazier, K. R. (2016). Jupiter: Cosmic Jekyll and Hyde. *Astrobiology,* 16(1): 23–38.

Greaves, J. S., Richards, A. M. S., Bains, W., Rimmer, P. B., Sagawa, H., Clements, D. L., Seager, S., Petkowski, J. J., Sousa-Silva, C., Ranjan, S., Drabek-Maunder, E., Fraser, H. J., Cartwright, A., Mueller-Wodarg, I., Zhan, Z., Friberg, P., Coulson, I., Lee, E., and Hoge, J. (2020). Phosphine gas in the cloud decks of Venus. *Nat. Astron.,* DOI: 10.1038/s41550-020-1174-4.

Greenberg, R. (2010). Transport Rates of Radiolytic Substances into Europa's Ocean: Implications for the Potential Origin and Maintenance of Life. *Astrobiology,* 10(3):275–283.

Grenfell, J. L. (2017). A review of exoplanetary biosignatures. *Phys. Rep.,* 713:1–17.

Grenfell, J. L., Leconte, J., Forget, F., Godolt, M., Carrión-González, Ó., Noack, L., Tian, F., Rauer, H., Gaillard, F., Bolmont, E., Charnay, B., and Turbet, M. (2020). Possible Atmospheric Diversity of Low Mass Exoplanets - Some Central Aspects. *Space Sci. Rev.,* 216:98.

Grießmeier, J. M. (2015). Detection Methods and Relevance of Exoplanetary Magnetic Fields. In Lammer, H. and Khodachenko, M., editors, *Characterizing Stellar and Exoplanetary Environments,* pages 213–237. Springer, Cham, Switzerland.

Grießmeier, J.-M., Tabataba-Vakili, F., Stadelmann, A., Grenfell, J. L., and Atri, D. (2016). Galactic cosmic rays on extrasolar Earth-like planets. II. Atmospheric implications. *Astron. Astrophys.,* 587:A159.

Griffin, D. R. (2001). *Animal Minds: Beyond Cognition to Consciousness*. The University of Chicago Press, Chicago, IL.

Griffith, E. C. and Vaida, V. (2012). In situ observation of peptide bond formation at the water-air interface. *Proc. Natl. Acad. Sci. USA,* 109(39):15697–15701.

Griffith, E. C., Tuck, A. F., and Vaida, V. (2012). Ocean-atmosphere interactions in the emergence of complexity in simple chemical systems. *Acc. Chem. Res.,* 45(12):2106–2113.

Griffith, R. L., Wright, J. T., Maldonado, J., Povich, M. S., Sigurdsson, S., and Mullan, B. (2015). The Ĝ Infrared Search for Extraterrestrial Civilizations with Large Energy Supplies. III. The Reddest Extended Sources in WISE. *Astrophys. J. Suppl.,* 217(2):25.

Grimaldi, C. and Marcy, G. W. (2018). Bayesian approach to SETI. *Proc. Natl. Acad. Sci. USA,* 115(42):E9755–E9764.

Grimm, S. L., Demory, B.-O., Gillon, M., Dorn, C., Agol, E., Burdanov, A., Delrez, L., Sestovic, M., Triaud, A. H. M. J., Turbet, M., Bolmont, É., Caldas,

A., de Wit, J., Jehin, E., Leconte, J., Raymond, S. N., Van Grootel, V., Burgasser, A. J., Carey, S., . . . Queloz, D. (2018). The nature of the TRAPPIST-1 exoplanets. *Astron. Astrophys.,* 613:A68.

Gronoff, G., Arras, P., Baraka, S., Bell, J. M., Cessateur, G., Cohen, O., Curry, S. M., Drake, J. J., Elrod, M., Erwin, J., Garcia-Sage, K., Garraffo, C., Glocer, A., Heavens, N. G., Lovato, K., Maggiolo, R., Parkinson, C. D., Wedlund, C. S., Weimer, D. R., and Moore, W. B. (2020). Atmospheric escape processes and planetary atmospheric evolution. *J. Geophys. Res. Space Phys.,* 125(8):e2019JA027639.

Gros, C. (2005). Expanding Advanced Civilizations in the Universe. *J. Br. Interplanet. Soc.,* 58:108–110.

Gros, C. (2017). Universal scaling relation for magnetic sails: momentum braking in the limit of dilute interstellar media. *J. Phys. Commun.,* 1(4):045007.

Grosberg, R. K. and Strathmann, R. R. (2007). The evolution of multicellularity: A minor major transition? *Annu. Rev. Ecol. Evol. Syst.,* 38:621–654.

Guillochon, J. and Loeb, A. (2015a). SETI via Leakage from Light Sails in Exoplanetary Systems. *Astrophys. J. Lett.,* 811(2):L20.

Guillochon, J. and Loeb, A. (2015b). The Fastest Unbound Stars in the Universe. *Astrophys. J.,* 806(1):124.

Gulick, A. (1955). Phosphorus as a factor in the origin of life. *Am. Sci.,* 43(3):479–489.

Gumsley, A. P., Chamberlain, K. R., Bleeker, W., Söderlund, U., de Kock, M. O., Larsson, E. R., and Bekker, A. (2017). Timing and tempo of the Great Oxidation Event. *Proc. Natl. Acad. Sci. USA,* 114(8):1811–1816.

Gunell, H., Maggiolo, R., Nilsson, H., Wieser, G. S., Slapak, R., Lindkvist, J., Hamrin, M., and De Keyser, J. (2018). Why an intrinsic magnetic field does not protect a planet against atmospheric escape. *Astron. Astrophys.,* 614: L3.

Günther, M. N., Zhan, Z., Seager, S., Rimmer, P. B., Ranjan, S., Stassun, K. G., Oelkers, R. J., Daylan, T., Newton, E., Kristiansen, M. H., Olah, K., Gillen, E., Rappaport, S., Ricker, G. R., Vanderspek, R. K., Latham, D. W., Winn, J. N., Jenkins, J. M., Glidden, . . . Ting, E. B. (2020). Stellar Flares from the First TESS Data Release: Exploring a New Sample of M Dwarfs. *Astron. J.,* 159(2):60.

Güntürkün, O., Ströckens, F., Scarf, D., and Colombo, M. (2017). Apes, feathered apes, and pigeons: Differences and similarities. *Curr. Opin. Behav. Sci.,* 16: 35–40.

Gupta, M., Prasad, N. G., Dey, S., Joshi, A., and Vidya, T. N. C. (2017). Niche construction in evolutionary theory: the construction of an academic niche? *J. Genet.,* 96(3):491–504.

Guyon, O. (2018). Extreme Adaptive Optics. *Annu. Rev. Astron. Astrophys.,* 56:315–355.

Guzik, P., Drahus, M., Rusek, K., Waniak, W., Cannizzaro, G., and Pastor-Marazuela, I. (2020). Initial characterization of interstellar comet 2I/Borisov. *Nat. Astron.,* 4:53–57.

Hadariová, L., Vesteg, M., Hampl, V., and Krajčovič, J. (2018). Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. *Curr. Genet.,* 64(2):365–387.

Haddock, S. H. D., Moline, M. A., and Case, J. F. (2010). Bioluminescence in the Sea. *Annu. Rev. Mar. Sci.,* 2:443–493.

Hadley, N. F. and Szarek, S. R. (1981). Productivity of Desert Ecosystems. *BioScience,* 31(10):747–753.

Halevy, I. and Bachan, A. (2017). The geologic history of seawater pH. *Science,* 355(6329):1069–1071.

Hampl, V., Čepička, I., and Eliáš, M. (2019). Was the mitochondrion necessary to start eukaryogenesis? *Trends Microbiol.,* 27(2):96–104.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comp. Cogn. Behav. Rev.,* 4: 17–28.

Hand, K. P., Carlson, R. W., and Chyba, C. F. (2007). Energy, Chemical Disequilibrium, and Geological Constraints on Europa. *Astrobiology,* 7(6):1006–1022.

Hands, T. O. and Dehnen, W. (2020). Capture of interstellar objects: a source of long-period comets. *Mon. Not. R. Astron. Soc. Lett.,* 493(1):L59–L64.

Hanlon, R. T. and Messenger, J. B. (2018). *Cephalopod Behaviour.* Cambridge University Press, Cambridge, UK (2nd edition).

Hanna, D. S., Ball, J., Covault, C. E., Carson, J. E., Driscoll, D. D., Fortin, P., Gingrich, D. M., Jarvis, A., Kildea, J., Lindner, T., Mueller, C., Mukherjee, R., Ong, R. A., Ragan, K., Williams, D. A., and Zweerink, J. (2009). OSETI with STACEE: A Search for Nanosecond Optical Transients from Nearby Stars. *Astrobiology,* 9(4):345–357.

Hansen, C. J., Andersen, A. C., and Christlieb, N. (2014). Stellar abundances and presolar grains trace the nucleosynthetic origin of molybdenum and ruthenium. *Astron. Astrophys.,* 568:A47.

Hao, J., Giovenco, E., Pedreira-Segade, U., Montagnac, G. and Daniel, I. (2018). Compatibility of amino acids in ice Ih: Implications for the origin of life. *Astrobiology,* 18(4):381–392.

Hao, J., Knoll, A. H., Huang, F., Hazen, R. M., and Daniel, I. (2020a). Cycling phosphorus on the Archean Earth: Part I. Continental weathering and riverine transport of phosphorus. *Geochim. Cosmochim. Acta,* 273:70–84.

Hao, J., Knoll, A. H., Huang, F., Hazen, R. M., and Daniel, I. (2020b). Cycling phosphorus on the Archean Earth: Part II. Phosphorus limitation on primary production in Archean ecosystems. *Geochim. Cosmochim. Acta,* 280:360–377.

Haqq-Misra, J. D. and Baum, S. D. (2009). The Sustainability Solution To The Fermi Paradox. *J. Br. Interplanet. Soc.,* 62:47–51.

Haqq-Misra, J. and Kopparapu, R. K. (2012). On the likelihood of non-terrestrial artifacts in the Solar System. *Acta Astronaut.,* 72:15–20.

Haqq-Misra, J., Kopparapu, R. K., and Schwieterman, E. (2020). Observational Constraints on the Great Filter. *Astrobiology,* 20(5):572–579.

Hare, B. (2017). Survival of the friendliest: *Homo sapiens* evolved via selection for prosociality. *Annu. Rev. Psychol.,* 68:155–186.

Harmand, S., Lewis, J. E., Feibel, C. S., Lepre, C. J., Prat, S., Lenoble, A., Boës, X., Quinn, R. L., Brenet, M., Arroyo, A., Taylor, N., Clément, S., Daver, G., Brugal, J.-P., Leakey, L., Mortlock, R. A., Wright, J. D., Lokorodi, S., Kirwa, . . . Roche, H. (2015). 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya. *Nature,* 521(7552):310–315.

Harness, A., Shaklan, S., Kasdin, N. J., Galvin, M., Willems, P., Balasubramanian, K., White, V., Yee, K., Muller, R., Dumont, P., and Vuong, S. (2019). Demonstration of 1e-10 contrast at the inner working angle of a starshade in broadband light and at a flight-like Fresnel number. In *Techniques and Instrumentation for Detection of Exoplanets IX,* volume 11117 of *Proc. SPIE,* page 111170L.

Harp, G. R., Richards, J., Shostak, S., Tarter, J. C., Vakoch, D. A., and Munson, C. (2016a). Radio SETI Observations of the Anomalous Star KIC 8462852. *Astrophys. J.,* 825(2):155.

Harp, G. R., Richards, J., Tarter, J. C., Dreher, J., Jordan, J., Shostak, S., Smolek, K., Kilsdonk, T., Wilcox, B. R., Wimberly, M. K. R., Ross, J., Barott, W. C., Ackermann, R. F., and Blair, S. (2016b). SETI Observations of Exoplanets with the Allen Telescope Array. *Astron. J.,* 152(6):181.

Harp, G. R., Gray, R. H., Richards, J., Shostak, G. S., and Tarter, J. C. (2020). An ATA search for a repetition of the Wow signal. *Astron. J.,* 160(4):162.

Harris, M. J. (1986). On the detectability of antimatter propulsion spacecraft. *Astrophys. Space Sci.,* 123(2):297–303.

Harris, M. J. (2002). Limits from CGRO-EGRET Data on the use of Antimatter as a Power Source by Extraterrestrial Civilizations. *J. Br. Interplanet. Soc.,* 55:383–393.

Harrison, J. P., Gheeraert, N. Tsigelnitskiy, D., and Cockell, C. S. (2013). The limits for life under multiple extremes. *Trends Microbiol.,* 21(4):204–212.

Harrison, S. A. and Lane, N. (2018). Life as a guide to prebiotic nucleotide synthesis. *Nat. Commun.,* 9:5176.

Harrison, T. M. (2020). *Hadean Earth*. Springer, Cham, Switzerland.

Hart, M. H. (1975). Explanation for the Absence of Extraterrestrials on Earth. *Q. Jl. R. Astr. Soc.,* 16:128–135.

Hartman, H. (1975). Speculations on the origin and evolution of metabolism. *J. Mol. Evol.,* 4(4):359–370.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science,* 298(5598):1569–1579.

Hawkesworth, C. J., Cawood, P. A., Dhuime, B., and Kemp, T. I. S. (2017). Earth's continental lithosphere through time. *Annu. Rev. Earth Planet. Sci.,* 45:169–198.

Hawkesworth, C., Cawood, P. A., and Dhuime, B. (2019). Rates of generation and growth of the continental crust. *Geosci. Front.,* 10(1):165–173.

Hayes, A. G. (2016). The Lakes and Seas of Titan. *Annu. Rev. Earth Planet. Sci.,* 44:57–83.

Hazen, R. M., Griffin, P. L., Carothers, J. M., and Szostak, J. W. (2007). Functional information and the emergence of biocomplexity. *Proc. Natl. Acad. Sci. USA,* 104(Suppl. 1):8574–8581.

Hazen, R. M. and Ferry, J. M. (2010). Mineral evolution: Mineralogy in the fourth dimension. *Elements,* 6(1):9–12.

Hazen, R. M. (2017). Chance, necessity and the origins of life: a physical sciences perspective. *Phil. Trans. R. Soc. A,* 375(2109):20160353.

Hebsgaard, M. B., Phillips, M. J., and Willerslev, E. (2005). Geologically ancient DNA: fact or artefact? *Trends Microbiol.,* 13(5):212–220.

Hegde, S., Paulino-Lima, I. G., Kent, R., Kaltenegger, L., and Rothschild, L. (2015). Surface biosignatures of exo-Earths: Remote detection of extraterrestrial life. *Proc. Natl. Acad. Sci. USA,* 112(13):3886–3891.

Heggie, D. C. (1975). Binary evolution in stellar dynamics. *Mon. Not. R. Astron. Soc.,* 173(3):729–787.

Hein, A. M., Perakis, N., Eubanks, T. M., Hibberd, A., Crowl, A., Hayward, K., Kennedy, R. G., and Osborne, R. (2019). Project Lyra: Sending a spacecraft to 1I/'Oumuamua (former A/2017 U1), the interstellar asteroid. *Acta Astronaut.,* 161:552–561.

Hein, A. M., Lingam, M., Eubanks, T. M., Hibberd, A., Fries, D., and Blase, W. P. (2020). A precursor balloon mission for Venusian astrobiology. *Astrophys. J. Lett.,* 903(2):L36.

Heller, R. and Armstrong, J. (2014). Superhabitable Worlds. *Astrobiology,* 14(1): 50–66.

Heller, R., Williams, D., Kipping, D., Limbach, M. A., Turner, E., Greenberg, R., Sasaki, T., Bolmont, E., Grasset, O., Lewis, K., Barnes, R., and Zuluaga, J. I. (2014). Formation, habitability, and detection of extrasolar moons. *Astrobiology,* 14(9):798–835.

Heller, R. and Hippke, M. (2017). Deceleration of High-velocity Interstellar Photon Sails into Bound Orbits at α Centauri. *Astrophys. J. Lett.,* 835(2):L32.

Heller, R., Anglada-Escudé, G., Hippke, M., and Kervella, P. (2020). Low-cost precursor of an interstellar mission. *Astron. Astrophys.,* 641:A45.

Henderson, L. J. (1913). *The Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter.* Macmillan, New York, NY.

Hendrix, A. R., Hurford, T. A., Barge, L. M., Bland, M. T., Bowman, J. S., Brinckerhoff, W., Buratti, B. J., Cable, M. L., Castillo-Rogez, J., Collins, G. C., Diniega, S., German, C. R., Hayes, A. G., Hoehler, T., Hosseini, S., Howett, C. J. A., McEwen, A. S., Neish, C. D., Neveu, M., . . . Vance, S. D. (2019). The NASA Roadmap to Ocean Worlds. *Astrobiology,* 19(1):1–27.

Hendry, A. P. (2017). *Eco-evolutionary Dynamics*. Princeton University Press, Princeton, NJ.

Heng, K. (2017). *Exoplanetary Atmospheres: Theoretical Concepts and Foundations*. Princeton University Press, Princeton, NJ.

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.* Princeton University Press, Princeton, NJ.

Henry, A. G., Büdel, T., and Bazin, P.-L. (2018). Towards an understanding of the costs of fire. *Quat. Int.,* 493:96–105.

Henry, T. J., Jao, W.-C., Subasavage, J. P., Beaulieu, T. D., Ianna, P. A., Costa, E., and Méndez, R. A. (2006). The solar neighborhood XVII: Parallax results from the CTIOPI 0.9 m program—Twenty new members of the RECONS 10 parsec sample. *Astron. J.,* 132(6):2360–2371.

Henry, T. J., Jao, W.-C., Winters, J. G., Dieterich, S. B., Finch, C. T., Ianna, P. A., Riedel, A. R., Silverstein, M. L., Subasavage, J. P., and Vrijmoet, E. H. (2018). The solar neighborhood XLIV: RECONS discoveries within 10 parsecs. *Astron. J.,* 155(6):265.

Heraclitus of Ephesus. (1892). The Fragments. In Burnet, J., *Early Greek Philosophy,* pg. 134. Adam & Charles Black, London, UK.

Herbst, E. and van Dishoeck, E. F. (2009). Complex Organic Interstellar Molecules. *Annu. Rev. Astron. Astrophys.,* 47:427–480.

Herculano-Houzel, S. (2016). *The Human Advantage: A New Understanding of How Our Brain Became Remarkable*. The MIT Press, Cambridge, MA.

Herrero, A., Stavans, J., and Flores, E. (2016). The multicellular nature of filamentous heterocyst-forming cyanobacteria. *FEMS Microbiol. Rev.,* 40(6):831–854.

Herries, A. I. R., Martin, J. M., Leece, A. B., Adams, J. W., Boschian, G., Joannes-Boyau, R., Edwards, T. R., Mallett, T., Massey, J., Murszewski, A., Neubauer, S., Pickering, R., Strait, D. S., Armstrong, B. J., Baker, S., Caruana, M. V., Denham, T., Hellstrom, J., Moggi-Cecchi, J., . . . Menter, C. (2020).

Contemporaneity of *Australopithecus, Paranthropus,* and early *Homo erectus* in South Africa. *Science,* 368(6486):eaaw7293.

Herzing, D. L. (2014). Profiling nonhuman intelligence: An exercise in developing unbiased tools for describing other "types" of intelligence on Earth. *Acta Astronaut.,* 94(2):676–680.

Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science,* 344(6190):1243091.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press, Cambridge, MA.

Higgs, P. G. and Pudritz, R. E. (2009). A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology,* 9(5):483–490.

Higgs, P. G. and Lehman, N. (2015). The RNA world: Molecular cooperation at the origins of life. *Nat. Rev. Genet.,* 16:7–17.

Hills, J. G. and Goda, M. P. (1993). The fragmentation of small asteroids in the atmosphere. *Astron. J.,* 105(3):1114–1144.

Hinkel, N. R., Hartnett, H. E., and Young, P. A. (2020). The influence of stellar phosphorus on our understanding of exoplanets and astrobiology. *Astrophys. J. Lett.,* 900(2):L38.

Hippke, M., Leyland, P., and Learned, J. G. (2018). Benchmarking inscribed matter probes. *Acta Astronaut.,* 151:32–36.

Hippke, M. (2018a). Benchmarking information carriers. *Acta Astronaut.,* 151: 53–62.

Hippke, M. (2018b). Interstellar communication. II. Application to the solar gravitational lens. *Acta Astronaut.,* 142:64–74.

Hippke, M. (2020). Interstellar Communication Network. I. Overview and Assumptions. *Astron. J.,* 159(3):85.

Hitchcock, D. R. and Lovelock, J. E. (1967). Life detection by atmospheric analysis. *Icarus,* 7:149–159.

Hlubik, S., Cutts, R., Braun, D. R., Berna, F., Feibel, C. S., and Harris, J. W. K. (2019). Hominin fire use in the Okote member at Koobi Fora, Kenya: New evidence for the old debate. *J. Hum. Evol.,* 133:214–229.

Ho, M.-Y., Shen, G., Canniffe, D. P., Zhao, C., and Bryant, D. A. (2016). Light-dependent chlorophyll f synthase is a highly divergent paralog of PsbA of photosystem II. *Science,* 353(6302):aaf9178.

Hoang, T. (2017). Relativistic Gas Drag on Dust Grains and Implications. *Astrophys. J.,* 847(1):77.

Hoang, T. and Loeb, A. (2017). Electromagnetic Forces on a Relativistic Spacecraft in the Interstellar Medium. *Astrophys. J.,* 848(1):31.

Hoang, T., Lazarian, A., Burkhart, B., and Loeb, A. (2017). The Interaction of Relativistic Spacecrafts with the Interstellar Medium. *Astrophys. J.,* 837(1):5.

Hoang, T. and Loeb, A. (2020). Detectability of Thermal Emission from Sub-Relativistic Objects. *arXiv e-prints,* arXiv:2007.04892.

Hobsbawm, E. (1995). *The Age of Extremes: The Short Twentieth Century 1914–1991.* Abacus, London, UK.

Hockett, C. F. (1960). The Origin of Speech. *Sci. Am.,* 203(3):88–96.

Hoehler, T. M., Amend, J. P., and Shock, E. L. (2007). A "Follow the Energy" Approach for Astrobiology. *Astrobiology,* 7(6):819–823.

Hoehler, T. M. and Jørgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.,* 11(2):83–94.

Hohmann-Marriott, M. F. and Blankenship, R. E. (2011). Evolution of photosynthesis. *Annu. Rev. Plant Biol.,* 62:515–548.

Hollinger, M. (2016). Life from Elsewhere–Early History of the Maverick Theory of Panspermia. *Sudhoffs Arch.,* 100(2):188–205.

Holman, M. J. and Murray, N. W. (2005). The Use of Transit Timing to Detect Terrestrial-Mass Extrasolar Planets. *Science,* 307(5713):1288–1291.

Höning, D., Tosi, N., Hansen-Goos, H., and Spohn, T. (2019). Bifurcation in the growth of continental crust. *Phys. Earth Planet. Inter.,* 287:37–50.

Hörandl, E. and Speijer, D. (2018). How oxygen gave rise to eukaryotic sex. *Proc. R. Soc. B,* 285(1872):20172706.

Hordijk, W. and Steel, M. (2017). Chasing the tail: The emergence of autocatalytic networks. *Biosystems,* 152:1–10.

Horneck, G., Klaus, D. M., and Mancinelli, R. L. (2010). Space Microbiology. *Microbiol. Mol. Biol. Rev.,* 74(1):121–156.

Horning, D. P. and Joyce, G. F. (2016). Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl. Acad. Sci. USA,* 113(35):9786–9791.

Hörst, S. M. (2017). Titan's atmosphere and climate. *J. Geophys. Res. Planets,* 122(3):432–482.

Hoshika, S., Leal, N. A., Kim, M.-J., Kim, M.-S., Karalkar, N. B., Kim, H.-J., Bates, A. M., Watkins Jr, N. E., SantaLucia, H. A., Meyer, A. J., DasGupta, S., Piccirilli, J. A., Ellington, A. D., SantaLucia Jr, J., Georgiadis, M. M., and Benner, S. A. (2019). Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science,* 363(6429):884–887.

Howard, A., Horowitz, P., Mead, C., Sreetharan, P., Gallicchio, J., Howard, S., Coldwell, C., Zajac, J., and Sliski, A. (2007). Initial results from Harvard all-sky optical SETI. *Acta Astronaut.,* 61:78–87.

Howard, A. W., Horowitz, P., Wilkinson, D. T., Coldwell, C. M., Groth, E. J., Jarosik, N., Latham, D. W., Stefanik, R. P., Willman, Jr., A. J., Wolff, J., and Zajac, J. M. (2004). Search for Nanosecond Optical Pulses from Nearby Solar-Type Stars. *Astrophys. J.,* 613(2):1270–1284.

Howard, W. S., Tilley, M. A., Corbett, H., Youngblood, A., Loyd, R. O. P., Ratzloff, J. K., Law, N. M., Fors, O., del Ser, D., Shkolnik, E. L., Ziegler, C., Goeke, E. E., Pietraallo, A. D., and Haislip, J. (2018). The First Naked-eye Superflare Detected from Proxima Centauri. *Astrophys. J. Lett.,* 860(2):L30.

Howard, W. S., Corbett, H., Law, N. M., Ratzloff, J. K., Glazier, A., Fors, O., del Ser, D. and Haislip, J. (2019). EvryFlare. I. Long-term Evryscope Monitoring of Flares from the Cool Stars across Half the Southern Sky. *Astrophys. J.,* 881(1):9.

Hoyle, F. and Wickramasinghe, N. C. (2000). *Astronomical Origins of Life: Steps Towards Panspermia*. Kluwer, Dordrecht, Netherlands.

Hrdy, S. B. (2009). *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Harvard University Press, Cambridge, MA.

Huang, S.-S. (1960). Life Outside the Solar System. *Sci. Am.,* 202(4):55–63.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.

Huber, C. and Wächtershäuser, G. (2006). α-hydroxy and α-amino acids under possible Hadean, volcanic origin-of-life conditions. *Science,* 314(5799):630–632.

Hud, N. V., Cafferty, B. J., Krishnamurthy, R., and Williams, L. D. (2013). The origin of RNA and "my grandfather's axe." *Chem. Biol.,* 20(4):466–474.

Hud, N. V. (2018). Searching for lost nucleotides of the pre-RNA world with a self-refining model of early Earth. *Nat. Commun.,* 9:5171.

Hudson, H. S. (2015). Solar extreme events. *J. Phys. Conf. Ser.,* 632(1):012058.

Hudson, R., de Graaf, R., Rodin, M. S., Ohno, A., Lane, N., McGlynn, S. E., Yamada, Y. M. A., Nakamura, R., Barge, L. M., Braun, D., and Sojo, V. (2020). CO_2 reduction driven by a pH gradient. *Proc. Natl. Acad. Sci. USA,* 117(37):22873–22879.

Hudson, S. R., Startsev, E., and Feibush, E. (2014). A new class of magnetic confinement device in the shape of a knot. *Phys. Plasmas,* 21(1):010705.

Husnik, F. and McCutcheon, J. P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.,* 16(2):67–79.

Hussmann, H., Sohl, F., and Spohn, T. (2006). Subsurface oceans and deep interiors of medium-sized outer planet satellites and large trans-neptunian objects. *Icarus,* 185(1):258–273.

Ilardo, M., Bose, R., Meringer, M., Rasulev, B., Grefenstette, N., Stephenson, J., Freeland, S., Gillams, R. J., Butch, C. J., and Cleaves, H. J. (2019). Adaptive

properties of the genetically encoded amino acid alphabet are inherited from its subsets. *Sci. Rep.,* 9:12468.

Ilic, O. and Atwater, H. A. (2019). Self-stabilizing photonic levitation and propulsion of nanostructured macroscopic objects. *Nat. Photonics,* 13:289–295.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., . . . Takai, K. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature,* 577:519–525.

Imara, N. and Di Stefano, R. (2018). Searching for Exoplanets around X-Ray Binaries with Accreting White Dwarfs, Neutron Stars, and Black Holes. *Astrophys. J.,* 859(1):40.

Inoue, M. and Yokoo, H. (2011). Type III Dyson Sphere of Highly Advanced Civilisations around a Super Massive Black Hole. *J. Br. Interplanet. Soc.,* 64(3):58–62.

Inoue, S. and Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Curr. Biol.,* 17(23):R1004–R1005.

Isaacson, H., Siemion, A. P. V., Marcy, G. W., Lebofsky, M., Price, D. C., MacMahon, D., Croft, S., DeBoer, D., Hickish, J., Werthimer, D., Sheikh, S., Hellbourg, G., and Enriquez, J. E. (2017). The Breakthrough Listen Search for Intelligent Life: Target Selection of Nearby Stars and Galaxies. *Publ. Astron. Soc. Pac.,* 129(975):054501.

Islam, S., Bučar, D.-K., and Powner, M. W. (2017). Prebiotic selection and assembly of proteinogenic amino acids and natural nucleotides from complex mixtures. *Nat. Chem.,* 9(6):584–589.

Islam, S. and Powner, M. W. (2017). Prebiotic systems chemistry: Complexity overcoming clutter. *Chem,* 2(4):470–501.

Isson, T. T., Planavsky, N. J., Coogan, L. A., Stewart, E. M., Ague, J. J., Bolton, E. W., Zhang, S., McKenzie, N. R., and Kump, L. R. (2020). Evolution of the Global Carbon Cycle and Climate Regulation on Earth. *Glob. Biogeochem. Cy.,* 34(2):e2018GB006061.

Ivanov, V. D., Beamín, J. C., Cáceres, C., and Minniti, D. (2020). A qualitative classification of extraterrestrial civilizations. *Astron. Astrophys.,* 639:A94.

Jaakkola, S. T., Ravantti, J. J., Oksanen, H. M., and Bamford, D. H. (2016). Buried Alive: Microbes from Ancient Halite. *Trends Microbiol.,* 24(2):148–160.

Jackson, A. A. and Whitmire, D. P. (1978). Laser Powered Interstellar Rocket. *J. Br. Interplanet. Soc.,* 31:335–337.

Jackson, J. B. (2016). Natural pH gradients in hydrothermal alkali vents were unlikely to have played a role in the origin of life. *J. Mol. Evol.,* 83:1–11.

Jacob, D. J. (1999). *Introduction to Atmospheric Chemistry*. Princeton University Press, Princeton, NJ.

Jacob, F. (1982). *The Possible and the Actual*. University of Washington Press, Seattle, WA.

Jacobson, H. R., Thanathibodee, T., Frebel, A., Roederer, I. U., Cescutti, G., and Matteucci, F. (2014). The Chemical Evolution of Phosphorus. *Astrophys. J. Lett.,* 796(2):L24.

Jafarpour, F., Biancalani, T., and Goldenfeld, N. (2017). Noise-induced symmetry breaking far from equilibrium and the emergence of biological homochirality. *Phys. Rev. E,* 95(3):032407.

Jakosky, B. M. and Shock, E. L. (1998). The biological potential of Mars, the early Earth, and Europa. *J. Geophys. Res. Planets,* 103(E8):19359–19364.

Jakosky, B. M. (2019). The CO_2 inventory on Mars. *Planet. Space Sci.,* 175: 52–59.

Janhunen, P. (2004). Electric Sail for Spacecraft Propulsion. *J. Propuls. Power,* 20(4):763–764.

Janhunen, P. and Sandroos, A. (2007). Simulation study of solar wind push on a charged wire: basis of solar wind electric sail propulsion. *Ann. Geophys.,* 25(3):755–767.

Janhunen, P., Toivanen, P. K., Polkko, J., Merikallio, S., Salminen, P., Haeggström, E., Seppänen, H., Kurppa, R., Ukkonen, J., Kiprich, S., Thornell, G., Kratz, H., Richter, L., Krömer, O., Rosta, R., Noorma, M., Envall, J., Lätt, S., Mengali, G., . . . Obraztsov, A. (2010). Electric solar wind sail: Toward test missions. *Rev. Sci. Instrum.,* 81(11):111301.

Janik, V. M. (2014). Cetacean vocal learning and communication. *Curr. Opin. Neurobiol.,* 28:60–65.

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science,* 366(6461):50–54.

Javaux, E. J. and Lepot, K. (2018). The Paleoproterozoic fossil record: Implications for the evolution of the biosphere during Earth's middle-age. *Earth-Sci. Rev.,* 176:68–86.

Javaux, E. J. (2019). Challenges in evidencing the earliest traces of life. *Nature,* 572(7770):451–460.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jebbar, M., Hickman-Lewis, K., Cavalazzi, B., Taubner, R.-S., Rittmann, Simon K.-M. R., and Antunes, A. (2020). Microbial Diversity and Biosignatures: An Icy Moons Perspective. *Space Sci. Rev.,* 216(1):10.

Jelen, B. I., Giovannelli, D., and Falkowski, P. G. (2016). The Role of Microbial Electron Transfer in the Coevolution of the Biosphere and Geosphere. *Annu. Rev. Microbiol.,* 70:45–62.

Jewitt, D., Hui, M.-T., Kim, Y., Mutchler, M., Weaver, H., and Agarwal, J. (2020). The Nucleus of Interstellar Comet 2I/Borisov. *Astrophys. J. Lett.,* 888 (2):L23.

Jia, T. Z., Chandru, K., Hongo, Y., Afrin, R., Usui, T., Myojo, K., and Cleaves, H. J. (2019). Membraneless polyester microdroplets as primordial compartments at the origins of life. *Proc. Natl. Acad. Sci. USA,* 116(32):15830–15835.

Jin, W., Li, W., Orenstein, M., and Fan, S. (2020). Inverse design of lightweight broadband reflector for relativistic lightsail propulsion. *ACS Photonics,* 7(9): 2350–2355.

Johnson, B. W. and Wing, B. A. (2020). Limited Archaean continental emergence reflected in an early Archaean ^{18}O-enriched ocean. *Nat. Geosci.,* 13: 243–248.

Johnson, R. E., Quickenden, T. I., Cooper, P. D., McKinley, A. J., and Freeman, C. G. (2003). The production of oxidants in Europa's surface. *Astrobiology,* 3(4):823–850.

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends Cogn. Sci.,* 8(2):71–78.

Johnstone, C. P., Güdel, M., Brott, I., and Lüftinger, T. (2015). Stellar winds on the main-sequence. II. The evolution of rotation and winds. *Astron. Astrophys.,* 577:A28.

Jones, B. W. (2003). The search for extraterrestrial intelligence. *Phys. Educ.,* 38(1): 8–13.

Jones, C., Nomosatryo, S., Crowe, S. A., Bjerrum, C. J., and Canfield, D. E. (2015). Iron oxides, divalent cations, silica, and the early earth phosphorus crisis. *Geology,* 43(2):135–138.

Jones, E. G. and Lineweaver, C. H. (2010). To What Extent Does Terrestrial Life "Follow The Water"? *Astrobiology,* 10(3):349–361.

Jones, R. M., Goordial, J. M., and Orcutt, B. N. (2018). Low Energy Subsurface Environments as Extraterrestrial Analogs. *Front. Microbiol.,* 9:1605.

Jordan, S. F., Rammu, H., Zheludev, I. N., Hartley, A. M., Marechal, A., and Lane, N. (2019). Promotion of protocell self-assembly from mixed amphiphiles at the origin of life. *Nat. Ecol. Evol.,* 3:1705–1714.

Joshi, M. M., Haberle, R. M., and Reynolds, R. T. (1997). Simulations of the Atmospheres of Synchronously Rotating Terrestrial Planets Orbiting M Dwarfs: Conditions for Atmospheric Collapse and the Implications for Habitability. *Icarus,* 129(2):450–465.

Joyce, G. F., Visser, G. M., van Boeckel, C. A., van Boom, J. H., Orgel, L. E., and van Westrenen, J. (1984). Chiral selection in poly(C)-directed synthesis of oligo(G). *Nature,* 310(5978):602–604.

Joyce, G. F. (1994). Foreword. In Deamer, D. W. and Fleischaker, G., editors, *Origins of Life: The Central Concepts*. Jones and Bartlett, Boston, MA.

Joyce, G. F. (2002). The antiquity of RNA-based evolution. *Nature,* 418(6894):214–221.

Joyce, G. F. and Szostak, J. W. (2018). Protocells and RNA self-replication. *Cold Spring Harb. Perspect. Biol.,* 10(9):a034801.

Juba, B. (2011). *Universal Semantic Communication*. Springer-Verlag, Berlin, Germany.

Judson, O. (2017). The energy expansions of evolution. *Nat. Ecol. Evol.,* 1:0138.

Jugaku, J. and Nishimura, S. (2004). A Search for Dyson Spheres Around Late-type Stars in the Solar Neighborhood. In Norris, R. and Stootman, F., editors, *IAU Symp. 213: Bioastronomy 2002: Life Among the Stars,* pages 437–438. Astronomical Society of the Pacific, San Francisco, CA.

Kabadayi, C. and Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science,* 357(6347):202–204.

Kadoya, S., Catling, D. C., Nicklas, R. W., Puchtel, I. S., and Anbar, A. D. (2020). Mantle data imply a decline of oxidizable volcanic gases could have triggered the Great Oxidation. *Nat. Commun.,* 11:2774.

Kadoya, S., Krissansen-Totton, J., and Catling, D. C. (2020). Probable cold and alkaline surface environment of the Hadean Earth caused by impact ejecta weathering. *Geochem. Geophys. Geosyst.,* 21(1):e2019GC008734.

Kahana, A., Schmitt-Kopplin, P., and Lancet, D. (2019). Enceladus: First observed primordial soup could arbitrate origin-of-life debate. *Astrobiology,* 19(10): 1263–1278.

Kalousová, K., Sotin, C., Choblet, G., Tobie, G., and Grasset, O. (2018). Two-phase convection in Ganymede's high-pressure ice layer - Implications for its geological evolution. *Icarus,* 299:133–147.

Kaltenegger, L. (2017). How to Characterize Habitable Worlds and Signs of Life. *Annu. Rev. Astron. Astrophys.,* 55:433–485.

Kamerlin, S. C. L., Sharma, P. K., Prasad, R. B., and Warshel, A. (2013). Why nature really chose phosphate. *Q. Rev. Biophys.,* 46(1):1–132.

Kamminga, H. (1982). Life from space–A history of panspermia. *Vistas Astron.,* 26(2):67–86.

Kardashev, N. S. (1964). Transmission of Information by Extraterrestrial Civilizations. *Soviet Ast.,* 8(2):217–221.

Kardashev, N. S. (1979). Optimal wavelength region for communication with extraterrestrial intelligence: $\lambda = 1.5$ mm. *Nature,* 278(5699):28–30.

Kardashev, N. S. (1985). On the inevitability and the possible structures of supercivilizations. In Papagiannis, M. D., editor, *IAU Symp. 112: The Search for Extraterrestrial Life: Recent Developments,* pages 497–504. D. Reidel Publishing Co., Dordrecht, Netherlands.

Kardashev, N. S. (1997). Cosmology and Civilizations. *Astrophys. Space Sci.,* 252: 25–40.

Kargel, J. S., Kaye, J. Z., Head, J. W., Marion, G. M., Sassen, R., Crowley, J. K., Ballesteros, O. P., Grant, S. A., and Hogenboom, D. L. (2000). Europa's Crust and Ocean: Origin, Composition, and the Prospects for Life. *Icarus,* 148(1):226–265.

Karl, D. M. (2000). Phosphorus, the staff of life. *Nature,* 406(6791):31–33.

Karl, D. M. and Björkman, K. M. (2015). Dynamics of Dissolved Organic Phosphorus. In Hansell, D. A. and Carlson, C. A., editors, *Biogeochemistry of Marine Dissolved Organic Matter,* pages 233–334. Academic Press, Waltham, MA (2nd edition).

Kasting, J. F., Whitmire, D. P., and Reynolds, R. T. (1993). Habitable zones around main sequence stars. *Icarus,* 101(1):108–128.

Kasting, J. (2010). *How to Find a Habitable Planet*. Princeton University Press, Princeton, NJ.

Kasting, J. F. (2019). The Goldilocks Planet? How Silicate Weathering Maintains Earth "Just Right". *Elements,* 15(4):235–240.

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *J. Theor. Biol.,* 119(1):1–24.

Kauffman, S. A. (2019). *A World beyond Physics: The Emergence and Evolution of Life*. Oxford University Press, Oxford, UK.

Kay, C., Opher, M., and Kornbleuth, M. (2016). Probability of CME Impact on Exoplanets Orbiting M Dwarfs and Solar-like Stars. *Astrophys. J.,* 826(2): 195.

Kay, C., Airapetian, V. S., Lüftinger, T., and Kochukhov, O. (2019). Frequency of Coronal Mass Ejection Impacts with Early Terrestrial Planets and Exoplanets around Active Solar-like Stars. *Astrophys. J. Lett.,* 886(2):L37.

Kecskes, C. (1998). The possibility of finding traces of extraterrestrial intelligence on asteroids. *J. Br. Interplanet. Soc.,* 51(5):175–179.

Keller, M. A., Turchyn, A. V., and Ralser, M. (2014). Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Mol. Syst. Biol.,* 10(4):725.

Keller, M. A., Kampjut, D., Harrison, S. A., and Ralser, M. (2017). Sulfate radicals enable a non-enzymatic Krebs cycle precursor. *Nat. Ecol. Evol.,* 1:0083.

Kempes, C. P., Wolpert, D., Cohen, Z., and Pérez-Mercader, J. (2017). The thermodynamic efficiency of computations made in cells across the range of life. *Phil. Trans. R. Soc. A,* 375(2109):20160343.

Kempton, E. M. -R., Bean, J. L., Louie, D. R., Deming, D., Koll, D. D. B., Mansfield, M., Christiansen, J. L., López-Morales, M., Swain, M. R., Zellem, R. T., Ballard, S., Barclay, T., Barstow, J. K., Batalha, N. E., Beatty, T. G., Berta-Thompson, Z., Birkby, J., Buchhave, L. A., Charbonneau, D., . . . von Essen, C. (2018). A Framework for Prioritizing the TESS Planetary Candidates Most Amenable to Atmospheric Characterization. *Publ. Astron. Soc. Pac.,* 130(993):114401.

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., and Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends Cogn. Sci.,* 22(7):651–665.

Kepler, J. (1965). *Conversation with the Sidereal Messenger.* Translated by E. Rosen. Johnson Reprint Corp., New York, NY.

Kessler, D. J. and Cour-Palais, B. G. (1978). Collision frequency of artificial satellites: The creation of a debris belt. *J. Geophys. Res.,* 83(A6):2637–2646.

Kharecha, P., Kasting, J., and Siefert, J. (2005). A coupled atmosphere-ecosystem model of the early Archean Earth. *Geobiology,* 3(2):53–76.

Kiang, N. Y., Segura, A., Tinetti, G., Govindjee, Blankenship, R. E., Cohen, M., Siefert, J., Crisp, D., and Meadows, V. S. (2007). Spectral signatures of photosynthesis. II. Coevolution with other stars and the atmosphere on extrasolar worlds. *Astrobiology,* 7(1):252–274.

Kiang, N. Y., Siefert, J., Govindjee, and Blankenship, R. E. (2007). Spectral Signatures of Photosynthesis. I. Review of Earth Organisms. *Astrobiology,* 7(1): 222–251.

Kiang, N. Y., Domagal-Goldman, S., Parenteau, M. N., Catling, D. C., Fujii, Y., Meadows, V. S., Schwieterman, E. W., and Walker, S. I. (2018). Exoplanet Biosignatures: At the Dawn of a New Era of Planetary Observations. *Astrobiology,* 18(6):619–629.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., Mutke, J., and Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proc. Natl. Acad. Sci. USA,* 106(23):9322–9327.

Kim, H.-J., Furukawa, Y., Kakegawa, T., Bita, A., Scorei, R., and Benner, S. A. (2016). Evaporite Borate-Containing Mineral Ensembles Make Phosphate Available and Regiospecifically Phosphorylate Ribonucleosides: Borate as a Multifaceted Problem Solver in Prebiotic Chemistry. *Angew. Chem.,* 128(51):6048–16052.

Kim, H.-J. and Benner, S. A. (2017). Prebiotic stereoselective synthesis of purine and noncanonical pyrimidine nucleotide from nucleobases and phosphorylated carbohydrates. *Proc. Natl. Acad. Sci. USA,* 114(43):11315–11320.

Kim, H., Smith, H. B., Mathis, C., Raymond, J., and Walker, S. I. (2019). Universal scaling across biochemical networks on Earth. *Sci. Adv.,* 5(1):eaau0149.

Kimbel, W. H. and Villmoare, B. (2016). From *Australopithecus* to *Homo*: The transition that wasn't. *Phil. Trans. R. Soc. B,* 371(1698):20150248.

Kimura, J. and Kitadai, N. (2015). Polymerization of Building Blocks of Life on Europa and Other Icy Moons. *Astrobiology,* 15(6):430–441.

Kingsolver, J. G. and Huey, R. B. (2008). Size, temperature, and fitness: three rules. *Evol. Ecol. Res.,* 10(2):251–268.

Kingsolver, J. G. (2009). The Well-Temperatured Biologist. *Am. Nat.,* 174(6):755–768.

Kipp, M. A. and Stüeken, E. E. (2017). Biomass recycling and Earth's early phosphorus cycle. *Sci. Adv.,* 3(11):eaao4795.

Kipping, D. M. and Teachey, A. (2016). A cloaking device for transiting planets. *Mon. Not. R. Astron. Soc.,* 459(2):1233–1241.

Kipping, D. (2019). Transiting Quasites as a Possible Technosignature. *Res. Notes AAS,* 3(7):91.

Kipping, D. (2020). An objective Bayesian analysis of life's early start and our late arrival. *Proc. Natl. Acad. Sci. USA,* 117(22):11995–12003.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *J. R. Soc. Interface,* 15(138):20170792.

Kirkpatrick, J. D., Martin, E. C., Smart, R. L., Cayago, A. J., Beichman, C. A., Marocco, F., Gelino, C. R., Faherty, J. K., Cushing, M. C., Schneider, A. C., Mace, G. N., Tinney, C. G., Wright, E. L., Lowrance, P. J., Ingalls, J. G., Vrba, F. J., Munn, J. A., Dahm, S. E., and McLean, I. S. (2019). Preliminary Trigonometric Parallaxes of 184 Late-T and Y Dwarfs and an Analysis of the Field Substellar Mass Function into the "Planetary" Mass Regime. *Astrophys. J., Suppl. Ser.,* 240(2):19.

Kissel, M. and Fuentes, A. (2018). "Behavioral modernity" as a process, not an event, in the human niche. *Time & Mind,* 11(2):163–183.

Kitadai, N. and Maruyama, S. (2018). Origins of building blocks of life: A review. *Geosci. Front.,* 9(4):1117–1153.

Kitadai, N., Nakamura, R., Yamamoto, M., Takai, K., Yoshida, N., and Oono, Y. (2019). Metals likely promoted protometabolism in early ocean alkaline hydrothermal systems. *Sci. Adv.,* 5(6):eaav7848.

Kite, E. S. and Ford, E. B. (2018). Habitability of Exoplanet Waterworlds. *Astrophys. J.,* 864(1):75.

Kite, E. S., Gaidos, E., and Onstott, T. C. (2018). Valuing Life-Detection Missions. *Astrobiology,* 18(7):834–840.

Kite, E. S. (2019). Geologic constraints on early Mars climate. *Space Sci. Rev.,* 215(1):10.

Kivelson, M. G. and Ridley, A. J. (2008). Saturation of the polar cap potential: Inference from Alfvén wing arguments. *J. Geophys. Res. Space Phys.,* 113(A5):A05214.

Klein, R. G. (2000). Archeology and the evolution of human behavior. *Evol. Anthropol.,* 9(1):17–36.

Klein, R. G. (2008). Out of Africa and the evolution of human behavior. *Evol. Anthropol.,* 17(6):267–281.

Klein, R. G. (2009). *The Human Career: Human Biological and Cultural Origins*. The University of Chicago Press, Chicago, IL (3rd edition).

Knight, R. D., Freeland, S. J., and Landweber, L. F. (2001). Rewiring the keyboard: Evolvability of the genetic code. *Nat. Rev. Genet.,* 2(1):49–58.

Knoll, A. H. and Bambach, R. K. (2000). Directionality in the history of life: Diffusion from the left wall or repeated scaling of the right? *Paleobiology,* 26(sp4):1–14.

Knoll, A. H., Bambach, R. K., Payne, J. L., Pruss, S., and Fischer, W. W. (2007). Paleophysiology and end-Permian mass extinction. *Earth Planet. Sci. Lett.,* 256:295–313.

Knoll, A. H. (2011). The multiple origins of complex multicellularity. *Annu. Rev. Earth Planet. Sci.,* 39:217–39.

Knoll, A. H. (2015). *Life on a Young Planet: The First Three Billion Years of Evolution on Earth*. Princeton University Press, Princeton, NJ (2nd edition).

Knoll, A. H. (2017). Food for early animal evolution. *Nature,* 548(7669): 528–530.

Knoll, A. H. and Nowak, M. A. (2017). The timetable of evolution. *Sci. Adv.,* 3(5):e1603076.

Knoll, A. H. (2021). *A Brief History of Earth: Four Billion Years in Eight Chapters*. HarperCollins, New York, NY.

Kobayashi, K., Kasamatsu, T., Kaneko, T., Koike, J., Oshima, T., Saito, T., Yamamoto, T., and Yanagawa, H. (1995). Formation of amino acid precursors in cometary ice environments by cosmic radiation. *Adv. Space Res.,* 16(2):21–26.

Koga, S., Williams, D. S., Perriman, A. W., and Mann, S. (2011). Peptide-nucleotide microdroplets as a step towards a membrane-free protocell model. *Nat. Chem.,* 3:720–724.

Koga, T. and Naraoka, H. (2017). A new family of extraterrestrial amino acids in the Murchison meteorite. *Sci. Rep.,* 7:636.

Kohda, M., Hotta, T., Takeyama, T., Awata, S., Tanaka, H., Asai, J.-y., and Jordan, A. L. (2019). If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals? *PLoS Biol.,* 17(2):e3000021.

Koll, D. D. B. and Abbot, D. S. (2016). Temperature Structure and Atmospheric Circulation of Dry Tidally Locked Rocky Exoplanets. *Astrophys. J.,* 825(2):99.

Koll, D. D. B., Malik, M., Mansfield, M., Kempton, E. M. -R., Kite, E., Abbot, D., and Bean, J. L. (2019). Identifying Candidate Atmospheres on Rocky M Dwarf Planets via Eclipse Photometry. *Astrophys. J.,* 886(2):140.

Komacek, T. D., Fauchez, T. J., Wolf, E. T., and Abbot, D. S. (2020). Clouds will Likely Prevent the Detection of Water Vapor in *JWST* Transmission Spectra of Terrestrial Exoplanets. *Astrophys. J. Lett.,* 888(2):L20.

Kondepudi, D. and Prigogine, I. (2015). *Modern Thermodynamics: From Heat Engines to Dissipative Structures*. Wiley, Chichester, UK (2nd edition).

Koonin, E. V. and Yutin, N. (2014). The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.,* 6(4):a016188.

Koonin, E. V. (2016). Viruses and mobile elements as drivers of evolutionary transitions. *Phil. Trans. R. Soc. B,* 371(1701):20150442.

Koonin, E. V. and Novozhilov, A. S. (2017). Origin and evolution of the universal genetic code. *Annu. Rev. Genet.,* 51:45–62.

Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V., and Deshpande, R. (2013). Habitable zones around main-sequence stars: New estimates. *Astrophys. J.,* 765(2):131.

Korenaga, J. (2008). Plate tectonics, flood basalts and the evolution of Earths oceans. *Terra Nova,* 20(6):419–439.

Korenaga, J. (2013). Initiation and Evolution of Plate Tectonics on Earth: Theories and Observations. *Annu. Rev. Earth Planet. Sci.,* 41:117–151.

Korenaga, J. (2018). Crustal evolution and mantle dynamics through Earth history. *Phil. Trans. R. Soc. A,* 376(2132):20170408.

Korpela, E. J., Sallmen, S. M., and Leystra Greene, D. (2015). Modeling Indications of Technology in Planetary Transit Light Curves-Dark-side Illumination. *Astrophys. J.,* 809(2):139.

Koshland, D. E. (2002). The seven pillars of life. *Science,* 295(5563):2215–2216.

Krauss, L. M. and Starkman, G. D. (2000). Life, the Universe, and Nothing: Life and Death in an Ever-expanding Universe. *Astrophys. J.,* 531(1): 22–30.

Kreidberg, L. and Loeb, A. (2016). Prospects for Characterizing the Atmosphere of Proxima Centauri b. *Astrophys. J. Lett.,* 832(1):L12.

Kreidberg, L., Koll, D. D. B., Morley, C., Hu, R., Schaefer, L., Deming, D., Stevenson, K. B., Dittmann, J., Vanderburg, A., Berardo, D., Guo, X., Stassun, K., Crossfield, I., Charbonneau, D., Latham, D. W., Loeb, A., Ricker, G., Seager, S., and Vanderspek, R. (2019). Absence of a thick atmosphere on the terrestrial exoplanet LHS 3844b. *Nature,* 573(7772):87–90.

Kreysing, M., Keil, L., Lanzmich, S., and Braun, D. (2015). Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length. *Nat. Chem.,* 7:203–208.

Krijt, S., Bowling, T. J., Lyons, R. J., and Ciesla, F. J. (2017). Fast Lithopanspermia in the Habitable Zone of the TRAPPIST-1 System. *Astrophys. J. Lett.,* 839(2):L21.

Krissansen-Totton, J., Bergsman, D. S., and Catling, D. C. (2016). On Detecting Biospheres from Chemical Thermodynamic Disequilibrium in Planetary Atmospheres. *Astrobiology,* 16(1):39–67.

Krissansen-Totton, J., Arney, G. N., and Catling, D. C. (2018a). Constraining the climate and ocean pH of the early Earth with a geological carbon cycle model. *Proc. Natl. Acad. Sci. USA,* 115(16):4105–4110.

Krissansen-Totton, J., Olson, S., and Catling, D. C. (2018b). Disequilibrium biosignatures over Earth history and implications for detecting exoplanet life. *Sci. Adv.,* 4(1):eaao5747.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science,* 354(6308):110–114.

Krupenye, C. and Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdiscip. Rev. Cogn. Sci.,* 10(6):e1503.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, IL.

Kuhn, J. R. and Berdyugina, S. V. (2015). Global warming as a detectable thermodynamic marker of Earth-like extrasolar civilizations: the case for a telescope like Colossus. *Int. J. Astrobiol.,* 14(3):401–410.

Kulkarni, N., Lubin, P., and Zhang, Q. (2018). Relativistic Spacecraft Propelled by Directed Energy. *Astron. J.,* 155(4):155.

Kurosawa, K., Sugita, S., Ishibashi, K., Hasegawa, S., Sekine, Y., Ogawa, N. O., Kadono, T., Ohno, S., Ohkouchi, N., Nagaoka, Y., and Matsui, T. (2013). Hydrogen cyanide production due to mid-size impacts in a redox-neutral N_2-rich atmosphere. *Orig. Life Evol. Biosph.,* 43(3):221–245.

Kurzban, R., Burton-Chellew, M. N., and West, S. A. (2015). The evolution of altruism in humans. *Annu. Rev. Psychol.,* 66:575–599.

Laakso, T. A. and Schrag, D. P. (2017). A theory of atmospheric oxygen. *Geobiology,* 15(3):366–384.

Laakso, T. A. and Schrag, D. P. (2018). Limitations on Limitation. *Global Biogeochem. Cy.,* 32(3):486–496.

Laakso, T. A. and Schrag, D. P. (2019). A small marine biosphere in the Proterozoic. *Geobiology,* 17(2):161–171.

Laakso, T. A., Sperling, E. A., Johnston, D. T., and Knoll, A. H. (2020). Ediacaran reorganization of the marine phosphorus cycle. *Proc. Natl. Acad. Sci. USA,* 117(22):11961–11967.

Lacki, B. C. (2016). Type III Societies (Apparently) Do Not Exist. *arXiv e-prints,* arXiv:1604.07844.

Lacki, B. C. (2019). A shiny new method for SETI: Specular reflections from interplanetary artifacts. *Publ. Astron. Soc. Pac.,* 131(1002):084401.

Lacki, B. C., Brzycki, B., Croft, S., Czech, D., DeBoer, D., DeMarines, J., Gajjar, V., Isaacson, H., Lebofsky, M., MacMahon, D. H. E., Price, D. C., Sheikh, S. Z., Siemion, A. P. V., Drew, J., and Worden, S. P. (2020). One of everything: The Breakthrough Listen *Exotica Catalog*. arXiv e-prints. arXiv:2006.11304.

Laland, K., Matthews, B., and Feldman, M. W. (2016). An introduction to niche construction theory. *Evol. Ecol.,* 30(2):191–202.

Laland, K. N. (2017). *Darwin's Unfinished Symphony: How Culture Made the Human Mind*. Princeton University Press, Princeton, NJ.

Lambert, J.-F. (2008). Adsorption and polymerization of amino acids on mineral surfaces: A review. *Orig. Life Evol. Biosph.,* 38(3):211–242.

Lammer, H., Bredehöft, J. H., Coustenis, A., Khodachenko, M. L., Kaltenegger, L., Grasset, O., Prieur, D., Raulin, F., Ehrenfreund, P., Yamauchi, M., Wahlund, J.-E., Grießmeier, J.-M., Stangl, G., Cockell, C. S., Kulikov, Y. N., Grenfell, J. L., and Rauer, H. (2009). What makes a planet habitable? *Astron. Astrophys. Rev.,* 17(2):181–249.

Lammer, H. (2013). *Origin and Evolution of Planetary Atmospheres: Implications for Habitability*. Springer, Heidelberg, Germany.

Lammer, H., Zerkle, A. L., Gebauer, S., Tosi, N., Noack, L., Scherf, M., Pilat-Lohinger, E., Güdel, M., Grenfell, J. L., Godolt, M., and Nikolaou, A. (2018). Origin and evolution of the atmospheres of early Venus, Earth and Mars. *Astron. Astrophys. Rev.,* 26:2.

Lammer, H., Sproß, L., Grenfell, J. L., Scherf, M., Fossati, L., Lendl, M., and Cubillos, P. E. (2019). The role of N_2 as a geo-biosignature for the detection and characterization of Earth-like habitats. *Astrobiology,* 19(7):927–950.

Lancet, D., Zidovetzki, R., and Markovitch, O. (2018). Systems protobiology: Origin of life in lipid catalytic networks. *J. R. Soc. Interface,* 15(144): 20180159.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.,* 5(3):183–191.

Landgraf, M., Baggaley, W. J., Grün, E., Krüger, H., and Linkert, G. (2000). Aspects of the mass distribution of interstellar dust grains in the solar system from in situ measurements. *J. Geophys. Res.,* 105(A5):10343–10352.

Landis, G. (1998). The Fermi paradox: An approach based on percolation theory. *J. Br. Interplanet. Soc.,* 51(5):163–166.

Lane, N. (2002). *Oxygen: The Molecule that Made the World*. Oxford University Press, Oxford, UK.

Lane, N. and Martin, W. (2010). The energetics of genome complexity. *Nature,* 467(7318):929–934.

Lane, N. (2015). *The Vital Question: Energy, Evolution, and the Origins of Complex Life*. W. W. Norton & Co., New York, NY.

Lane, N. (2017a). Proton gradients at the origin of life. *BioEssays,* 39(6):1600217.

Lane, N. (2017b). Serial endosymbiosis or singular event at the origin of eukaryotes? *J. Theor. Biol.,* 434:58–67.

Lapôtre, M. G. A., O'Rourke, J. G., Schaefer, L. K., Siebach, K. L., Spalding, C., Tikoo, S. M., and Wordsworth, R. D. (2020). Probing space to understand Earth. *Nat. Rev. Earth Environ.,* 1:170–181.

Larkum, A. W. D., Ritchie, R. J., and Raven, J. A. (2018). Living off the Sun: chlorophylls, bacteriochlorophylls and rhodopsins. *Photosynthetica,* 56(1): 11–43.

Laumer, C E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., Andrade, S. C. S., Sterrer, W., Sørensen, M. V., and Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B,* 286(1906):20190831.

Lazcano, A. and Miller, S. L. (1994). How long did it take for life to begin and evolve to cyanobacteria? *J. Mol. Evol.,* 39(6):546–554.

Lazcano, A. (2010). Historical development of origins research. *Cold Spring Harb. Perspect. Biol.,* 2(11):a002089.

Lazcano, A. and Peretó, J. (2017). On the origin of mitosing cells: A historical appraisal of Lynn Margulis endosymbiotic theory. *J. Theor. Biol.,* 434:80–87.

Lazcano, A. (2018). Prebiotic evolution and self-assembly of nucleic acids. *ACS Nano,* 12(10):9643–9647.

Learned, J. G., Pakvasa, S. and Zee, A. (2009). Galactic neutrino communication. *Phys. Lett. B,* 671(1):15–19.

Lebofsky, M., Croft, S., Siemion, A. P. V., Price, D. C., Enriquez, J. E., Isaacson, H., MacMahon, D. H. E., Anderson, D., Brzycki, B., Cobb, J., Czech, D., DeBoer, D., DeMarines, J., Drew, J., Foster, G., Gajjar, V., Gizani, N., Hellbourg, G., Korpela, E. J., . . . Zhang, Y. G. (2019). The Breakthrough Listen search for intelligent life: Public data, formats, reduction, and archiving. *Publ. Astron. Soc. Pac.,* 131(1006):124505.

Leconte, J., Forget, F., Charnay, B., Wordsworth, R., Selsis, F., Millour, E., and Spiga, A. (2013). 3D climate modeling of close-in land planets: Circulation patterns, climate moist bistability, and habitability. *Astron. Astrophys.,* 554: A69.

Lederberg, J. (1965). Signs of life: Criterion-system of exobiology. *Nature,* 207(4992):9–13.

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., and Ghadiri, M. R. (1996). A self-replicating peptide. *Nature,* 382(6591):525–528.

Léger, A., Selsis, F., Sotin, C., Guillot, T., Despois, D., Mawet, D., Ollivier, M., Labèque, A., Valette, C., Brachet, F., Chazelas, B., and Lammer, H. (2004). A new family of planets? "Ocean-planets." *Icarus,* 169(2):499–504.

Lehmer, O. R., Catling, D. C., Parenteau, M. N., and Hoehler, T. M. (2018). The productivity of oxygenic photosynthesis around cool, M dwarf stars. *Astrophys. J.,* 859(2):171.

Lem, S. (2020). *His Master's Voice*. Translated by M. Kandel. The MIT Press, Cambridge, MA.

Lemarchand, G. A. (1994). Passive and active SETI strategies using the synchronization of SN1987A. *Astrophys. Space Sci.,* 214(2):209–223.

Lenardic, A., Crowley, J. W., Jellinek, A. M., and Weller, M. (2016). The Solar system of forking paths: Bifurcations in planetary evolution and the search for life-bearing planets in our Galaxy. *Astrobiology,* 16(7):551–559.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. John Wiley & Sons, Inc., New York, NY.

Lennon, J. T. and Jones, S. E. (2011). Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nat. Rev. Microbiol.,* 9:119–130.

Lenormand, T., Engelstädter, J., Johnston, S. E., Wijnker, E., and Haag, C. R. (2016). Evolutionary mysteries in meiosis. *Phil. Trans. R. Soc. B,* 371(1706): 20160001.

Lenton, T. M., Boyle, R. A., Poulton, S. W., Shields-Zhou, G. A., and Butterfield, N. J. (2014). Co-evolution of eukaryotes and ocean oxygenation in the Neoproterozoic era. *Nat. Geosci.,* 7(4):257–265.

Lenton, T. M. (2020). On the use of models in understanding the rise of complex life. *Interface Focus,* 10(4):20200018.

Leopardi, G. (2013). *Zibaldone*. Edited by M. Caesar and F. D'Intino. Translated by K. Baldwin, R. Dixon, D. Gibbons, A. Goldstein, G. Slowey, M. Thom, and P. Williams. Farrar, Straus and Giroux, New York, NY.

Lesnikowski, A., Bickel, V. T., and Angerhausen, D. (2019, December 8–14). *Unsupervised distribution learning for lunar surface anomaly detection*. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Leung, M., Meadows, V. S., and Lustig-Yaeger, J. (2020). High-resolution spectral discriminants of ocean loss for M-dwarf terrestrial exoplanets. *Astron. J.,* 160(1):11.

Le Vay, K. and Mutschler, H. (2019). The difficult case of an RNA-only origin of life. *Emerg. Top. Life Sci.,* 3(5):469–475.

Levchenko, I., Xu, S., Mazouffre, S., Lev, D., Pedrini, D., Goebel, D., Garrigues, L., Taccogna, F., and Bazaka, K. (2020). Perspectives, frontiers, and new horizons for plasma-based space electric propulsion. *Phys. Plasmas,* 27(2):020601.

Lever, M. A., Rogers, K. L., Lloyd, K. G., Overmann, J., Schink, B., Thauer, R. K., Hoehler, T. M., and Jørgensen, B. B. (2015). Life under extreme energy limitation: A synthesis of laboratory- and field-based investigations. *FEMS Microbiol. Rev.,* 39(5):688–728.

Levin, G. V. and Straat, P. A. (2016). The case for extant life on Mars and its possible detection by the Viking Labeled Release experiment. *Astrobiology,* 16(10): 798–810.

Levin, N. E. (2015). Environment and climate of early human evolution. *Annu. Rev. Earth Planet. Sci.,* 43:405–429.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.,* 15(3):237–240.

Levins, R. and Lewontin, R. (1985). *The Dialectical Biologist*. Harvard University Press, Cambridge, MA.

Levy, M. and Miller, S. L. (1998). The stability of the RNA bases: Implications for the origin of life. *Proc. Natl. Acad. Sci. USA,* 95(14):7933–7938.

Lewis, S. L. and Maslin, M. A. (2015). Defining the Anthropocene. *Nature,* 519(7542):171–180.

Lewontin, R. (2000). *The Triple Helix: Gene, Organism, and Environment*. Harvard University Press, Cambridge, MA.

Li, G. and Batygin, K. (2014). On the spin-axis dynamics of a Moonless Earth. *Astrophys. J.,* 790(1):69.

Lichtenberg, T., Golabek, G. J., Burn, R., Meyer, M. R., Alibert, Y., Gerya, T. V., and Mordasini, C. (2019). A water budget dichotomy of rocky protoplanets from ^{26}Al-heating. *Nat. Astron.,* 3:307–313.

Lifson, S. (1997). On the crucial stages in the origin of animate matter. *J. Mol. Evol.,* 44(1):1–8.

Limaye, S. S., Mogul, R., Smith, D. J., Ansari, A. H., Słowik, G. P., and Vaishampayan, P. (2018). Venus' spectral signatures and the potential for life in the clouds. *Astrobiology,* 18(9):1181–1198.

Lin, H. W., Gonzalez Abad, G., and Loeb, A. (2014). Detecting industrial pollution in the atmospheres of Earth-like exoplanets. *Astrophys. J. Lett.,* 792(1):L7.

Lin, H. W. and Loeb, A. (2015). Statistical signatures of panspermia in exoplanet surveys. *Astrophys. J. Lett.,* 810(1):L3.

Lincoln, T. A. and Joyce, G. F. (2009). Self-sustained replication of an RNA enzyme. *Science,* 323(5918):1229–1232.

Lincowski, A. P., Lustig-Yaeger, J., and Meadows, V. S. (2019). Observing isotopologue bands in terrestrial exoplanet atmospheres with the James Webb Space Telescope—Implications for identifying past atmospheric and ocean loss. *Astron. J.,* 158(1):26.

Lineweaver, C. H. and Davis, T. M. (2002). Does the rapid appearance of life on earth suggest that life is common in the Universe? *Astrobiology,* 2(3): 293–304.

Lineweaver, C. H., Fenner, Y., and Gibson, B. K. (2004). The Galactic Habitable Zone and the age distribution of complex life in the Milky Way. *Science,* 303(5654):59–62.

Lingam, M. (2016a). Analytical approaches to modelling panspermia—Beyond the mean-field paradigm. *Mon. Not. R. Astron. Soc.,* 455(3):2792–2803.

Lingam, M. (2016b). Interstellar travel and Galactic colonization: Insights from percolation theory and the Yule process. *Astrobiology,* 16(6):418–426.

Lingam, M., Hirvijoki, E., Pfefferlé, D., Comisso, L., and Bhattacharjee, A. (2017). Nonlinear resistivity for magnetohydrodynamical models. *Phys. Plasmas,* 24(4):042120.

Lingam, M. and Loeb, A. (2017a). Enhanced interplanetary panspermia in the TRAPPIST-1 system. *Proc. Natl. Acad. Sci. USA,* 114(26):6689–6693.

Lingam, M. and Loeb, A. (2017b). Fast radio bursts from extragalactic light sails. *Astrophys. J. Lett.,* 837(2):L23.

Lingam, M. and Loeb, A. (2017c). Natural and artificial spectral edges in exoplanets. *Mon. Not. R. Astron. Soc. Lett.,* 470(1):L82-L86.

Lingam, M. and Loeb, A. (2017d). Reduced diversity of life around Proxima Centauri and TRAPPIST-1. *Astrophys. J. Lett.,* 846(2):L21.

Lingam, M. and Loeb, A. (2017e). Risks for life on habitable planets from superflares of their host stars. *Astrophys. J. Lett.,* 848(1):41.

Lingam, M., Dong, C., Fang, X., Jakosky, B. M., and Loeb, A. (2018). The propitious role of solar energetic particles in the origin of life. *Astrophys. J.,* 853(1):10.

Lingam, M. and Loeb, A. (2018a). Implications of captured interstellar objects for panspermia and extraterrestrial life. *Astron. J.,* 156(5):193.

Lingam, M. and Loeb, A. (2018b). Implications of tides for life on exoplanets. *Astrobiology,* 18(7):967–982.

Lingam, M. and Loeb, A. (2018c). Is extraterrestrial life suppressed on subsurface ocean worlds due to the paucity of bioessential elements? *Astron. J.,* 156(4):151.

Lingam, M. and Loeb, A. (2018d). Is life most likely around Sun-like stars? *J. Cosmol. Astropart. Phys.,* 05:020.

Lingam, M. and Loeb, A. (2018e). Limitations of chemical propulsion for interstellar escape from habitable zones around low-mass stars. *Res. Notes AAS,* 2(3):154.

Lingam, M. and Loeb, A. (2018f). Optimal target stars in the search for life. *Astrophys. J. Lett.,* 857(2):L17.

Lingam, M. and Loeb, A. (2018g). Physical constraints on the likelihood of life on exoplanets. *Int. J. Astrobiol.,* 17(2):116–126.

Lingam, M. (2019a). Interstellar travel and Galactic colonization: Insights from percolation theory and the Yule process. *Astrobiology,* 16(6):418–426.

Lingam, M. (2019b). Revisiting the biological ramifications of variations in Earth's magnetic field. *Astrophys. J. Lett.,* 874(2):L28.

Lingam, M. and Loeb, A. (2019a). Brown dwarf atmospheres as the potentially most detectable and abundant sites for life. *Astrophys. J.,* 883(2):143.

Lingam, M. and Loeb, A. (2019b). Dependence of biological activity on the surface water fraction of planets. *Astron. J.,* 157(1):25.

Lingam, M. and Loeb, A. (2019c). Photosynthesis on habitable planets around low-mass stars. *Mon. Not. R. Astron. Soc.,* 485(4):5924–5928.

Lingam, M. and Loeb, A. (2019d). Physical constraints for the evolution of life on exoplanets. *Rev. Mod. Phys.,* 91(2):021002.

Lingam, M. and Loeb, A. (2019e). Relative Likelihood of success in the search for primitive versus intelligent extraterrestrial life. *Astrobiology,* 19(1):28–39.

Lingam, M. and Loeb, A. (2019f). Role of stellar physics in regulating the critical steps for life. *Int. J. Astrobiol.,* 18(6):527–546.

Lingam, M. and Loeb, A. (2019g). Subsurface exolife. *Int. J. Astrobiol.,* 18(2): 112–141.

Lingam, M. (2020). Implications of abiotic oxygen buildup for Earth-like complex life. *Astron. J.,* 159(4):144.

Lingam, M. and Loeb, A. (2020a). Aquatic biospheres on temperate planets around Sun-like stars and M-dwarfs. arXiv e-prints. arXiv:2005.14387.

Lingam, M. and Loeb, A. (2020b). Constraints on aquatic photosynthesis for terrestrial planets around other stars. *Astrophys. J. Lett.,* 889(1):L15.

Lingam, M. and Loeb, A. (2020c). Electric sails are potentially more effective than light sails near most stars. *Acta Astronaut.,* 168:146–154.

Lingam, M. and Loeb, A. (2020d). On the habitable lifetime of terrestrial worlds with high radionuclide abundances. *Astrophys. J. Lett.,* 889(1):L20.

Lingam, M. and Loeb, A. (2020e). Photosynthesis on exoplanets and exomoons from reflected light. *Int. J. Astrobiol.,* 19(3):210–219.

Lingam, M. and Loeb, A. (2020f). Propulsion of spacecrafts to relativistic speeds using natural astrophysical sources. *Astrophys. J.,* 894(1):36.

Lingam, M. and Loeb, A. (2020g). What's in a name: The etymology of astrobiology. *Int. J. Astrobiol.,* 19(5):379–385.

Lingam, M. and Loeb, A. (2020h). Potential for liquid water biochemistry deep under the surfaces of the Moon, Mars and beyond. *Astrophys. J. Lett.,* 901(1):L11.

Lingam, M. and Loeb, A. (2020i). Constraints on the abundance of $0.01c$ stellar engines in the Milky Way. *Astrophys. J.,* 905(2):175.

Lingappa, U. F., Monteverde, D. R., Magyar, J. S., Valentine, J. S., and Fischer, W. W. (2019). How manganese empowered life with dioxygen (and vice versa). *Free Radic. Biol. Med.,* 140:113–125.

Linsky, J. (2019). *Host Stars and Their Effects on Exoplanet Atmospheres: An Introductory Overview*. Springer, Cham, Switzerland.

Lipman, D., Isaacson, H., Siemion, A. P. V., Lebofsky, M., Price, D. C., MacMahon, D., Croft, S., DeBoer, D., Hickish, J., Werthimer, D., Hellbourg, G., Enriquez, J. E., and Gizani, N. (2019). The Breakthrough Listen search for intelligent

life: Searching Boyajian's star for laser line emission. *Publ. Astron. Soc. Pac.,* 131(997):034202.

Lissauer, J. J., Barnes, J. W. and Chambers, J. E. (2012). Obliquity variations of a moonless Earth. *Icarus,* 217(1):77–87.

Lisse, C. M., Desch, S. J., Unterborn, C. T., Kane, S. R., Young, P. R., Hartnett, H. E., Hinkel, N. R., Shim, S.-H., Mamajek, E. E., and Izenberg, N. R. (2020). A geologically robust procedure for observing rocky exoplanets to ensure that detection of atmospheric oxygen is a modern Earth-like biosignature. *Astrophys. J. Lett.,* 898(1):L17.

Liu, Z., Wu, L.-F., Xu, J., Bonfio, C., Russell, D. A., and Sutherland, J. D. (2020). Harnessing chemical energy for the activation and joining of prebiotic building blocks. *Nat. Chem.,* 12(11):1023–1028.

Livio, M. (1999). How rare are extraterrestrial civilizations, and when did they emerge? *Astrophys. J.,* 511(1):429–431.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. USA,* 113(21):5970–5975.

Loeb, A. and Zaldarriaga, M. (2007). Eavesdropping on radio broadcasts from Galactic civilizations with upcoming observatories for redshifted 21 cm radiation. *J. Cosmol. Astropart. Phys.,* 01:020.

Loeb, A. and Turner, E. L. (2012). Detection technique for artificially illuminated objects in the outer Solar system and beyond. *Astrobiology,* 12(4):290–294.

Loeb, A. and Furlanetto, S. (2013). *The First Galaxies in the Universe*. Princeton University Press, Princeton, NJ.

Loeb, A., Batista, R. A., and Sloan, D. (2016). Relative likelihood for life as a function of cosmic time. *J. Cosmol. Astropart. Phys.,* 08:040.

Loeb, J. (1912). *The Mechanistic Conception of Life: Biological Essays*. The University of Chicago Press, Chicago, IL.

Lombard, J., López-García, P., and Moreira, D. (2012). The early evolution of lipid membranes and the three domains of life. *Nat. Rev. Microbiol.,* 10(7):507–515.

Long, K. F., Obousy, R. K., and Hein, A. (2011). Project Icarus: Optimisation of nuclear fusion propulsion for interstellar missions. *Acta Astronaut.,* 68: 1820–1829.

Long, K. F. (2012). *Deep Space Propulsion: A Roadmap to Interstellar Flight*. Springer, New York, NY.

Loper, R. D. (2019). Carrington-class events as a great filter for electronic civilizations in the Drake equation. *Publ. Astron. Soc. Pac.,* 131(998):044202.

Lopez, J. V., Peixoto, R. S., and Rosado, A. S. (2019). Inevitable future: space colonization beyond Earth with microbes first. *FEMS Microbiol. Ecol.,* 95(10):fiz127.

López-García, P. and Moreira, D. (2015). Open questions on the origin of eukaryotes. *Trends Ecol. Evol.,* 30(11):697–708.

López-García, P., Eme, L., and Moreira, D. (2017). Symbiosis in eukaryotic evolution. *J. Theor. Biol.,* 434:20–33.

López-García, P. and Moreira, D. (2020). The syntrophy hypothesis for the origin of eukaryotes revisited. *Nat. Microbiol.,* 5(5):655–667.

Lorenz, R. D. (2019). *Exploring Planetary Climate: A History of Scientific Discovery on Earth, Mars, Venus and Titan*. Cambridge University Press, Cambridge, UK.

Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol.,* 17(2):e3000106.

Lovelock, J. E. (1965). A physical basis for life detection experiments. *Nature,* 207(4997):568–570.

Lovelock, J. (2000). *The Ages of Gaia: A Biography of Our Living Earth*. Oxford University Press, Oxford, UK (2nd edition).

Lovis, C. and Fischer, D. (2010). Radial velocity techniques for exoplanets. In Seager, S., editor, *Exoplanets,* pages 27–53. The University of Arizona Press, Tucson, AZ.

Lovley, D. R. (2017). Happy together: Microbial communities that hook up to swap electrons. *ISME J.,* 11(2):327–336.

Lugaro, M., Ott, U., and Kereszturi, Á. (2018). Radioactive nuclei from cosmochronology to habitability. *Prog. Part. Nucl. Phys.,* 102:1–47.

Lubin, P. (2016a). Implications of directed energy for SETI. *Proceedings Volume 9981, Planetary Defense and Space Environment Applications,* 99810H.

Lubin, P. (2016b). A roadmap to interstellar flight. *J. Br. Interplanet. Soc.,* 69:40–72.

Luger, R. and Barnes, R. (2015). Extreme water loss and abiotic O_2 buildup on planets throughout the habitable zones of M dwarfs. *Astrobiology,* 15(2):119–143.

Luger, R., Lustig-Yaeger, J., Fleming, D. P., Tilley, M. A., Agol, E., Meadows, V. S., Deitrick, R., and Barnes, R. (2017). The pale green dot: A method to characterize Proxima Centauri b using exo-aurorae. *Astrophys. J.,* 837:63.

Luhman, K. L. (2014). Discovery of a $\sim$250 K brown dwarf at 2 pc from the Sun. *Astrophys. J. Lett.,* 786(2):L18.

Luisi, P. L. (2016). *The Emergence of Life: From Chemical Origins to Synthetic Biology*. Cambridge University Press, Cambridge, UK (2nd edition).

Lunine, J. I. (2017). Ocean worlds exploration. *Acta Astronaut.,* 131:123–130.

Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature,* 476(7361):442–445.

Luo, Z.-X., Gatesy, S. M., Jenkins, F. A., Amaral, W. W., and Shubin, N. H. (2015). Mandibular and dental characteristics of Late Triassic mammaliaform Haramiyavia and their ramifications for basal mammal evolution. *Proc. Natl. Acad. Sci. USA,* 112(51):E7101–E7109.

Lustig-Yaeger, J., Meadows, V. S., and Lincowski, A. P. (2019). The detectability and characterization of the TRAPPIST-1 exoplanet atmospheres with JWST. *Astron. J.,* 158(1):27.

Lynch, M. and Marinov, G. K. (2017). Membranes, energetics, and evolution across the prokaryote-eukaryote divide. *eLife,* 6:e20437.

Lyons, N. A. and Kolter, R. (2015). On the evolution of bacterial multicellularity. *Curr. Opin. Microbiol.,* 24:21–28.

Lyons, T. W., Reinhard, C. T., and Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature,* 506(7488):307–315.

Macchi, A., Borghesi, M., and Passoni, M. (2013). Ion acceleration by superintense laser-plasma interaction. *Rev. Mod. Phys.,* 85(2):751–793.

Maccone, C. (2009). *Deep Space Flight and Communications: Exploiting the Sun as a Gravitational Lens*. Springer-Praxis, Chichester, UK.

Maccone, C. (2010). The statistical Drake Equation. *Acta Astronaut.,* 67(11):1366–1383.

Macdonald, M. and McInnes, C. (2011). Solar sail science mission applications and advancement. *Adv. Space Res.,* 48(11):1702–1716.

Machery, E. (2012). Why I stopped worrying about the definition of life . . . and why you should as well. *Synthese,* 185(1):145–164.

Madhusudhan, N. (2019). Exoplanetary atmospheres: Key insights, challenges and prospects. *Annu. Rev. Astron. Astrophys.,* 57:617–663.

Maehara, H., Shibayama, T., Notsu, S., Notsu, Y., Nagao, T., Kusaba, S., Honda, S., Nogami, D., and Shibata, K. (2012). Superflares on solar-type stars. *Nature,* 485(7399):478–481.

Maehara, H., Shibayama, T., Notsu, Y., Notsu, S., Honda, S., Nogami, D., and Shibata, K. (2015). Statistical properties of superflares on solar-type stars based on 1-min cadence data. *Earth Planets Space,* 67:59.

Magnabosco, C., Lin, L.-H., Dong, H., Bomberg, M., Ghiorse, W., Stan-Lotter, H., Pedersen, K., Kieft, T. L., van Heerden, E., and Onstott, T. C. (2018). The biomass and biodiversity of the continental subsurface. *Nat. Geosci.,* 11(10):707–717.

Magurran, A. E. (2004). *Measuring Biological Diversity*. Blackwell Publishing, Malden, MA.

Maire, J., Wright, S. A., Dorval, P., Drake, F. D., Duenas, A., Isaacson, H., Marcy, G. W., Siemion, A., Stone, R. P. S., Tallis, M., Treffers, R. R., and Werthimer, D. (2016). A near-infrared SETI experiment: Commissioning, data analysis, and performance results. *Proceedings Volume 9908, Ground-based and Airborne Instrumentation for Astronomy VI,* 990810.

Maire, J., Wright, S. A., Barrett, C. T., Dexter, M. R., Dorval, P., Duenas, A., Drake, F. D., Hultgren, C., Isaacson, H. Marcy, G. W., Meyer, E., Ramos, J. R., Shirman, N., Siemion, A., Stone, R. P. S., Tallis, M., Tellis, N. K.,

Treffers, R. R., and Werthimer, D. (2019). Search for nanosecond near-infrared transients around 1280 celestial objects. *Astron. J.,* 158(5):203.

Makovetskii, P. V. (1978). Effectiveness of linking the beacons of extraterrestrial civilizations to natural phenomena. *Radiofizika,* 21(1):139–141.

Makukov, M. A. and Shcherbak, V. I. (2018). SETI in vivo: Testing the we-are-them hypothesis. *Int. J. Astrobiol.,* 17(2):127–146.

Mall, A., Sobotta, J., Huber, C., Tschirner, C., Kowarschik, S., Bačnik, K., Mergelsberg, M., Boll, M., Hügler, M., Eisenreich, W., and Berg, I. A. (2018). Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium. *Science,* 359(6375):563–567.

Manchester, Z. and Loeb, A. (2017). Stability of a light sail riding on a laser beam. *Astrophys. J. Lett.,* 837(2):L20.

Mann, J. and Patterson, E. M. (2013). Tool use by aquatic animals. *Phil. Trans. R. Soc. B,* 368(1630):20120424.

Manning, P. (2020). *A History of Humanity: The Evolution of the Human System*. Cambridge University Press, Cambridge, UK.

Margot, J.-L., Croft, S., Lazio, T. J. W., Tarter, J., and Korpela, E. J. (2019). The radio search for technosignatures in the decade 2020–2030. *Bull. Am. Astron. Soc.,* 51(3):298.

Marino, L. and Colvin, C. M. (2015). Thinking pigs: A comparative review of cognition, emotion, and personality in *Sus domesticus*. *Int. J. Comp. Psychol.,* 28:23859.

Mariscal, C., Barahona, A., Aubert-Kato, N., Aydinoglu, A. U., Bartlett, S., Cárdenas, M. L., Chandru, K., Cleland, C., Cocanougher, B. T., Comfort, N., Cornish-Bowden, A., Deacon, T., Froese, T., Giovannelli, D., Hernlund, J., Hut, P., Kimura, J., Maurel, M.-C., Merino, N., . . . Cleaves, H. J. (2019). Hidden concepts in the history and philosophy of origins-of-life studies: A workshop report. *Orig. Life Evol. Biosph.,* 49(3):111–145.

Mariscal, C. and Doolittle, W. F. (2020). Life and life only: A radical alternative to life definitionism. *Synthese,* 197:2975–2989.

Marley, M. S. and Robinson, T. D. (2015). On the cool side: Modeling the atmospheres of brown dwarfs and giant planets. *Annu. Rev. Astron. Astrophys.,* 53:279–323.

Marshall, C. R. (2006). Explaining the Cambrian "explosion" of animals. *Annu. Rev. Earth Planet. Sci.,* 34:355–384.

Marshall, S. M., Murray, A. R. G., and Cronin, L. (2017). A probabilistic framework for identifying biosignatures using pathway complexity. *Phil. Trans. R. Soc. A,* 375(2109):20160342.

Martel, J., Young, D., Peng, H.-H., Wu, C.-Y., and Young, J. D. (2012). Biomimetic properties of minerals and the search for life in the Martian meteorite ALH84001. *Annu. Rev. Earth Planet. Sci.,* 40:167–193.

Martijn, J. and Ettema, T. J. G. (2013). From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem. Soc. Trans.,* 41(1):451–457.

Martin, A. and McMinn, A. (2018). Sea ice, extremophiles and life on extraterrestrial ocean worlds. *Int. J. Astrobiol.,* 17(1):1–16.

Martin, A. R. (1973). Magnetic intake limitations on interstellar ramjets. *Astronautica Acta,* 18:1–10.

Martin, A. R. and Bond, A. (1979). Nuclear pulse propulsion: A historical review of an advanced propulsion concept. *J. Br. Interplanet. Soc.,* 32:283–310.

Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature,* 392(6671):37–41.

Martin, W., Baross, J., Kelley, D., and Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.,* 6(11):805–814.

Martin, W. F., Garg, S., and Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Phil. Trans. R. Soc. B,* 370(1678):20140330.

Martin, W. F., Tielens, A. G. M., Mentel, M., Garg, S. G., and Gould, S. B. (2017). The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol. Mol. Biol. Rev.,* 81(3):e00008–17.

Martin, W. F., Bryant, D. A., and Beatty, J. T. (2018). A physiological perspective on the origin and evolution of photosynthesis. *FEMS Microbiol. Rev.,* 42(2): 205–231.

Martin, W. F. (2020). Older than genes: The acetyl CoA pathway and origins. *Front. Microbiol.,* 11:817.

Maruyama, S., Ikoma, M., Genda, H., Hirose, K., Yokoyama, T., and Santosh, M. (2013). The naked planet Earth: Most essential pre-requisite for the origin and evolution of life. *Geosci. Front.,* 4(2):141–165.

Maruyama, S., Santosh, M., and Azuma, S. (2018). Initiation of plate tectonics in the Hadean: Eclogitization triggered by the ABEL Bombardment. *Geosci. Front.,* 9(4):1033–1048.

Mashian, N. and Loeb, A. (2016). CEMP stars: Possible hosts to carbon planets in the early Universe. *Mon. Not. R. Astron. Soc.,* 460(3):2482–2491.

Maslin, M. A., Shultz, S., and Trauth, M. H. (2015). A synthesis of the theories and concepts of early human evolution. *Phil. Trans. R. Soc. B,* 370(1663):20140064.

Mast, C. B., Schink, S., Gerland, U., and Braun, D. (2013). Escalation of polymerization in a thermal gradient. *Proc. Natl. Acad. Sci. USA,* 110(20):8030–8035.

Mastrapa, R. M. E., Glanzberg, H., Head, J. N., Melosh, H. J., and Nicholson, W. L. (2001). Survival of bacteria exposed to extreme acceleration: Implications for panspermia. *Earth Planet. Sci. Lett.,* 189:1–8.

Matloff, G. L. (2005). *Deep Space Probes: To the Outer Solar System and Beyond*. Praxis Publishing, Chichester, UK (2nd edition).

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Co., Dordecht, Netherlands.

Maurel, M.-C. and Leclerc, F. (2016). From foundation stones to life: Concepts and results. *Elements,* 12(6):407–412.

Mautner, M. N. (1997). Directed panspermia. 3. strategies and motivation for seeding star-forming clouds. *J. Br. Interplanet. Soc.,* 50:93–102.

Mayor, M. and Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature,* 378(6555):355–359.

Mayr, E. (1985). The probability of extraterrestrial intelligent life. In Regis, E., editor, *Extraterrestrials: Science and Alien Intelligence,* pages 23–30. Cambridge University Press, Cambridge, UK.

Mazin, B., Artigau, E., Bailey, V., Baranec, C., Beichman, C., Benneke, B., Birkby, J., Brandt, T., Chilcote, J., Chun, M., Close, L., Currie, T., Crossfield, I., Dekany, R., Delorme, J. R., Dong, C., Dong, R., Doyon, R., Dressing, C., . . . Wright, S. (2019). Directly imaging rocky planets from the ground. *Bull. Am. Astron. Soc.,* 51(3):128.

Mazouffre, S. (2016). Electric propulsion for satellites and spacecraft: Established technologies and novel approaches. *Plasma Sources Sci. Technol.,* 25(3):033002.

McBrearty, S. and Brooks, A. S. (2000). The revolution that wasn't: A new interpretation of the origin of modern human behavior. *J. Hum. Evol.,* 39(5): 453–563.

McCabe, M. and Lucas, H. (2010). On the origin and evolution of life in the Galaxy. *Int. J. Astrobiol.,* 9(4):217–226.

McCollom, T. M. (2013). Miller-Urey and beyond: What have we learned about prebiotic organic synthesis reactions in the past 60 years? *Annu. Rev. Earth Planet. Sci.,* 41:207–229.

McCoy, D. E., Schiestl, M., Neilands, P., Hassall, R., Gray, R. D., and Taylor, A. H. (2019). New Caledonian crows behave optimistically after using tools. *Curr. Biol.,* 29(16):2737–2742.

McGhee, G. (2011). *Convergent Evolution: Limited Forms Most Beautiful.* The MIT Press, Cambridge, MA.

McInnes, C. R. (1999). *Solar Sailing: Technology, Dynamics and Mission Applications.* Praxis Publishing Ltd., Chichester, UK.

McKay, C. P. and Borucki, W. J. (1997). Organic synthesis in experimental impact shocks. *Science,* 276(5311):390–392.

McKay, C. P. (2004). What is life–and how do we search for it in other worlds? *PLoS Biol.,* 2(9):e302.

McKay, C. P. (2010). An origin of life on Mars. *Cold Spring Harb. Perspect. Biol.,* 2(4):a003509.

McKay, C. P. (2014). Requirements and limits for life in the context of exoplanets. *Proc. Natl. Acad. Sci. USA,* 111(35):12628–12633.

McKay, D. S., Gibson, E. K., Thomas-Keprta, K. L., Vali, H., Romanek, C. S., Clemett, S. J., Chillier, X. D. F., Maechling, C. R., and Zare, R. N. (1996).

Search for past life on Mars: Possible relic biogenic activity in Martian meteorite ALH84001. *Science,* 273(5277):924–930.

McPherron, S. P., Alemseged, Z., Marean, C. W., Wynn, J. G., Reed, D., Geraads, D., Bobe, R., and Béarat, H. A. (2010). Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. *Nature,* 466(7308):857–860.

McPherson, R. A. (2007). A review of vegetation-atmosphere interactions and their influences on mesoscale phenomena. *Prog. Phys. Geogr.,* 31(3):261–285.

McTier, M. A. S., Kipping, D. M., and Johnston, K. (2020). Eight in ten stars in the Milky Way Bulge experience stellar encounters within 1000 AU in a gigayear. *Mon. Not. R. Astron. Soc.,* 495(2):2105–2111.

Meadows, V. S., Reinhard, C. T., Arney, G. N., Parenteau, M. N., Schwieterman, E. W., Domagal-Goldman, S. D., Lincowski, A. P., Stapelfeldt, K. R., Rauer, H., DasSarma, S., Hegde, S., Narita, N., Deitrick, R., Lustig-Yaeger, J., Lyons, T. W., Siegler, N., and Grenfell, J. L. (2018). Exoplanet biosignatures: Understanding oxygen as a biosignature in the context of its environment. *Astrobiology,* 18(6):630–662.

Meadows, V. S., Arney, G. N., Schmidt, B. E., and Des Marais, D. J. (Eds.) (2020). *Planetary astrobiology*. The University of Arizona Press, Tucson, AZ.

Meech, K. J., Weryk, R., Micheli, M., Kleyna, J. T., Hainaut, O. R., Jedicke, R., Wainscoat, R. J., Chambers, K. C., Keane, J. V., Petric, A., Denneau, L., Magnier, E., Berger, T., Huber, M. E., Flewelling, H., Waters, C., Schunova-Lilly, E., and Chastel, S. (2017). A brief visit from a red and extremely elongated interstellar asteroid. *Nature,* 552(7685):378–381.

Melosh, H. J. (1984). Impact ejection, spallation, and the origin of meteorites. *Icarus,* 59(2):234–260.

Melosh, H. J. (2003). Exchange of meteorites (and life?) between stellar systems. *Astrobiology,* 3(1):207–215.

Ménez, B., Pisapia, C., Andreani, M., Jamme, F., Vanbellingen, Q. P., Brunelle, A., Richard, L., Dumas, P., and Réfrégiers, M. (2018). Abiotic synthesis of amino acids in the recesses of the oceanic lithosphere. *Nature,* 564(7734):59–63.

Menor-Salván, C. and Marín-Yaseli, Y. (2012). Prebiotic chemistry in eutectic solutions at the water-ice matrix. *Chem. Soc. Rev.,* 41(16):5404–5415.

Menor-Salván, C. (Ed.). (2018). *Prebiotic Chemistry and Chemical Evolution of Nucleic Acids*. Springer, Cham, Switzerland.

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J., Burk-Herrick, A., Westerman, M., . . . Murphy, W. J. (2011). Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science,* 334(6055):521–524.

Mereschkowsky, C. (1905). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Zentbl.,* 25:593–604.

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S., and Giovannelli, D. (2019). Living at the extremes: Extremophiles and the limits of life in a planetary context. *Front. Microbiol.,* 10:780.

Meshik, A. P. (2005). The workings of an ancient nuclear reactor. *Sci. Am.,* 293(5):82–91.

Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. The University of Chicago Press, Chicago, IL.

Mesoudi, A. and Thornton, A. (2018). What is cumulative cultural evolution? *Proc. R. Soc. B,* 285(1880):20180712.

Messerschmitt, D. G. (2015). Design for minimum energy in interstellar communication. *Acta Astronaut.,* 107:20–39.

Messerschmitt, D. G., Lubin, P., and Morrison, I. (2020). Challenges in scientific data communication from low-mass interstellar probes. *Astrophys. J. Suppl. Ser.,* 249(2):36.

Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J. M. B., Carbonell, E., Viola, B., Kelso, J., Prüfer, K., and Pääbo, S. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature,* 531(7595): 504–507.

Mileikowsky, C., Cucinotta, F. A., Wilson, J. W., Gladman, B., Horneck, G., Lindegren, L., Melosh, J., Rickman, H., Valtonen, M., and Zheng, J. Q. (2000a). Natural transfer of viable microbes in space. 1. From Mars to Earth and Earth to Mars. *Icarus,* 145(2):391–427.

Mileikowsky, C., Cucinotta, F. A., Wilson, J. W., Gladman, B., Horneck, G., Lindegren, L., Melosh, J., Rickman, H., Valtonen, M., and Zheng, J. Q. (2000b). Risks threatening viable transfer of microbes between bodies in our Solar system. *Planet. Space Sci.,* 48(11):1107–1115.

Miller, G. (2000). *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. Doubleday, New York, NY.

Miller, S. L. (1953). A production of amino acids under possible primitive Earth conditions. *Science,* 117(3046):528–529.

Mills, D. B., Francis, W. R., and Canfield, D. E. (2018). Animal origins and the Tonian Earth system. *Emerging Topics in Life Sciences,* 2(2):289–298.

Mills, D. B. (2020). The origin of phagocytosis in Earth history. *Interface Focus,* 10(4):20200019.

Mills, D. R., Peterson, R. L., and Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA,* 58(1):217–224.

Milshteyn, D., Damer, B., Havig, J., and Deamer, D. (2018). Amphiphilic compounds assemble into membranous vesicles in hydrothermal hot spring water but not in seawater. *Life,* 8(2):11.

Mirzaghaderi, G. and Hörandl, E. (2016). The evolution of meiotic sex and its alternatives. *Proc. R. Soc. B,* 283(1838):20161221.

Mitchell, P. (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature,* 191(4784):144–148.

Mithen, S. (2006). *The Singing Neanderthals: The Origins of Music, Language, Mind, and Body*. Harvard University Press, Cambridge, MA.

Mix, L. J. (2015). Defending definitions of life. *Astrobiology,* 15(1):15–19.

Miyakawa, S., Cleaves, H. J., and Miller, S. L. (2002). The cold origin of life: A. Implications based on the hydrolytic stabilities of hydrogen cyanide and formamide. *Orig. Life Evol. Biosph.,* 32(3):195–208.

Miyake, F., Nagaya, K., Masuda, K., and Nakamura, T. (2012). A signature of cosmic-ray increase in AD 774–775 from tree rings in Japan. *Nature,* 486(7402):240–242.

Mizuuchi, R., Blokhuis, A., Vincent, L., Nghe, P., Lehman, N., and Baum, D. (2019). Mineral surfaces select for longer RNA molecules. *Chem. Commun.,* 55(14):2090–2093.

Moger-Reischer, R. Z. and Lennon, J. T. (2019). Microbial ageing and longevity. *Nat. Rev. Microbiol.,* 17(11):679–690.

Moissl-Eichinger, C., Cockell, C., and Rettberg, P. (2016). Venturing into new realms? Microorganisms in space. *FEMS Microbiol. Rev.,* 40(5):722–737.

Mollière, P. and Snellen, I. A. G. (2019). Detecting isotopologues in exoplanet atmospheres using ground-based high-dispersion spectroscopy. *Astron. Astrophys.,* 622:A139.

Monod, J. (1971). *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. Translated by A. Wainhouse. Knopf, New York, NY.

Monteux, J., Golabek, G. J., Rubie, D. C., Tobie, G., and Young, E. D. (2018). Water and the interior structure of terrestrial planets and icy bodies. *Space Sci. Rev.,* 214(1):39.

Moore, W. B., Lenardic, A., Jellinek, A. M., Johnson, C. L., Goldblatt, C., and Lorenz, R. D. (2017). How habitable zones and super-Earths lead us astray. *Nat. Astron.,* 1:0043.

Morasch, M., Liu, J., Dirscherl, C. F., Ianeselli, A., Kühnlein, A., Le Vay, K., Schwintek, P., Islam, S., Corpinot, M. K., Scheu, B., Dingwell, D. B., Schwille, P., Mutschler, H., Powner, M. W., Mast, C. B., and Braun, D. (2019). Heated gas bubbles enrich, crystallize, dry, phosphorylate and encapsulate prebiotic molecules. *Nat. Chem.,* 11(9):779–788.

Morbidelli, A., Lunine, J. I., O'Brien, D. P., Raymond, S. N., and Walsh, K. J. (2012). Building terrestrial planets. *Annu. Rev. Earth Planet. Sci.,* 40:251–275.

Morley, C. V., Fortney, J. J., Marley, M. S., Zahnle, K., Line, M., Kempton, E., Lewis, N., and Cahoy, K. (2015). Thermal emission and reflected light spectra of super Earths with flat transmission spectra. *Astrophys. J.,* 815(2):110.

Morley, C. V., Kreidberg, L., Rustamkulov, Z., Robinson, T., and Fortney, J. J. (2017). Observing the atmospheres of known temperate Earth-sized planets with JWST. *Astrophys. J.,* 850(2):121.

Morley, C. V., Skemer, A. J., Allers, K. N., Marley, M. S., Faherty, J. K., Visscher, C., Beiler, S. A., Miles, B. E., Lupu, R., Freedman, R. S., Fortney, J. J., Geballe, T. R., and Bjoraker, G. L. (2018). An L band spectrum of the coldest brown dwarf. *Astrophys. J.,* 858(2):97.

Morono, Y., Ito, M., Hoshino, T., Terada, T., Hori, T., Ikehara, M., D'Hondt, S., and Inagaki, F. (2020). Aerobic microbial life persists in oxic marine sediment as old as 101.5 million years. *Nat. Commun.,* 11:3626.

Morowitz, H. and Sagan, C. (1967). Life in the clouds of Venus? *Nature,* 215(5107):1259–1260.

Morris, B. E. L., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial syntrophy: Interaction for the common good. *FEMS Microbiol. Rev.,* 37(3):384–406.

Morris, S. C. (2003). *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge University Press, Cambridge, UK.

Morris, S. C. (2015). *The Runes of Evolution: How the Universe Became Self-Aware*. Templeton Press, West Conshohocken, PA.

Morrison, P., Billingham, J., and Wolfe, J. (1979). The search for extraterrestrial intelligence—SETI. *Acta Astronaut.,* 6(1):11–31.

Moschou, S.-P., Drake, J. J., Cohen, O., Alvarado-Gómez, J. D., Garraffo, C., and Fraschetti, F. (2019). The stellar CME-flare relation: What do historic observations reveal? *Astrophys. J.,* 877(2):105.

Motsch, S., Tremmel, P., and Richert, C. (2020). Regioselective formation of RNA strands in the absence of magnesium ions. *Nucleic Acids Res.,* 48(3):1097–1107.

Muchowska, K. B., Varma, S. J., Chevallot-Beroux, E., Lethuillier-Karl, L., Li, G., and Moran, J. (2017). Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol.,* 1:1716–1721.

Muchowska, K. B., Varma, S. J., and Moran, J. (2019). Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature,* 569(7754):104–107.

Muchowska, K. B., Varma, S. J., and Moran, J. (2020). Nonenzymatic metabolic reactions and life's origins. *Chem. Rev.*, 120(15):7708–7744.

Mulkidjanian, A. Y. and Junge, W. (1997). On the origin of photosynthesis as inferred from sequence analysis. *Photosynth. Res.,* 51(1):27–42.

Mulkidjanian, A. Y., Bychkov, A. Y., Dibrova, D. V., Galperin, M. Y., and Koonin, E. V. (2012). Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl. Acad. Sci. USA,* 109(14):E821–E830.

Mullan, D. J. and Bais, H. P. (2018). Photosynthesis on a planet orbiting an M dwarf: Enhanced effectiveness during flares. *Astrophys. J.,* 865(2):101.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res.,* 1(1):2–9.

Muller, M. N., Wrangham, R. W., and Pilbeam, D. R. (Eds.) (2017). *Chimpanzees and Human Evolution*. Harvard University Press, Cambridge, MA.

Muñoz Caro, G. M., Meierhenrich, U. J., Schutte, W. A., Barbier, B., Arcones Segovia, A., Rosenbauer, H., Thiemann, W. H.-P., Brack, A., and Greenberg, J. M. (2002). Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature,* 416(6879):403–406.

Murray, C. D. and Dermott, S. F. (1999). *Solar System Dynamics*. Cambridge University Press, Cambridge, UK.

Muscente, A. D., Boag, T. H., Bykova, N., and Schiffbauer, J. D. (2018). Environmental disturbance, resource availability, and biologic turnover at the dawn of animal life. *Earth-Sci. Rev.,* 177:248–264.

Mutschler, H., Wochner, A., and Holliger, P. (2015). Freeze-thaw cycles as drivers of complex ribozyme assembly. *Nat. Chem.,* 7(6):502–508.

Myilswamy, K. V., Krishnan, A., and Povinelli, M. L. (2020). Photonic crystal lightsail with nonlinear reflectivity for increased stability. *Opt. Express,* 28(6):8223–8232.

Nadell, C. D., Drescher, K., and Foster, K. R. (2016). Spatial structure, cooperation and competition in biofilms. *Nat. Rev. Microbiol.,* 14:589–600.

Nam, I., Nam, H. G., and Zare, R. N. (2018). Abiotic synthesis of purine and pyrimidine ribonucleosides in aqueous microdroplets. *Proc. Natl. Acad. Sci. USA,* 115(1):36–40.

Napier, W. M. (2004). A mechanism for interstellar panspermia. *Mon. Not. R. Astron. Soc.,* 348(1):46–51.

Narusawa, S.-y., Aota, T., and Kishimoto, R. (2018). Which colors would extraterrestrial civilizations use to transmit signals? The "magic wavelengths" for optical SETI. *New. Astron.,* 60:61–64.

NASA Astrobiology Strategy (2016). Executive summary. *Astrobiology,* 16(8): 654–656.

Nelson, J. P. (2020). Mythic forecasts: Researcher portrayals of extraterrestrial life discovery. *Int. J. Astrobiol.,* 19(1):16–24.

Nelson, K. E., Levy, M., and Miller, S. L. (2000). Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proc. Natl. Acad. Sci. USA,* 97(8):3868–3871.

Neubauer, S., Hublin, J.-J., and Gunz, P. (2018). The evolution of modern human brain shape. *Sci. Adv.,* 4(1):eaao5961.

Neveu, M., Hays, L. E., Voytek, M. A., New, M. H., and Schulte, M. D. (2018). The ladder of life detection. *Astrobiology,* 18(11):1375–1402.

Newman, W. I. and Sagan, C. (1981). Galactic civilizations - Population dynamics and interstellar diffusion. *Icarus,* 46(3):293–327.

Ngwira, C. M., Pulkkinen, A., Kuznetsova, M. M., and Glocer, A. (2014). Modeling extreme "Carrington-type" space weather events using three-dimensional global MHD simulations. *J. Geophys. Res. Space Phys.,* 119(6):4456–4474.

Nicholson, A. and Forgan, D. (2013). Slingshot dynamics for self-replicating probes and the effect on exploration timescales. *Int. J. Astrobiol.,* 12(4):337–344.

Nicholson, W. L. (2009). Ancient micronauts: Interplanetary transport of microbes by cosmic impacts. *Trends Microbiol.,* 17(6):243–250.

Nielsen, P. E., Egholm, M., Berg, R. H., and Buchardt, O. (1991). Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science,* 254(5037):1497–1500.

Niklas, K. J. and Newman, S. A. (2013). The origins of multicellular organisms. *Evol. Dev.,* 15(1):41–52.

Niklas, K. J. and Newman, S. A. (Eds.) (2016). *Multicellularity: Origins and Evolution*. The MIT Press, Cambridge, MA.

Nimmo, F. and Pappalardo, R. T. (2016). Ocean worlds in the outer solar system. *J. Geophys. Res. Planets,* 121(8):1378–1399.

Nishino, Y. and Seto, N. (2018). The search for extra-Galactic intelligence signals synchronized with binary neutron star mergers. *Astrophys. J. Lett.,* 862(2):L21.

Niven, L. (1970). *Ringworld*. Ballantine Books, New York, NY.

Noack, L., Höning, D., Rivoldini, A., Heistracher, C., Zimov, N., Journaux, B., Lammer, H., Van Hoolst, T., and Bredehöft, J. H. (2016). Water-rich planets: How habitable is a water layer deeper than on Earth? *Icarus,* 277:215–236.

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M.-N., and Jenner, K. C. S. (2000). Cultural revolution in whale songs. *Nature,* 408(6812):537.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

Nowak, M. A. and Ohtsuki, H. (2008). Prevolutionary dynamics and the origin of evolution. *Proc. Natl. Acad. Sci. USA,* 105(39):14924–14927.

Nowak, H., Schneebeli-Hermann, E., and Kustatscher, E. (2019). No mass extinction for land plants at the Permian-Triassic transition. *Nat. Commun.,* 10:384.

Nunoura, T., Chikaraishi, Y., Izaki, R., Suwa, T., Sato, T., Harada, T., Mori, K., Kato, Y., Miyazaki, M., Shimamura, S., Yanagawa, K., Shuto, A., Ohkouchi, N., Fujita, N., Takaki, Y., Atomi, H., and Takai, K. (2018). A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science,* 359(6375):559–563.

Nürnberg, D. J., Morton, J., Santabarbara, S., Telfer, A., Joliot, P., Antonaru, L. A., Ruban, A. V., Cardona, T., Krausz, E., Boussac, A., Fantuzzi, A., and

Rutherford, A. W. (2018). Photochemistry beyond the red limit in chlorophyll f-containing photosystems. *Science,* 360(6394):1210–1213.

Nurse, P. (2008). Life, logic and information. *Nature,* 454(7203):424–426.

Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J., and Chivas, A. R. (2016). Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature,* 537(7621):535–538.

Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J., Rothacker, L., and Chivas, A. R. (2019). Cross-examining Earth's oldest stromatolites: Seeing through the effects of heterogeneous deformation, metamorphism and metasomatism affecting Isua (Greenland) ~3700 Ma sedimentary rocks. *Precambrian Res.,* 331:105347.

Oakley, K. P. (1968). *Man the Tool-Maker*. The University of Chicago Press, Chicago, IL.

Öberg, K. (2016). Photochemistry and astrochemistry: Photochemical pathways to interstellar complex organic molecules. *Chem. Rev.,* 116(17):9631–9663.

O'Brien, D. P., Izidoro, A., Jacobson, S. A., Raymond, S. N., and Rubie, D. C. (2018). The delivery of water during terrestrial planet formation. *Space Sci. Rev.,* 214(1):47.

Odert, P., Leitzinger, M., Hanslmeier, A., and Lammer, H. (2017). Stellar coronal mass ejections - I. Estimating occurrence frequencies and mass-loss rates. *Mon. Not. R. Astron. Soc.,* 472(1):876–890.

Odling-Smee, F. J., Laland, K. N., and Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton University Press, Princeton, NJ.

Ó Fionnagáin, D. and Vidotto, A. A. (2018). The solar wind in time: A change in the behaviour of older winds? *Mon. Not. R. Astron. Soc.,* 476(2):2465–2475.

O'Malley, M. A. and Powell, R. (2016). Major problems in evolutionary transitions: How a metabolic perspective can enrich our understanding of macroevolution. *Biol. Philos.,* 31(2):159–189.

O'Malley-James, J. T. and Kaltenegger, L. (2018). Biofluorescent worlds: Global biological fluorescence as a biosignature. *Mon. Not. R. Astron. Soc.,* 481(2): 2487–2496.

O'Malley-James, J. T. and Kaltenegger, L. (2019a). Biofluorescent worlds—II. Biological fluorescence induced by stellar UV flares, a new temporal biosignature. *Mon. Not. R. Astron. Soc.,* 488(4):4530–4545.

O'Malley-James, J. T. and Kaltenegger, L. (2019b). Lessons from early Earth: UV surface radiation should not limit the habitability of active M star systems. *Mon. Not. R. Astron. Soc.,* 485(4):5598–5603.

O'Neill, G. K. (1977). *The High Frontier: Human Colonies in Space*. William Morrow & Co., New York, NY.

Ohtomo, Y., Kakegawa, T., Ishida, A., Nagase, T., and Rosing, M. T. (2014). Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. *Nat. Geosci.,* 7(1):25–28.

Oliver, B. M. (1979). Rationale for the water hole. *Acta Astronaut.,* 6:71–79.

Olkowicz, S., Kocourek, M., Lučan, R. K., Porteš, M., Fitch, W. T., Herculano-Houzel, S., and Němec, P. (2016). Birds have primate-like numbers of neurons in the forebrain. *Proc. Natl. Acad. Sci. USA,* 113(26):7255–7260.

Olson, J. M. (2006). Photosynthesis in the Archean era. *Photosynth. Res.,* 88(2): 109–117.

Olson, S. L., Schwieterman, E. W., Reinhard, C. T., Ridgwell, A., Kane, S. R., Meadows, V. S., and Lyons, T. W. (2018). Atmospheric seasonality as an exoplanet biosignature. *Astrophys. J. Lett.,* 858(2):L14.

Olson, S. L., Jansen, M., and Abbot, D. S. (2020). Oceanographic considerations for exoplanet life detection. *Astrophys. J.,* 895(1):19.

Onofri, S., de la Torre, R., de Vera, J.-P., Ott, S., Zucconi, L., Selbmann, L., Scalzi, G., Venkateswaran, K. J., Rabbow, E., Sánchez Iñigo, F. J., and Horneck, G. (2012). Survival of rock-colonizing organisms after 1.5 years in outer space. *Astrobiology,* 12(5):508–516.

Onstott, T. C., Ehlmann, B. L., Sapers, H., Coleman, M., Ivarsson, M., Marlow, J. J., Neubeck, A., and Niles, P. (2019). Paleo-rock-hosted life on Earth and the search on Mars: A review and strategy for exploration. *Astrobiology,* 19(10):1230–1262.

Ooka, H., McGlynn, S. E., and Nakamura, R. (2019). Electrochemistry at deep-sea hydrothermal vents: Utilization of the thermodynamic driving force towards the autotrophic origin of life. *ChemElectroChem,* 6(5):1316–1323.

Oparin, A. I. (1938). *The Origin of Life*. Macmillan, New York, NY.

Orban, G. A. and Caruana, F. (2014). The neural basis of human tool use. *Front. Psychol.,* 5:310.

Orf, G. S., Gisriel, C., and Redding, K. E. (2018). Evolution of photosynthetic reaction centers: Insights from the structure of the heliobacterial reaction center. *Photosyn. Res.,* 138(1):11–37.

Orgel, L. E. (2004). Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.,* 39(2):99–123.

Orgel, L. E. (2008). The implausibility of metabolic cycles on the prebiotic Earth. *PLOS Biol.,* 6(1):e18.

Oró, J. (1960). Synthesis of adenine from ammonium cyanide. *Biochem. Biophys. Res. Commun.,* 2(6):407–412.

Oró, J. (1961). Comets and the formation of biochemical compounds on the primitive earth. *Nature,* 190(4774):389–390.

Orzechowska, G. E., Goguen, J. D., Johnson, P. V., Tsapin, A., and Kanik, I. (2007). Ultraviolet photolysis of amino acids in a 100 K water ice matrix: Application to the outer Solar system bodies. *Icarus,* 187(2):584–591.

Osinski, G. R., Cockell, C. S., Pontefract, A., and Sapers, H. M. (2020). The role of meteorite impacts in the origin of life. *Astrobiology,* 20(9):1121–1149.

Osmanov, Z. (2016). On the search for artificial Dyson-like structures around pulsars. *Int. J. Astrobiol.,* 15(2):127–132.

Osmanov, Z. and Berezhiani, V. I. (2018). On the possibility of the Dyson spheres observable beyond the infrared spectrum. *Int. J. Astrobiol.,* 17(4): 356–360.

Ostriker, J. P. and Turner, E. L. (1986). The inclusion of interaction terms into population dynamics equations of interstellar colonization. *J. Br. Interplanet. Soc.,* 39(3):141.

Osvath, M. (2009). Spontaneous planning for future stone throwing by a male chimpanzee. *Curr. Biol.,* 19(5):R190–R191.

Otto, S. P. (2009). The evolutionary enigma of sex. *Am. Nat.,* 174(S1):S1–S14.

Ovaskainen, O. (2002). Long-term persistence of species and the SLOSS problem. *J. Theor. Biol.,* 218(4):419–433.

Owen, J. E. and Alvarez, M. A. (2016). UV driven evaporation of close-in planets: Energy-limited, recombination-limited, and photon-limited flows. *Astrophys. J.,* 816(1):34.

Owen, J. E. (2019). Atmospheric escape and the evolution of close-in exoplanets. *Annu. Rev. Earth Planet. Sci.,* 47:67–90.

Oza, A. V., Leblanc, F., Johnson, R. E., Schmidt, C., Leclercq, L., Cassidy, T. A., and Chaufray, J.-Y. (2019). Dusk over dawn O_2 asymmetry in Europa's near-surface atmosphere. *Planet. Space Sci.,* 167:23–32.

Ozaki, K., Thompson, K. J., Simister, R. L., Crowe, S. A., and Reinhard, C. T. (2020). Anoxygenic photosynthesis and the delayed oxygenation of Earth's atmosphere. *Nat. Commun.,* 10:3026.

Pace, G. W. and Walker, J. C. G. (1975). Time markers in interstellar communication. *Nature,* 254(5499):400–401.

Pace, N. R. (2001). The universal nature of biochemistry. *Proc. Natl. Acad. Sci. USA,* 98(3):805–808.

Pályi, G. (2020). *Biological Chirality*. Academic Press, London, UK.

Papagiannis, M. D. (1978). Are we all alone, or could they be in the asteroid belt? *Q. Jl. R. Astr. Soc.,* 19:277–281.

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., and Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. USA,* 108(33):13624–13629.

Parkin, K. L. G. (2018). The Breakthrough Starshot system model. *Acta Astronaut.,* 152:370–384.

Parkos, D., Pikus, A., Alexeenko, A., and Melosh, H. J. (2018). HCN production via impact ejecta reentry during the Late Heavy Bombardment. *J. Geophys. Res. Planets,* 123(4):892–909.

Parnell, J. (2004). Plate tectonics, surface mineralogy, and the early evolution of life. *Int. J. Astrobiol.,* 3(2):131–137.

Parrish, J. T. (1993). Climate of the supercontinent Pangea. *J. Geol.,* 101(2): 215–233.

Pascal, R., Pross, A., and Sutherland, J. D. (2013). Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol.,* 3:130156.

Pasek, M. A. and Greenberg, R. (2012). Acidification of Europa's subsurface ocean as a consequence of oxidant delivery. *Astrobiology,* 12(2):151–159.

Pasek, M. A., Gull, M., and Herschy, B. (2017). Phosphorylation on the early Earth. *Chem. Geol.,* 475:149–170.

Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D., and Sutherland, J. D. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.,* 7(4):301–307.

Patty, C. H. L., ten Kate, I. L., Sparks, W. B., and Snik, F. (2018). Remote sensing of homochirality: A proxy for the detection of extraterrestrial life. In Polavarapu, P. L., editor, *Chiral Analysis: Advances in Spectroscopy, Chromatography and Emerging Methods,* pages 29–69. Elsevier, Amsterdam, Netherlands (2nd edition).

Pawar, S., Dell, A. I., Savage, V. M., and Knies, J. L. (2016). Real versus artificial variation in the thermal sensitivity of biological traits. *Am. Nat.,* 187(2): E41–E52.

Payne, J. L., McClain, C. R., Boyer, A. G., Brown, J. H., Finnegan, S., Kowalewski, M., Krause, R. A., Lyons, S. K., McShea, D. W., Novack-Gottshall, P. M., Smith, F. A., Spaeth, P., Stempien, J. A., and Wang, S. C. (2011). The evolutionary consequences of oxygenic photosynthesis: A body size perspective. *Photosynth. Res.,* 107(1):37–57.

Paytan, A. and McLaughlin, K. (2007). The oceanic phosphorus cycle. *Chem. Rev.,* 107(2):563–576.

Pearce, B. K. D., Pudritz, R. E., Semenov, D. A., and Henning, T. K. (2017). Origin of the RNA world: The fate of nucleobases in warm little ponds. *Proc. Natl. Acad. Sci. USA,* 114(43):11327–11332.

Pearce, B. K. D., Tupper, A. S., Pudritz, R.-E., and Higgs, P. G. (2018). Constraining the time interval for the origin of life on Earth. *Astrobiology,* 18(3): 343–364.

Pedreira-Segade, U., Hao, J., Montagnac, G., Cardon, H., and Daniel, I. (2019). Spontaneous polymerization of glycine under hydrothermal conditions. *ACS Earth Space Chem.,* 3(8):1669–1677.

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.,* 31(2):109–130.

Penn, J. L., Deutsch, C., Payne, J. L., and Sperling, E. A. (2018). Temperature-dependent hypoxia explains biogeography and severity of end-Permian marine mass extinction. *Science,* 362(6419):eaat1327.

Perakis, N. and Hein, A. M. (2016). Combining magnetic and electric sails for interstellar deceleration. *Acta Astronaut.,* 128:13–20.

Pérez-Villa, A., Pietrucci, F., and Saitta, A. M. (2020). Prebiotic chemistry and origins of life research with atomistic computer simulations. *Phys. Life Rev.,* 34:105–135.

Perry, C. J., Barron, A. B., and Chittka, L. (2017). The frontiers of insect cognition. *Curr. Opin. Behav. Sci.,* 16:111–118.

Perryman, M., Hartman, J., Bakos, G. Á., and Lindegren, L. (2014). Astrometric exoplanet detection with *GAIA*. *Astrophys. J.,* 797(1):14.

Perryman, M. (2018). *The Exoplanet Handbook*. Cambridge University Press, Cambridge, UK (2nd edition).

Persson, T., Sauciuc, G.-A., and Madsen, E. A. (2018). Spontaneous cross-species imitation in interactions between chimpanzees and zoo visitors. *Primates,* 59(1):19–29.

Peslier, A. H., Schönbächler, M., Busemann, H., and Karato, S.-I. (2017). Water in the Earth's interior: Distribution and origin. *Space Sci. Rev.,* 212:743–810.

Petkowski, J. J., Bains, W., and Seager, S. (2020). On the potential of silicon as a building block for life. *Life,* 10(6):84.

Petrenko, V. F. and Whitworth, R. W. (1999). *Physics of Ice*. Oxford University Press, Oxford, UK.

Petroff, P., Hessels, J. W. T., and Lorimer, D. R. (2019). Fast Radio Bursts. *Astron. Astrophys. Rev.,* 27:4.

Philip, G. K. and Freeland, S. J. (2011). Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology,* 11(3):235–240.

Phipps, C., Birkan, M., Bohn, W., Eckel, H.-A., Horisawa, H., Lippert, T., Michaelis, M., Rezunkov, Y., Sasoh, A., Schall, W., Scharring, S., and Sinko, J. (2010). Review: Laser-ablation propulsion. *J. Propuls. Power,* 26(4): 609–637.

Pierrehumbert, R. T. (2010). *Principles of Planetary Climate*. Cambridge University Press, Cambridge, UK.

Pierrehumbert, R. T. and Hammond, M. (2019). Atmospheric circulation of tide-locked exoplanets. *Annu. Rev. Fluid Mech.,* 51:275–303.

Pinchuk, P., Margot, J.-L., Greenberg, A. H., Ayalde, T., Bloxham, C., Boddu, A., Gerardo Chinchilla-Garcia, L., Cliffe, M., Gallagher, S., Hart, K., Hesford, B., Mizrahi, I., Pike, R., Rodger, D., Sayki, B., Schneck, U., Tan, A., Xiao, Y., and Lynch, R. S. (2019). A search for technosignatures from TRAPPIST-1, LHS 1140, and 10 planetary systems in the Kepler field with the Green Bank Telescope at 1.15–1.73 GHz. *Astron. J.,* 157(3):122.

Pinker, S. (2007). *The Language Instinct: How the Mind Creates Language*. Harper Perennial Modern Classics, New York, NY.

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proc. Natl. Acad. Sci. USA,* 107(Suppl. 2):8993–8999.

Piran, T. and Jimenez, R. (2014). Possible role of Gamma Ray Bursts on life extinction in the Universe. *Phys. Rev. Lett.,* 113(23):231102.

Pittis, A. A. and Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature,* 531(7592):101–104.

Pizzarello, S. and Shock, E. (2017). Carbonaceous chondrite meteorites: The chronicle of a potential evolutionary path between stars and life. *Orig. Life Evol. Biosph.,* 47(3):249–260.

Planavsky, N. J., Asael, D., Hofmann, A., Reinhard, C. T., Lalonde, S. V., Knudsen, A., Wang, X., Ossa Ossa, F., Pecoits, E., Smith, A. J. B., Beukes, N. J., Bekker, A., Johnson, T. M., Konhauser, K. O., Lyons, T. W., and Rouxel, O. J. (2014). Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat. Geosci.,* 7(4):283–286.

Planavsky, N. J., Cole, D. B., Isson, T. T., Reinhard, C. T., Crockford, P. W., Sheldon, N. D., and Lyons, T. W. (2018). A case for low atmospheric oxygen levels during Earth's middle history. *Emerging Topics in Life Sciences,* 2(2):149–159.

Plasson, R., Kondepudi, D. K., Bersini, H., Commeyras, A., and Asakura, K. (2007). Emergence of homochirality in far-from-equilibrium systems: Mechanisms and role in prebiotic chemistry. *Chirality,* 19(8):589–600.

Plotnik, J. M., de Waal, F. B. M., and Reiss, D. (2006). Self-recognition in an Asian elephant. *Proc. Natl. Acad. Sci. USA,* 103(45):17053–17057.

Pöhlker, C., Wiedemann, K. T., Sinha, B., Shiraiwa, M., Gunthe, S. S., Smith, M., Su, H., Artaxo, P., Chen, Q., Cheng, Y., Elbert, W., Gilles, M. K., Kilcoyne, A. L. D., Moffet, R. C., Weigand, M., Martin, S. T., Pöschl, U., and Andreae, M. O. (2012). Biogenic potassium salt particles as seeds for secondary organic aerosol in the Amazon. *Science,* 337(6098):1075–1078.

Pohorille, A. and Pratt, L. R. (2012). Is water the universal solvent for life? *Orig. Life Evol. Biosph.,* 42(5):405–409.

Pohorille, A. and Sokolowska, J. (2020). Evaluating biosignatures for life detection. *Astrobiology,* 20(10):1236–1250.

Poole, A. M. and Gribaldo, S. (2014). Eukaryotic origins: How and when was the mitochondrion acquired? *Cold Spring Harb. Perspect. Biol.,* 6(12):a015990.

Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge, New York, NY.

Por, E. H. and Haffert, S. Y. (2020). The Single-mode Complex Amplitude Refinement (SCAR) coronagraph. I. Concept, theory, and design. *Astron. Astrophys.,* 635:A55.

Porter, S. M. (2020). Insights into eukaryogenesis from the fossil record. *Interface Focus,* 10(4):20190105.

Postberg, F., Khawaja, N., Abel, B., Choblet, G., Glein, C. R., Gudipati, M. S., Henderson, B. L., Hsu, H.-W., Kempf, S., Klenner, F., Moragas-Klostermeyer, G., Magee, B., Nölle, L., Perry, M., Reviol, R., Schmidt, J., Srama, R., Stolz, F., Tobie, G., . . . Waite, J. H. (2018). Macromolecular organic compounds from the depths of Enceladus. *Nature,* 558(7711): 564–568.

Poudyal, R. R., Guth-Metzler, R. M., Veenis, A. J., Frankel, E. A., Keating, C. D., and Bevilacqua, P. C. (2019). Template-directed RNA polymerization and enhanced ribozyme catalysis inside membraneless compartments formed by coacervates. *Nat. Commun.,* 10:490.

Powner, M. W., Gerland, B., and Sutherland, J. D. (2009). Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature,* 459(7244):239–242.

Prantzos, N. (2008). On the "Galactic Habitable Zone." *Space Sci. Rev.,* 135:313–322.

Preiner, M., Asche, S., Becker, S., Betts, H. C., Boniface, A., Camprubi, E., Chandru, K., Erastova, V., Garg, S. G., Khawaja, N., Kostyrka, G., Machné, R., Moggioli, G., Muchowska, K. B., Neukirchen, S., Peter, B., Pichlhöfer, E., Radványi, A., Rossetto, D., . . . Xavier, J. C. (2020a). The future of origin of life research: Bridging decades-old divisions. *Life,* 10(3):20.

Preiner, M., Igarashi, K., Muchowska, K. B., Yu, M., Varma, S. J., Kleinermanns, K., Nobu, M. K., Kamagata, Y., Tüysüz, H., Moran, J., and Martin, W. F. (2020b). A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism. *Nat. Ecol. Evol.,* 4:534–542.

Pressman, A., Blanco, C., and Chen, I. A. (2015). The RNA world as a model system to study the origin of life. *Curr. Biol.,* 25(19):R953–R963.

Price, C. A., Weitz, J. S., Savage, V. M., Stegen, J., Clarke, A., Coomes, D. A., Dodds, P. S., Etienne, R. S., Kerkhoff, A. J., McCulloh, K., Niklas, K. J., Olff, H., and Swenson, N. G. (2012). Testing the metabolic theory of ecology. *Ecol. Lett.,* 15(12):1465–1474.

Price, D. C., Enriquez, J. E., Brzycki, B., Croft, S., Czech, D., DeBoer, D., DeMarines, J., Foster, G., Gajjar, V., Gizani, N., Hellbourg, G., Isaacson, H., Lacki, B., Lebofsky, M., MacMahon, D. H. E., de Pater, I., Siemion, A. P. V., Werthimer, D., Green, J. A., . . . Worden, S. P. (2020). The Breakthrough Listen search for intelligent life: Observations of 1327 nearby stars over 1.10–3.45 GHz. *Astron. J.,* 159(3):86.

Price, M. C., Solscheid, C., Burchell, M. J., Josse, L., Adamek, N., and Cole, M. J. (2013). Survival of yeast spores in hypervelocity impact events up to velocities of 7.4 km s^{-1}. *Icarus,* 222(1):263–272.

Priest, E. (2014). *Magnetohydrodynamics of the Sun*. Cambridge University Press, Cambridge, UK.

Prior, H., Schwarz, A., and Güntürkün, O. (2008). Mirror-induced behavior in the magpie (*Pica pica*): Evidence of self-recognition. *PLoS Biol.,* 6(8):e202.

Pross, A. (2016). *What Is Life?: How Chemistry Becomes Biology*. Oxford University Press, Oxford, UK.

Prothero, D. R. (2007). *Evolution: What the Fossils Say and Why It Matters*. Columbia University Press, New York, NY.

Purvis, A. and Hector, A. (2000). Getting the measure of biodiversity. *Nature,* 405(6783):212–219.

Race, M., Denning, K., Bertka, C. M., Dick, S. J., Harrison, A. A., Impey, C., and Mancinelli, R. (2012). Astrobiology and society: Building an interdisciplinary research community. *Astrobiology,* 12(10):958–965.

Ragsdale, S. W. and Pierce, E. (2008). Acetogenesis and the Wood-Ljungdahl pathway of CO_2 fixation. *Biochim. Biophys. Acta—Proteins Proteomics,* 1784(12): 1873–1898.

Rahvar, S. (2016). Gravitational microlensing events as a target for the SETI project. *Astrophys. J.,* 828(1):19.

Raines, C. A. (2003). The Calvin cycle revisited. *Photosynth. Res.,* 75(1):1–10.

Rainey, P. B. and De Monte, S. (2014). Resolving conflicts during the evolutionary transition to multicellular life. *Annu. Rev. Ecol. Evol. Syst.,* 45:599–620.

Rajamani, S., Vlassov, A., Benner, S., Coombs, A., Olasagasti, F., and Deamer, D. (2008). Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Orig. Life Evol. Biosph.,* 38(1):57–74.

Ramirez, R. M. and Kaltenegger, L. (2014). The habitable zones of pre-main-sequence stars. *Astrophys. J. Lett.,* 797(2):L25.

Ramirez, R. M. (2018). A more comprehensive habitable zone for finding life on other planets. *Geosciences,* 8:280.

Ramirez, R. M. and Levi, A. (2018). The ice cap zone: A unique habitable zone for ocean worlds. *Mon. Not. R. Astron. Soc.,* 477(4):4627–4640.

Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Phys. Life Rev.,* 24:1–16.

Ranjan, S., Wordsworth, R., and Sasselov, D. D. (2017). The surface UV environment on planets orbiting M dwarfs: Implications for prebiotic chemistry and the need for experimental follow-up. *Astrophys. J.,* 843(2):110.

Ranjan, S., Todd, Z. R., Rimmer, P. B., Sasselov, D. D., and Babbin, A. R. (2019). Nitrogen oxide concentrations in natural waters on early Earth. *Geochem. Geophys. Geosyst.,* 20(4):2021–2039.

Rapf, R. J. and Vaida, V. (2016). Sunlight as an energetic driver in the synthesis of molecules necessary for life. *Phys. Chem. Chem. Phys.,* 18(30):20067–20084.

Ratcliff, W. C., Denison, R. F., Borrello, M., and Travisano, M. (2012). Experimental evolution of multicellularity. *Proc. Natl. Acad. Sci. USA,* 109(5):1595–1600.

Raup, D. M. (1992). Nonconscious intelligence in the Universe. *Acta Astronaut.,* 26(3):257–261.

Ravi, V. (2019). The prevalence of repeating Fast Radio Bursts. *Nat. Astron.,* 3: 928–931.

Raymond, J. and Segrè, D. (2006). The effect of oxygen on biochemical networks and the evolution of complex life. *Science,* 311(5768):1764–1767.

Reames, D. V. (2013). The two sources of solar energetic particles. *Space Sci. Rev.,* 175:53–92.

Reaves, M. L., Sinha, S., Rabinowitz, J. D., Kruglyak, L., and Redfield, R. J. (2012). Absence of detectable arsenate in DNA from arsenate-grown GFAJ-1 cells. *Science,* 337(6093):470–473.

Reed, C. M. and Durlach, N. I. (1998). Note on information transfer rates in human communication. *Presence,* 7(5):509–518.

Rees, M. (2018). *On the Future: Prospects for Humanity*. Princeton University Press, Princeton, NJ.

Reich, D. (2018). *Who We Are and How We Got Here*. Oxford University Press, Oxford, UK.

Reiners, A. and Christensen, U. R. (2010). A magnetic field evolution scenario for brown dwarfs and giant planets. *Astron. Astrophys.,* 522:A13.

Reines, A. E. and Marcy, G. W. (2002). Optical search for extraterrestrial intelligence: A spectroscopic search for laser emission from nearby stars. *Publ. Astron. Soc. Pac.,* 114(794):416–426.

Reinhard, C. T., Olson, S. L., Schwieterman, E. W., and Lyons, T. W. (2017). False negatives for remote life detection on ocean-bearing planets: Lessons from the early Earth. *Astrobiology,* 17(4):287–297.

Reinhard, C. T., Planavsky, N. J., Gill, B. C., Ozaki, K., Robbins, L. J., Lyons, T. W., Fischer, W. W., Wang, C., Cole, D. B., and Konhauser, K. O. (2017). Evolution of the global phosphorus cycle. *Nature,* 541(7637):386–389.

Reinhard, C. T., Planavsky, N. J., Ward, B. A., Love, G. D., Le Hir, G., and Ridgwell, A. (2020). The impact of marine nutrient abundance on early eukaryotic ecosystems. *Geobiology,* 18(2):139–151.

Reiss, D. and Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proc. Natl. Acad. Sci. USA,* 98(10):5937–5942.

Reynaud, E., Lesourd, M., Navarro, J., and Osiurak, F. (2016). On the neurocognitive origins of human tool use: A critical review of neuroimaging data. *Neurosci. Biobehav. Rev.,* 64:421–437.

Reynolds, R. T., Squyres, S. W., Colburn, D. S., and McKay, C. P. (1983). On the habitability of Europa. *Icarus,* 56(2):246–254.

Ribas, I., Bolmont, E., Selsis, F., Reiners, A., Leconte, J., Raymond, S. N., Engle, S. G., Guinan, E. F., Morin, J., Turbet, M., Forget, F., and Anglada-Escudé, G. (2016). The habitability of Proxima Centauri b. I. Irradiation, rotation

and volatile inventory from formation to the present. *Astron. Astrophys.,* 596:A111.

Rice, M. and Laughlin, G. (2019). Hidden planets: Implications from 'Oumuamua and DSHARP. *Astrophys. J. Lett.,* 884(1):L22.

Richerson, P. J. and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*. The University of Chicago Press, Chicago, IL.

Richter, D. J. and King, N. (2013). The genomic and cellular foundations of animal origins. *Annu. Rev. Genet.,* 47:509–537.

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., Fernandes, P., Raynal, J.-P., Geraads, D., Ben-Ncer, A., Hublin, J.-J., and McPherron, S. P. (2017). The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature,* 546(7657): 293–296.

Rimmer, P. B., Xu, J., Thompson, S. J., Gillen, E., Sutherland, J. D., and Queloz, D. (2018). The origin of RNA precursors on exoplanets. *Sci. Adv.,* 4(8):eaar3302.

Rimmer, P. B. and Shorttle, O. (2019). Origin of life's building blocks in carbon- and nitrogen-rich surface hydrothermal vents. *Life,* 9(1):12.

Rios, A. C. and Tor, Y. (2013). On the origin of the canonical nucleobases: An assessment of selection pressures across chemical and early biological evolution. *Isr. J. Chem.,* 53(6):469–483.

Ritson, D. J., Mojzsis, S. J., and Sutherland, J. D. (2020). Supply of phosphate to early Earth by photogeochemistry after meteoritic weathering. *Nat. Geosci.,* 13(5):344–348.

Robertson, M. P. and Joyce, G. F. (2012). The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.,* 4(5):a003608.

Robinson, T. D., Meadows, V. S., and Crisp, D. (2010). Detecting oceans on extrasolar planets using the glint effect. *Astrophys. J. Lett.,* 721(1):L67–L71.

Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J., and Gardner, E. (2008). A comprehensive comparison of the Sun to other stars: Searching for self-selection effects. *Astrophys. J.,* 684(1): 691–706.

Rodler, F. and López-Morales, M. (2014). Feasibility studies for the detection of O_2 in an Earth-like exoplanet. *Astrophys. J.,* 781(1):54.

Roger, A. J., Muñoz-Gómez, S., and Kamikawa, R. (2017). The origin and diversification of mitochondria. *Curr. Biol.,* 27(21):R1177–R1192.

Rogers, L. A. (2015). Most 1.6 Earth-radius planets are not rocky. *Astrophys. J.,* 801(1):41.

Rokas, A. (2008). The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu. Rev. Genet.,* 42:235–251.

Rood, R. T. and Trefil, J. S. (1981). *Are We Alone? The Possibility of Extraterrestrial Civilizations*. Charles Scribner's Sons, New York, NY.

Rosati, A. G. (2017). Foraging cognition: Reviving the ecological intelligence hypothesis. *Trends Cogn. Sci.,* 21(9):691–702.

Rose, C. and Wright, G. (2004). Inscribed matter as an energy-efficient means of communication with an extraterrestrial civilization. *Nature,* 431(7004):47–49.

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Rospars, J.-P. (2013). Trends in the evolution of life, brains and intelligence. *Int. J. Astrobiol.,* 12(3):186–207.

Ross, D. S. and Deamer, D. (2016). Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life,* 6(3):28.

Ross, D. S. and Deamer, D. (2019). Prebiotic oligomer assembly: What was the energy source? *Astrobiology,* 19(4):517–521.

Rossetti, C. (1865). Goblin market. In *Goblin Market and Other Poems*. Macmillan and Co., London, UK (2nd edition).

Roth, G. and Dicke, U. (2019). Origin and evolution of human cognition. *Prog. Brain Res.,* 250:285–316.

Rothschild, L. J. (1999). The influence of UV radiation on protistan evolution. *J. Eukaryot. Microbiol.,* 46(5):548–555.

Rothschild, L. J. and Mancinelli, R. L. (2001). Life in extreme environments. *Nature,* 409(6823):1092–1101.

Roy, K. I., Kennedy, R. G., and Fields, D. E. (2013). Shell worlds. *Acta Astronaut.,* 82(2):238–245.

Ruane, G., Wang, J., Mawet, D., Jovanovic, N., Delorme, J.-R., Mennesson, B., and Wallace, J. K. (2018). Efficient spectroscopy of exoplanets at small angular separations with vortex fiber nulling. *Astrophys. J.,* 867(2):143.

Ruff, S. W. and Farmer, J. D. (2016). Silica deposits on Mars with features resembling hot spring biosignatures at El Tatio in Chile. *Nat. Commun.,* 7:13554.

Ruff, S. W., Campbell, K. A., Van Kranendonk, M. J., Rice, M. S., and Farmer, J. D. (2020). The case for ancient hot springs in Gusev crater, Mars. *Astrobiology,* 20(4):475–499.

Rugheimer, S., Segura, A., Kaltenegger, L., and Sasselov, D. (2015). UV surface environment of Earth-like planets orbiting FGKM stars through geological evolution. *Astrophys. J.,* 806(1):137.

Ruiz-Mirazo, K., Peretó, J., and Moreno, A. (2010). Defining life or bringing biology to life. *Orig. Life Evol. Biosph.,* 40(2):203–213.

Ruiz-Mirazo, K., Briones, C., and de la Escosura, A. (2014). Prebiotic systems chemistry: New perspectives for the origins of life. *Chem. Rev.,* 114(1): 285–366.

Rūmī, J. A.-D. M. (2004). The fragile vial. In *The Essential Rumi,* pp. 14–15. Translated by C. Barks, R. Nicholson, A. J. Arberry, and J. Moyne. HarperOne, New York, NY.

Rushby, A. J., Claire, M. W., Osborn, H., and Watson, A. J. (2013). Habitable zone lifetimes of exoplanets around main sequence stars. *Astrobiology,* 13(9): 833–849.

Russell, D. A. (1983). Exponential evolution: Implications for intelligent extraterrestrial life. *Adv. Space Res.,* 3(9):95–103.

Russell, M. J. and Hall, A. J. (1997). The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc.,* 154(3):377–402.

Russell, M. J. and Martin, W. (2004). The rocky roots of the acetyl-CoA pathway. *Trends Biochem. Sci.,* 29(7):358–363.

Russell, M. J., Nitschke, W., and Branscomb, E. (2013). The inevitable journey to being. *Phil. Trans. R. Soc. B,* 368(1622):20120254.

Russell, M. J., Barge, L. M., Bhartia, R., Bocanegra, D., Bracher, P. J., Branscomb, E., Kidd, R., McGlynn, S., Meier, D. H., Nitschke, W., Shibuya, T., Vance, S., White, L., and Kanik, I. (2014). The drive to life on wet and icy worlds. *Astrobiology,* 14(4):308–343.

Russell, M. J. (2018). Green rust: The simple organizing "Seed" of all life? *Life,* 8(3):35.

Safina, C. (2015). *Beyond Words: What Animals Think and Feel*. Henry Holt & Co., New York, NY.

Sagan, C. (1963). Direct contact among Galactic civilizations by relativistic interstellar spaceflight. *Planet. Space Sci.,* 11(5):485–498.

Sagan, C. and Walker, R. G. (1966). The infrared detectability of Dyson civilizations. *Astrophys. J.,* 144(3):1216–1218.

Sagan, C. and Khare, B. N. (1971). Long-wavelength ultraviolet photoproduction of amino acids on the primitive Earth. *Science,* 173(3995):417–420.

Sagan, C. and Mullen, G. (1972). Earth and Mars: Evolution of atmospheres and surface temperatures. *Science,* 177(4043):52–56.

Sagan, C. (1973). *The Cosmic Connection: An Extraterrestrial Perspective*. Doubleday, New York, NY.

Sagan, C. and Salpeter, E. E. (1976). Particles, environments, and possible ecologies in the Jovian atmosphere. *Astrophys. J., Suppl. Ser.,* 32:737–755.

Sagan, C. (1980). *Broca's Brain: Reflections on the Romance of Science*. Ballantine Books, New York, NY.

Sagan, C., Druyan, A., and Soter, S. (Writers) & Malone, A. (Director). (1980, 28 September). The shores of the cosmic ocean (Season 1, Episode 1) [TV series episode]. In Andorfer, G. and McCain, R. (Producers), *Cosmos: A Personal Voyage*. PBS.

Sagan, C., Thompson, W. R., Carlson, R., Gurnett, D. and Hord, C. (1993). A search for life on Earth from the galileo spacecraft. *Nature,* 365(6448):715–721.

Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.,* 14(3):225–274.

Saito, Y. and Hyuga, H. (2013). Homochirality: Symmetry breaking in systems driven far from equilibrium. *Rev. Mod. Phys.,* 85(2):603–621.

Saladino, R., Crestini, C., Pino, S., Costanzo, G., and Di Mauro, E. (2012). Formamide and the origin of life. *Phys. Life Rev.,* 9(1):84–104.

Saladino, R., Di Mauro, E., and García-Ruiz, J. M. (2019). A universal geochemical scenario for formamide condensation and prebiotic chemistry. *Chem. Eur. J.,* 25(13):3181–3189.

Salazar, A. M., Olson, S. L., Komacek, T. D., Stephens, H., and Abbot, D. S. (2020). The effect of substellar continent size on ocean dynamics of Proxima Centauri b. *Astrophys. J. Lett.,* 896(1):L16.

Sallmen, S., Korpela, E. J., and Crawford-Taylor, K. (2019). Improved analysis of Clarke exobelt detectability. *Astron. J.,* 158(6):258.

Sánchez-Baracaldo, P., Raven, J. A., Pisani, D., and Knoll, A. H. (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl. Acad. Sci. USA,* 114(37):E7737–E7745.

Sánchez-Baracaldo, P. and Cardona, T. (2020). On the origin of oxygenic photosynthesis and cyanobacteria. *New Phytol.,* 225(4):1440–1446.

Sandberg, A., Armstrong, S. and Ćirković, M. M. (2016). That is not dead which can eternal lie: The aestivation hypothesis for resolving Fermi's paradox. *J. Br. Interplanet. Soc.,* 69(11):406–415.

Sandford, S. A., Nuevo, M., Bera, P. P., and Lee, T. J. (2020). Prebiotic astrochemistry and the formation of molecules of astrobiological interest in interstellar clouds and protostellar disks. *Chem. Rev.,* 120(11):4616–4659.

Sänger, E. (1953). Zur theorie der photonenraketen. *Ingenieur-Archiv,* 21(3): 213–226.

Sarkar, S., Das, S., Dagar, S., Joshi, M. P., Mungi, C. V., Sawant, A. A., Patki, G. M., and Rajamani, S. (2020). Prebiological membranes and their role in the emergence of early cellular life. *J. Membr. Biol.,* 253:589–608.

Sarmiento, J. L. and Gruber, N. (2006) *Ocean Biogeochemical Dynamics*. Princeton University Press, Princeton, NJ.

Sasselov, D. D., Grotzinger, J. P., and Sutherland, J. D. (2020). The origin of life as a planetary phenomenon. *Sci. Adv.,* 6(6):eaax3419.

Sato, N. (2019). *Endosymbiotic Theories of Organelles Revisited: Retrospects and Prospects*. Springer, Singapore.

Saur, J., Duling, S., Roth, L., Jia, X., Strobel, D. F., Feldman, P. D., Christensen, U. R., Retherford, K. D., McGrath, M. A., Musacchio, F., Wennmacher, A., Neubauer, F. M., Simon, S., and Hartkorn, O. (2015). The search for a subsurface ocean in Ganymede with Hubble Space Telescope observations of its auroral ovals. *J. Geophys. Res. Space Phys.,* 120(3):1715–1737.

Scalo, J., Kaltenegger, L., Segura, A., Fridlund, M., Ribas, I., Kulikov, Yu. N., Grenfell, J. L., Rauer, H., Odert, P., Leitzinger, M., Selsis, F., Khodachenko, M. L.,

Eiroa, C., Kasting, J., and Lammer, H. (2007). M stars as targets for terrestrial exoplanet searches and biosignature detection. *Astrobiology,* 7(1):85–166.

Scharf, C. and Cronin, L. (2016). Quantifying the origins of life on a planetary scale. *Proc. Natl. Acad. Sci. USA,* 113(29):8127–8132.

Scharf, C. (2019). Exoplanet exergy: Why useful work matters for planetary habitability. *Astrophys. J.,* 876(1):16.

Scheffer, L. K. (1994). Machine intelligence, the cost of interstellar travel and Fermi's paradox. *Q. Jl. R. Astr. Soc.,* 35(2):157–175.

Schirrmeister, B. E., Sanchez-Baracaldo, P., and Wacey, D. (2016). Cyanobacterial evolution during the Precambrian. *Int. J. Astrobiol.,* 15(3):187–204.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., and Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science,* 358(6363): 652–655.

Schlesinger, W. H. and Bernhardt, E. S. (2013). *Biogeochemistry: An Analysis of Global Change*. Academic Press, Waltham, MA (3rd edition).

Schmelz, M., Grueneisen, S., Kabalak, A., Jost, J., and Tomasello, M. (2017). Chimpanzees return favors at a personal cost. *Proc. Natl. Acad. Sci. USA,* 114(28):7462–7467.

Schmidt, G. A. and Frank, A. (2019). The Silurian hypothesis: Would it be possible to detect an industrial civilization in the geological record? *Int. J. Astrobiol.,* 18(2):142–150.

Schneider, J., Léger, A., Fridlund, M., White, G. J., Eiroa, C., Henning, T., Herbst, T., Lammer, H., Liseau, R., Paresce, F., Penny, A., Quirrenbach, A., Röttgering, H., Selsis, F., Beichman, C., Danchi, W., Kaltenegger, L., Lunine, J., Stam, D., and Tinetti, G. (2010). The far future of exoplanet direct characterization. *Astrobiology,* 10(1):121–126.

Schopf, J. W., Kitajima, K., Spicuzza, M. J., Kudryavtsev, A. B., and Valley, J. W. (2018). SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions. *Proc. Natl. Acad. Sci. USA,* 115(1):53–58.

Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, UK.

Schroeder, K. (2011). *The deepening paradox*. KarlSchroeder.com. https://www.kschroeder.com/weblog/the-deepening-paradox

Schubert, G., Turcotte, D. L., and Olson, P. (2001). *Mantle Convection in the Earth and Planets*. Cambridge University Press, Cambridge, UK.

Schulte, P. M. (2015). The effects of temperature on aerobic metabolism: Towards a mechanistic understanding of the responses of ectotherms to a changing environment. *J. Exp. Biol.,* 218(12):1856–1866.

Schulze-Makuch, D. and Bains, W. (2017). *The Cosmic Zoo: Complex Life on Many Worlds*. Springer, Cham, Switzerland.

Schulze-Makuch, D. and Bains, W. (2018). Time to consider search strategies for complex life on exoplanets. *Nat. Astron.,* 2:432–433.

Schulze-Makuch, D. and Irwin, L. N. (2018). *Life in the Universe: Expectations and Constraints*. Springer-Verlag, Berlin, Germany (3rd edition).

Schwartz, R. N. and Townes, C. H. (1961). Interstellar and interplanetary communication by optical masers. *Nature,* 190(4772):205–208.

Schwieterman, E. W., Kiang, N. Y., Parenteau, M. N., Harman, C. E., DasSarma, S., Fisher, T. M., Arney, G. N., Hartnett, H. E., Reinhard, C. T., Olson, S. L., Meadows, V. S., Cockell, C. S., Walker, S. I., Grenfell, J. L., Hegde, S., Rugheimer, S., Hu, R., and Lyons, T. W. (2018). Exoplanet biosignatures: A review of remotely detectable signs of life. *Astrobiology,* 18(6): 663–708.

Schwieterman, E. W., Reinhard, C. T., Olson, S. L., Harman, C. E., and Lyons, T. W. (2019). A limited habitable zone for complex life. *Astrophys. J.,* 878(1):19.

Schwieterman, E. W., Reinhard, C. T., Olson, S. L., Ozaki, K., Harman, C. E., Hong, P. K., and Lyons, T. W. (2019). Rethinking CO antibiosignatures in the search for life beyond the Solar system. *Astrophys. J.,* 874(1):9.

Schwille, P. (2017). How simple could life be? *Angew. Chem. Int. Ed.,* 56(37): 10998–11002.

Scorei, R. (2012). Is boron a prebiotic element? A mini-review of the essentiality of boron for the appearance of life on Earth. *Orig. Life Evol. Biosph.,* 42(1): 3–17.

Scott-Phillips, T. (2014). *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make It Special*. Palgrave Macmillan, London, UK.

Seager, S., Turner, E. L., Schafer, J., and Ford, E. B. (2005). Vegetation's red eedge: A possible spectroscopic biosignature of extraterrestrial plants. *Astrobiology,* 5(3):372–390.

Seager, S. (2010). *Exoplanet Atmospheres: Physical Processes*. Princeton University Press, Princeton, NJ.

Seager, S., Schrenk, M., and Bains, W. (2012). An astrophysical view of earth-based metabolic biosignature gases. *Astrobiology,* 12(1):61–82.

Seager, S., Bains, W., and Hu, R. (2013). A biomass-based model to estimate the plausibility of exoplanet biosignature gases. *Astrophys. J.,* 775(2):104.

Seager, S., Bains, W., and Petkowski, J. J. (2016). Toward a list of molecules as potential biosignature gases for the search for life on exoplanets and applications to terrestrial biochemistry. *Astrobiology,* 16(6):465–485.

Seager, S. (2018). The search for habitable planets with biosignature gases framed by a "Biosignature Drake Equation." *Int. J. Astrobiol.,* 17(4):294–302.

Seager, S., Huang, J., Petkowski, J. J., and Pajusalu, M. (2020). Laboratory studies on the viability of life in H_2-dominated exoplanet atmospheres. *Nat. Astron.,* 4: 802–806.

Seager, S., Petkowski, J. J., Gao, P., Bains, W., Bryan, N. C., Ranjan, S., and Greaves, J. (2021). The Venusian lower atmosphere haze as a depot for desiccated microbial life: A proposed life cycle for persistence of the Venusian aerial biosphere. *Astrobiology,* DOI:10.1089/ast.2020.2244

Sebé-Pedrós, A., Degnan, B. M., and Ruiz-Trillo, I. (2017). The origin of Metazoa: A unicellular perspective. *Nat. Rev. Genet.,* 18:498–512.

Segré, D., Ben-Eli, D., Deamer, D. W., and Lancet, D. (2001). The lipid world. *Orig. Life Evol. Biosph.,* 31:119–145.

Segura, A., Walkowicz, L. M., Meadows, V., Kasting, J., and Hawley, S. (2010). The effect of a strong stellar flare on the atmospheric chemistry of an Earth-like planet orbiting an M dwarf. *Astrobiology,* 10(7):751–771.

Seligman, D. and Laughlin, G. (2018). The feasibility and benefits of in situ exploration of 'Oumuamua-like objects. *Astron. J.,* 155(5):217.

Semenov, S. N., Kraft, L. J., Ainla, A., Zhao, M., Baghbanzadeh, M., Campbell, V. E., Kang, K., Fox, J. M., and Whitesides, G. M. (2016). Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature,* 537(7622):656–660.

Semyonov, O. G. (2017). Diamagnetic antimatter storage. *Acta Astronaut.,* 136: 190–203.

Sephton, M. A., Waite, J. H., and Brockwell, T. G. (2018). How to detect life on icy moons. *Astrobiology,* 18(7):843–855.

Servais, T. and Harper, D. A. T. (2018). The Great Ordovician Biodiversification Event (GOBE): Definition, concept and duration. *Lethaia,* 51(2):151–164.

Seto, N. (2019). Possibility of a coordinated signaling scheme in the Galaxy and SETI experiments. *Astrophys. J. Lett.,* 875(2):L10.

Shao, Y., Wyrwoll, K.-H., Chappell, A., Huang, J., Lin, Z., McTainsh, G. H., Mikami, M., Tanaka, T. Y., Wang, X., and Yoon, S. (2011). Dust cycle: An emerging core theme in Earth system science. *Aeolian Res.,* 2(4):181–204.

Shapiro, R. (2006). Small molecule interactions were central to the origin of life. *Q. Rev. Biol.,* 81(2):105–126.

Shapley, H. (1967). *Beyond the Observatory*. Charles Scribner's Sons, New York, NY.

Sharma, C. and Awasthi, S. K. (2017). Versatility of peptide nucleic acids (PNAs): Role in chemical biology, drug discovery, and origins of life. *Chem. Biol. Drug Des.,* 89(1):16–37.

Sharov, A. A. (2006). Genome increase as a clock for the origin and evolution of life. *Biol. Direct,* 1:17.

Sharpe, P. J. H. and DeMichele, D. W. (1977). Reaction kinetics of poikilotherm development. *J. Theor. Biol.,* 64(4):649–670.

Shatilovich, A. V., Tchesunov, A. V., Neretina, T. V., Grabarnik, I. P., Gubin, S. V., Vishnivetskaya, T. A., Onstott, T. C., Rivkina, E. M. (2018). Viable nematodes from late Pleistocene permafrost of the Kolyma River Lowland. *Dokl. Biol. Sci.,* 480(1):100–102.

Sheikh, S. Z., Wright, J. T., Siemion, A., and Enriquez, J. E. (2019). Choosing a maximum drift rate in a SETI search: Astrophysical considerations. *Astron. J.,* 884(1):14.

Sheikh, S. Z. (2020). Nine axes of merit for technosignature searches. *Int. J. Astrobiol.,* 19(3):237–243.

Sheikh, S. Z., Siemion, A., Enriquez, J. E., Price, D. C., Isaacson, H., Lebofsky, M., Gajjar, V., and Kalas, P. (2020). The Breakthrough Listen search for intelligent life: A 3.95–8.00 GHz search for radio technosignatures in the restricted Earth Transit Zone. *Astron. J.,* 160(1):29.

Shepherd, L. (1952). Interstellar flight. *J. Br. Interplanet. Soc.,* 11:149–167.

Shermer, M. (2002). Why ET hasn't called. *Sci. Am.,* 287:33.

Sherwood Lollar, B., Onstott, T. C., Lacrampe-Couloume, G., and Ballentine, C. J. (2014). The contribution of the Precambrian continental lithosphere to global H_2 production. *Nature,* 516(7531):379–382.

Shettleworth, S. J. (2010). *Cognition, Evolution, and Behavior.* Oxford University Press, Oxford, UK (2nd edition).

Shibayama, T., Maehara, H., Notsu, S., Notsu, Y., Nagao, T., Honda, S., Ishii, T. T., Nogami, D., and Shibata, K. (2013). Superflares on Solar-type stars observed with Kepler. I. Statistical properties of superflares. *Astrophys. J. Suppl.,* 209(1):5.

Shields, A. L., Ballard, S., and Johnson, J. A. (2016). The habitability of planets orbiting M-dwarf stars. *Phys. Rep.,* 663:1–38.

Shields, A. L. (2019). The climates of other worlds: A review of the emerging field of exoplanet climatology. *Astrophys. J. Suppl. Ser.,* 243(2):30.

Shiratori, T., Suzuki, S., Kakizawa, Y., and Ishida, K.-i. (2019). Phagocytosis-like cell engulfment by a planctomycete bacterium. *Nat. Commun.,* 10:5529.

Shkadov, L. M. (1987, October 10–17). *Possibility of controlling Solar system motion in the Galaxy* [Paper presentation]. IAF, International Astronautical Congress, 38th, Brighton, England.

Shklovskii, I. S. and Sagan, C. (1966). *Intelligent Life in the Universe*. Holden-Day, Inc., San Francisco, CA.

Shock, E. L., McCollom, T., and Schulte, M. D. (1998). The emergence of metabolism from within hydrothermal systems. In Wiegel, J. and Adams, M. W. W., editors, *Thermophiles: The Keys to the Molecular Evolution and the Origin of Life?* pages 59–76. CRC Press, Boca Raton, FL.

Shumaker, R. W., Walkup, K. R., and Beck, B. B. (2011). *Animal Tool Behavior: The Use and Manufacture of Tools by Animals*. The Johns Hopkins University Press, Baltimore, MD.

Siebrand, W. J. (1982). Radiospectroscopic study of astrophysical photon source flares and the SETI. *J. Br. Interplanet. Soc.,* 35(3):135–141.

Siegel, J., Wang, A. Y., Menabde, S. G., Kats, M. A., Jang, M. S., and Brar, V. W. (2019). Self-stabilizing laser sails based on optical metasurfaces. *ACS Photonics,* 6(8):2032–2040.

Siemion, A. P. V., Demorest, P., Korpela, E., Maddalena, R. J., Werthimer, D., Cobb, J., Howard, A. W., Langston, G., Lebofsky, M., Marcy, G. W., and Tarter, J. (2013). A 1.1–1.9 GHz SETI survey of the Kepler field. I. A search for narrow-band emission from select targets. *Astrophys. J.,* 767(1):94.

Siemion, A., Benford, J., Cheng-Jin, J., Chennamangalam, J., Cordes, J. M., Falcke, H. D. E., Garrington, S. T., Garrett, M. A., Gurvits, L., Hoare, M., Korpela, E., Lazio, J., Messerschmitt, D., Morrison, I., O'Brien, T., Paragi, Z., Penny, A., Spitler, L., Tarter, J. and Werthimer, D. (2014). Searching for extraterrestrial intelligence with the Square Kilometre Array. *PoS,* AASKA14:116.

Sievers, D. and von Kiedrowski, G. (1994). Self-replication of complementary nucleotide-based oligomers. *Nature,* 369(6477):221–224.

Sigman, D. M. and Hain, M. P. (2012) The biological productivity of the ocean. *Nat. Educ. Knowledge,* 3(10):21.

Simoncini, E., Virgo, N., and Kleidon, A. (2013). Quantifying drivers of chemical disequilibrium: Theory and application to methane in the Earth's atmosphere. *Earth Syst. Dynam.,* 4(2):317–331.

Simpson, F. (2017). Bayesian evidence for the prevalence of waterworlds. *Mon. Not. R. Astron. Soc.,* 468(3):2803–2815.

Simpson, G. G. (1964). The nonprevalence of humanoids. *Science,* 143(3608): 769–775.

Simpson, G. G. (1967). *The Meaning of Evolution*. Yale University Press, New Haven, CT.

Siraj, A. and Loeb, A. (2019a). Discovery of a meteor of interstellar origin. arXiv e-prints. arXiv:1904.07224.

Siraj, A. and Loeb, A. (2019b). Identifying interstellar objects trapped in the Solar system through their orbital parameters. *Astrophys. J. Lett.,* 872(1):L10.

Sleep, N. H., Bird, D. K., and Pope, E. C. (2011). Serpentinite and the dawn of life. *Phil. Trans. R. Soc. B,* 366(1580):2857–2869.

Sleep, N. H. (2018). Geological and geochemical constraints on the origin and evolution of life. *Astrobiology,* 18(9):1199–1219.

Slobodkin, A., Gavrilov, S., Ionov, V., and Iliyin, V. (2015). Spore-forming thermophilic bacterium within artificial meteorite survives entry into the earth's atmosphere on FOTON-M4 satellite landing module. *PLoS One,* 10(7): e0132611.

Slysh, V. I. (1985). A search in the infrared to microwave for astroengineering activity. In Papagiannis, M. D., editor, *IAU Symp. 112: The Search for Extraterrestrial*

Life: Recent Developments, pages 315–319. D. Reidel Publishing Co., Dordrecht, Netherlands.

Smart, J. M. (2012). The transcension hypothesis: Sufficiently advanced civilizations invariably leave our universe, and implications for METI and SETI. *Acta Astronaut.,* 78:55–68.

Smith, A. R., Carmody, R. N., Dutton, R. J., and Wrangham, R. W. (2015). The significance of cooking for early hominin scavenging. *J. Hum. Evol.,* 84:62–70.

Smith, D. J., Griffin, D. W., McPeters, R. D., Ward, P. D., and Schuerger, A. C. (2011). Microbial survival in the stratosphere and implications for global dispersal. *Aerobiologia,* 27(4):319–332.

Smith, F. A., Boyer, A. G., Brown, J. H., Costa, D. P., Dayan, T., Ernest, S. K. M., Evans, A. R., Fortelius, M., Gittleman, J. L., Hamilton, M. J., Harding, L. E., Lintulaakso, K., Lyons, K. S., McCain, C., Okie, J. G., Saarinen, J. J., Sibly, R. M., Stephens, P. R., Theodor, J., and Uhen, M. D. (2010). The evolution of maximum body size of terrestrial mammals. *Science,* 330(6008):1216–1219.

Smith, J. M. and Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford University Press, Oxford, UK.

Smith, K. C. and Mariscal, C. (Eds.) (2020). *Social and Conceptual Issues in Astrobiology*. Oxford University Press, Oxford, UK.

Smith, E. and Morowitz, H. J. (2004). Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. USA,* 101(36):13168–13173.

Smith, E. and Morowitz, H. J. (2016). *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere*. Cambridge University Press, Cambridge, UK.

Snellen, I., de Kok, R., Birkby, J. L., Brandl, B., Brogi, M., Keller, C., Kenworthy, M., Schwarz, H., and Stuik, R. (2015). Combining high-dispersion spectroscopy with high contrast imaging: Probing rocky planets around our nearest neighbors. *Astron. Astrophys.,* 576:A59.

Soai, K., Shibata, T., Morioka, H., and Choji, K. (1995). Asymmetric autocatalysis and amplification of enantiomeric excess of a chiral molecule. *Nature,* 378(6559):767–768.

Socas-Navarro, H. (2018). Possible photometric signatures of moderately advanced civilizations: The Clarke exobelt. *Astrophys. J.,* 855(2):110.

Sögütoglu, L.-C., Steendam, R. R. E., Meekes, H., Vlieg, E., and Rutjes, F. P. J. T. (2015). Viedma ripening: A reliable crystallisation method to reach single chirality. *Chem. Soc. Rev.,* 44(19):6723–6732.

Sojo, V., Herschy, B., Whicher, A., Camprubí, E., and Lane, N. (2016). The origin of life in alkaline hydrothermal vents. *Astrobiology,* 16(2):181–197.

Solé, R. V. and Munteanu, A. (2004). The large-scale organization of chemical reaction networks in astrophysics. *Europhys. Lett.,* 68(2):170–176.

Sotin, C., Grasset, O., and Mocquet, A. (2007). Mass radius curve for extrasolar Earth-like planets and ocean planets. *Icarus,* 191(1):337–351.

Sotos, J. G. (2019). Biotechnology and the lifetime of technical civilizations. *Int. J. Astrobiol.,* 18(5):445–454.

Sousa, F. L., Hordijk, W., Steel, M., and Martin, W. F. (2015). Autocatalytic sets in E. coli metabolism. *J. Syst. Chem,* 6:4.

Sousa-Silva, C., Seager, S., Ranjan, S., Petkowski, J. J., Zhan, Z., Hu, R., and Bains, W. (2020). Phosphine as a biosignature gas in exoplanet atmospheres. *Astrobiology,* 20(2):235–268.

Spang, A., Eme, L., Saw, J. H., Caceres, E. F., Zaremba-Niedzwiedzka, K., Lombard, J., Guy, L., and Ettema, T. J. G. (2018). Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.,* 14(3):e1007080.

Spang, A., Stairs, C. W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E. F., Greening, C., Baker, B. J., and Ettema, T. J. G. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.,* 4(7):1138–1148.

Sparks, W. B. and Ford, H. C. (2002). Imaging spectroscopy for extrasolar planet detection. *Astrophys. J.,* 578(1):543–564.

Sparks, W. B., Hough, J., Germer, T. A., Chen, F., DasSarma, S., DasSarma, P., Robb, F. T., Manset, N., Kolokolova, L., Reid, N., Macchetto, F. D., and Martin, W. (2009). Detection of circular polarization in light scattered from photosynthetic microbes. *Proc. Natl. Acad. Sci. USA,* 106(19):7816–7821.

Sparks, W. B., Schmidt, B. E., McGrath, M. A., Hand, K. P., Spencer, J. R., Cracraft, M., and Deustua, S. E. (2017). Active cryovolcanism on Europa? *Astrophys. J. Lett.,* 839(2):L18.

Sparks, W. B., White, R. L., Lupu, R. E., and Ford, H. C. (2018). The direct detection and characterization of M-dwarf planets using light echoes. *Astrophys. J.,* 854(2):134.

Spencer, D. F. and Jaffe, L. D. (1963). Feasibility of interstellar travel. *Astronautica Acta,* 9:49–58.

Spencer, D. A., Betts, B., Bellardo, J. M., Diaz, A., Plante, B., and Mansell, J. R. (2020). The LightSail 2 Solar sailing technology demonstration. *Adv. Space Res.,* DOI: 10.1016/j.asr.2020.06.029.

Sperling, E. A. and Stockey, R. G. (2018). The temporal and environmental context of early animal evolution: Considering all the ingredients of an "explosion." *Integr. Comp. Biol.,* 58(4):605–622.

Spiegel, D. S. and Turner, E. L. (2012). Bayesian analysis of the astrobiological implications of life's early emergence on Earth. *Proc. Natl. Acad. Sci. USA,* 109(2):395–400.

Spitzer, J., Pielak, G. J., and Poolman, B. (2015). Emergence of life: Physical chemistry changes the paradigm. *Biol. Direct,* 10:33.

Spohn, T. and Schubert, G. (2003). Oceans in the icy Galilean satellites of Jupiter? *Icarus,* 161(2):456–467.

Šponer, J. E., Šponer, J., and Di Mauro, E. (2017). New evolutionary insights into the non-enzymatic origin of RNA oligomers. *Wiley Interdiscip. Rev. RNA,* 8(3):e1400.

Springsteen, G., Yerabolu, J. R., Nelson, J., Rhea, C. J., and Krishnamurthy, R. (2018). Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun.,* 9:91.

Srivastava, P. R. and Swartzlander, G. A. (2020). Optomechanics of a stable diffractive axicon light sail. *Eur. Phys. J. Plus,* 135:570.

Stairs, C. W. and Ettema, T. J. G. (2020). The Archaeal roots of the eukaryotic dynamic actin cytoskeleton. *Curr. Biol.,* 30(10):R521–R526.

Stairs, S., Nikmal, A., Bučar, D.-K., Zheng, S.-L., Szostak, J. W., and Powner, M. W. (2017). Divergent prebiotic synthesis of pyrimidine and 8-oxo-purine ribonucleotides. *Nat. Commun.,* 8:15270.

Stalcup, A. M. (2010). Chiral separations. *Annu. Rev. Anal. Chem.,* 3:341–363.

Stamenković, V. and Seager, S. (2016). Emerging possibilities and insuperable limitations of exogeophysics: The example of plate tectonics. *Astrophys. J.,* 825(1):78.

Stamenković, V., Beegle, L. W., Zacny, K., Arumugam, D. D., Baglioni, P., Barba, N., Baross, J., Bell, M. S., Bhartia, R., Blank, J. G., Boston, P. J., Breuer, D., Brinckerhoff, W., Burgin, M. S., Cooper, I., Cormarkovic, V., Davila, A., Davis, R. M., Edwards, C., . . . Woolley, R. (2019). The next frontier for planetary and human exploration. *Nat. Astron.,* 3:116–120.

Stanton, R. H. (2019). Exploring optical SETI's middle ground. *Acta Astronaut.,* 156:92–99.

Stapledon, O. (1968). *'Last and First Men' and 'Star Maker'*. Dover Publications, Inc., New York, NY.

Stayton, C. T. (2015). What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus,* 5(6):20150039.

Steel, D. (1995). SETA and 1991 VG. *Observatory,* 115:78–83.

Steel, E. L., Davila, A., and McKay, C. P. (2017). Abiotic and biotic formation of amino acids in the Enceladus ocean. *Astrobiology,* 17(9):862–875.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science,* 347(6223):1259855.

Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., Donges,

J. F., Fetzer, I., Lade, S. J., Scheffer, M., Winkelmann, R., and Schellnhuber, H. J. (2018). Trajectories of the Earth system in the Anthropocene. *Proc. Natl. Acad. Sci. USA,* 115(33):8252–8259.

Stelmach, K. B., Neveu, M., Vick-Majors, T. J., Mickol, R. L., Chou, L., Webster, K. D., Tilley, M., Zacchei, F., Escudero, C., Flores Martinez, C. L., Labrado, A., and Fernández, E. J. G. (2018). Secondary electrons as an energy source for life. *Astrobiology,* 18(1):73–85.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. The MIT Press, Cambridge, MA.

Stern, R. J. (2016). Is plate tectonics needed to evolve technological species on exoplanets? *Geosci. Front.,* 7(4):573–580.

Stern, R. J. (2018). The evolution of plate tectonics. *Phil. Trans. R. Soc. A,* 376(2132):20170406.

Stern, R. J. and Gerya, T. (2018). Subduction initiation in nature and models: A review. *Tectonophysics,* 746:173–198.

Stern, S. A. (1986). The effects of mechanical interaction between the interstellar medium and comets. *Icarus,* 68(2):276–283.

Stevens, A., Forgan, D., and James, J. O. (2016). Observational signatures of self-destructive civilizations. *Int. J. Astrobiol.,* 15(4):333–344.

Stevenson, D. J. (1979). Turbulent thermal convection in the presence of rotation and a magnetic field: A heuristic theory. *Geophys. Astrophys. Fluid Dyn.,* 12(1):139–169.

Stevenson, D. J. (1999). Life-sustaining planets in interstellar space? *Nature,* 400(6739):32.

Stevenson, J., Lunine, J., and Clancy, P. (2015). Membrane alternatives in worlds without oxygen: Creation of an azotosome. *Sci. Adv.,* 1(1):e1400067.

Stewart, J. R. and Stringer, C. B. (2012). Human evolution out of Africa: The role of refugia and climate change. *Science,* 335(6074):1317–1321.

Stoeger, A. S., Mietchen, D., Oh, S., de Silva, S., Herbst, C. T., Kwon, S., and Fitch, W. T. (2012). An Asian elephant imitates human speech. *Curr. Biol.,* 22(22):2144–2148.

Stöffler, D., Horneck, G., Ott, S., Hornemann, U., Cockell, C. S., Moeller, R., Meyer, C., de Vera, J.-P., Fritz, J., and Artemieva, N. A. (2007). Experimental evidence for the potential impact ejection of viable microorganisms from Mars and Mars-like planets. *Icarus,* 186(2):585–588.

Stone, R. P. S., Wright, S. A., Drake, F., Muñoz, M., Treffers, R., and Werthimer, D. (2005). Lick Observatory optical SETI: Targeted search and new directions. *Astrobiology,* 5(5):604–611.

Stout, D. and Hecht, E. E. (2017). Evolutionary neuroscience of cumulative culture. *Proc. Natl. Acad. Sci. USA,* 114(30):7861–7868.

Stribling, R. and Miller, S. L. (1987). Energy yields for hydrogen cyanide and formaldehyde syntheses: The HCN and amino acid concentrations in the primitive ocean. *Orig. Life Evol. Biosph.,* 17(3):261–273.

Stringer, C. (2016). The origin and evolution of *Homo sapiens*. *Phil. Trans. R. Soc. B,* 371(1698):20150237.

Strona, G. and Bradshaw, C. J. A. (2018). Co-extinctions annihilate planetary life during extreme environmental change. *Sci. Rep.,* 8:16724.

Strulson, C. A., Molden, R. C., Keating, C. D., and Bevilacqua, P. C. (2012). RNA catalysis through compartmentalization. *Nat. Chem.,* 4(11):941–946.

Struve, O. (1952). Proposal for a project of high-precision stellar radial velocity work. *Observatory,* 72(870):199–200.

Stubbs, R. T., Yadav, M., Krishnamurthy, R., and Springsteen, G. (2020). A plausible metal-free ancestral analogue of the Krebs cycle composed entirely of α-ketoacids. *Nat. Chem.,* 12:1016–1022.

Stüeken, E. E., Anderson, R. E., Bowman, J. S., Brazelton, W. J., Colangelo-Lillis, J., Goldman, A. D., Som, S. M., and Baross, J. A. (2013). Did life originate from a global chemical reactor? *Geobiology,* 11(2):101–126.

Stull, M. A. (1979). On the significance of the apparent absence of extraterrestrials on Earth. *J. Br. Interplanet. Soc.,* 32:221–222.

Suddendorf, T. and Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behav. Brain Sci.,* 30(3):299–313.

Suddendorf, T. (2013). *The Gap: The Science of What Separates Us from Other Animals*. Basic Books, New York, NY.

Suffern, K. G. (1977). Some thoughts on Dyson spheres. *Proc. Astron. Soc. Aust.,* 3(2):177–179.

Sugahara, H. and Mimura, K. (2014). Glycine oligomerization up to triglycine by shock experiments simulating comet impacts. *Geochem. J.,* 48(1):51–62.

Suissa, G., Mandell, A. M., Wolf, E. T., Villanueva, G. L., Fauchez, T., and Koppapurapu, R. K. (2020). Dim Prospects for Transmission Spectra of Ocean Earths around M Stars. *Astrophys. J.,* 891(1):58.

Sultan, S. E. (2015). *Organism and Environment: Ecological Development, Niche Construction, and Adaptation*. Oxford University Press, Oxford, UK.

Summons, R. E., Albrecht, P., McDonald, G., and Moldowan, J. M. (2008). Molecular biosignatures. *Space Sci. Rev.,* 135:133–159.

Sutherland, J. D. (2010). Ribonucleotides. *Cold Spring Harb. Perspect. Biol.,* 2(4): a005439.

Sutherland, J. D. (2016). The origin of life—Out of the blue. *Angew. Chem. Int. Ed.,* 55(1):104–121.

Suzuki, R., Buck, J. R., and Tyack, P. L. (2006). Information entropy of humpback whale songs. *J. Acoust. Soc. Am.,* 119(3):1849–1866.

Swalwell, B., Dalla, S., Kahler, S., White, S. M., Ling, A., Viereck, R., and Veronig, A. (2018). The reported durations of GOES soft X-ray flares in different solar cycles. *Space Weather,* 16(6):660–666.

Syverson, D. D., Reinhard, C. T., Isson, T. T., Holstege, C., Katchinoff, J., Tutolo, B. M., Etschmann, B., Brugger, J., and Planavsky, N. J. (2020). Anoxic weathering of mafic oceanic crust promotes atmospheric oxygenation. arXiv e-prints. arXiv:2002.07667.

Számadó, S. and Szathmáry, E. (2006). Selective scenarios for the emergence of natural language. *Trends Ecol. Evol.,* 21(10):555–561.

Szathmáry, E. (2003). Why are there four letters in the genetic alphabet? *Nat. Rev. Genet.,* 4:995–1001.

Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. *Proc. Natl. Acad. Sci. USA,* 112(33):10104–10111.

Szostak, J. W. (2003). Functional information: Molecular messages. *Nature,* 423(6941):689.

Szostak, J. W. (2012a). Attempts to define life do not help to understand the origin of life. *J. Biomol. Struct. Dyn.,* 29(4): 599–600.

Szostak, J. W. (2012b). The eightfold path to non-enzymatic RNA replication. *J. Syst. Chem.,* 3:2.

Szostak, J. W. (2017). The narrow road to the deep past: In search of the chemistry of the origin of life. *Angew. Chem. Int. Ed.,* 56(37):11037–11043.

Szostak, J. W., Bartel, D. P., and Luisi, P. L. (2001). Synthesizing life. *Nature,* 409(6818):387–390.

Tabataba-Vakili, F., Grenfell, J. L., Grießmeier, J.-M., and Rauer, H. (2016). Atmospheric effects of stellar cosmic rays on Earth-like exoplanets orbiting M-dwarfs. *Astron. Astrophys.,* 585:A96.

Tagliabue, A., Bowie, A. R., Boyd, P. W., Buck, K. N., Johnson, K. S., and Saito, M. A. (2017). The integral role of iron in ocean biogeochemistry. *Nature,* 543(7643):51–59.

Taiz, L., Alkon, D., Draguhn, A., Murphy, A., Blatt, M., Hawes, C., Thiel, G., and Robinson, D. G. (2019). Plants neither possess nor require consciousness. *Trends Plant Sci.,* 24(8):677–687.

Takahashi, T., Mizuno, Y., and Shibata, K. (2016). Scaling relations in coronal mass ejections and energetic proton events associated with solar superflares. *Astrophys. J. Lett.,* 833(1):L8.

Takeuchi, N. and Hogeweg, P. (2012). Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life. *Phys. Life Rev.,* 9(3):219–263.

Takeuchi, Y., Furukawa, Y., Kobayashi, T., Sekine, T., Terada, N., and Kakegawa, T. (2020). Impact-induced amino acid formation on Hadean Earth and Noachian Mars. *Sci. Rep.,* 10:9220.

Tang, T. B. (1976). Supernovae as time markers in interstellar communication. *J. Br. Interplanet. Soc.,* 29:469–470.

Tarter, J. (2001). The search for extraterrestrial intelligence (SETI). *Annu. Rev. Astron. Astrophys.,* 39:511–548.

Tarter, J. C. (2007). The evolution of life in the Universe: Are we alone? *Highlights of Astronomy,* 14:14–29.

Tarter, J. C., Backus, P. R., Mancinelli, R. L., Aurnou, J. M., Backman, D. E., Basri, G. S., Boss, A. P., Clarke, A., Deming, D., Doyle, L. R., Feigelson, E. D., Freund, F., Grinspoon, D. H., Haberle, R. M., Hauck, S. A., Heath, M. J., Henry, T. J., Hollingsworth, J. L., Joshi, M. M., . . . Young, R. E. (2007). A reappraisal of the habitability of planets around M dwarf stars. *Astrobiology,* 7(1):30–65.

Tarter, J. C., Agrawal, A., Ackermann, R., Backus, P., Blair, S. K., Bradford, M. T., Harp, G. R., Jordan, J., Kilsdonk, T., Smolek, K. E., Richards, J., Ross, J., Shostak, G. S., and Vakoch, D. (2010). SETI turns 50: Five decades of progress in the search for extraterrestrial intelligence. *Proceedings Volume 7819, Instruments, Methods, and Missions for Astrobiology XIII,* 781902.

Tasker, E., Tan, J., Heng, K., Kane, S., Spiegel, D., Brasser, R., Casey, A., Desch, S., Dorn, C., Hernlund, J., Houser, C., Laneuville, M., Lasbleis, M., Libert, A.-S., Noack, L., Unterborn, C., and Wicks, J. (2017). The language of exoplanet ranking metrics needs to change. *Nat. Astron.,* 1:0042.

Taubner, R.-S., Olsson-Francis, K., Vance, S. D., Ramkissoon, N. K., Postberg, F., de Vera, J.-P., Antunes, A., Camprubi Casas, E., Sekine, Y., Noack, L., Barge, L., Goodman, J., Jebbar, M., Journaux, B., Karatekin, Ö., Klenner, F., Rabbow, E., Rettberg, P., Rückriemen-Bez, T., . . . Soderlund, K. M. (2020). Experimental and simulation efforts in the astrobiological exploration of exooceans. *Space Sci. Rev.,* 216(1):9.

Taverne, Y. J., Merkus, D., Bogers, A. J., Halliwell, B., Duncker, D. J., and Lyons, T. J. (2018). Reactive oxygen species: Radical factors in the evolution of animal life. *BioEssays,* 40(3):1700158.

Taylor, A. H. (2014). Corvid cognition. *Wiley Interdiscip. Rev. Cogn. Sci.,* 5(3): 361–372.

Teachey, A. and Kipping, D. M. (2018). Evidence for a large exomoon orbiting Kepler-1625b. *Sci. Adv.,* 4(10):eaav1784.

Teichert, J. S., Kruse, F. M., and Trapp, O. (2019). Direct prebiotic pathway to DNA nucleosides. *Angew. Chem. Int. Ed.,* 58(29):9944–9947.

Tellis, N. K. and Marcy, G. W. (2015). A search for optical laser emission using Keck HIRES. *Publ. Astron. Soc. Pac.,* 127(952):540.

Tellis, N. K. and Marcy, G. W. (2017). A search for laser emission with megawatt thresholds from 5600 FGKM stars. *Astron. J.,* 153(6):251.

Temple, R. (2007). The prehistory of panspermia: Astrophysical or metaphysical? *Int. J. Astrobiol.,* 6(2):169–180.

Tentrup, T. B. H., Hummel, T., Wolterink, T. A. W., Uppu, R., Mosk, A. P., and Pinkse, P. W. H. (2017). Transmitting more than 10 bit with a single photon. *Opt. Express,* 25(3):2826–2833.

Tepfer, D. and Leach, S. (2006). Plant seeds as model vectors for the transfer of life through space. *Astrophys. Space Sci.,* 306:69–75.

Thiel, J., Byrne, J. M., Kappler, A., Schink, B., and Pester, M. (2019). Pyrite formation from FeS and H_2S is mediated through microbial redox activity. *Proc. Natl. Acad. Sci. USA,* 116(14):6897–6902.

Thompson, P. O., Cummings, W. C., and Ha, S. J. (1986). Sounds, source levels, and associated behavior of humpback whales, Southeast Alaska. *J. Acoust. Soc. Am.,* 80(3):735–740.

Thomson, W. (1894). Presidential address to the British Association, Edinburgh, 1871. In *Popular Lectures and Addresses,* vol. 2, pp. 132–205. Macmillan & Co., London, UK.

Tian, F., Toon, O. B., Pavlov, A. A., and De Sterck, H. (2005). A hydrogen-rich early Earth atmosphere. *Science,* 308(5724):1014–1017.

Tian, F., France, K., Linsky, J. L., Mauas, P. J. D., and Vieytes, M. C. (2014). High stellar FUV/NUV ratio and oxygen contents in the atmospheres of potentially habitable planets. *Earth Planet. Sci. Lett.,* 385:22–27.

Tian, F., Kasting, J. F., and Zahnle, K. (2011). Revisiting HCN formation in Earth's early atmosphere. *Earth Planet. Sci. Lett.,* 308(3):417–423.

Tian, F. and Ida, S. (2015). Water contents of Earth-mass planets around M dwarfs. *Nat. Geosci.,* 8(3):177–180.

Tilgner, C. N. and Heinrichsen, I. (1998). A program to search for Dyson spheres With The Infrared Space Observatory. *Acta Astronaut.,* 42:607–612.

Tilley, M. A., Segura, A., Meadows, V. S., Hawley, S., and Davenport, J. (2019). Modeling repeated M-dwarf flaring at an Earth-like planet in the habitable zone: I. Atmospheric effects for an unmagnetized planet. *Astrobiology,* 19(1):64–86.

Timofeev, M. Y., Kardashev, N. S., and Promyslov, V. G. (2000). A search of the IRAS database for evidence of Dyson spheres. *Acta Astronaut.,* 46:655–659.

Tingay, S. J., Tremblay, C., Walsh, A. and Urquhart, R. (2016). An opportunistic search for extraterrestrial intelligence (SETI) with the Murchison Widefield Array. *Astrophys. J. Lett.,* 827(2):L22.

Tingay, S. J., Tremblay, C. D. and Croft, S. (2018). A search for extraterrestrial intelligence (SETI) toward the Galactic anticenter with the Murchison Widefield Array. *Astrophys. J.,* 856(1):31.

Tipler, F. J. (1980). Extraterrestrial intelligent beings do not exist. *Q. Jl. R. Astr. Soc.,* 21:267–281.

Tipler, F. J. (2003). Intelligent life in cosmology. *Int. J. Astrobiol.,* 2(2):141–148.

Tirard, S. (2017). J. B. S. Haldane and the origin of life. *J. Genet.,* 96(5):735–739.

Tirard, S., Morange, M., and Lazcano, A. (2010). The definition of life: A brief history of an elusive scientific endeavor. *Astrobiology,* 10(10):1003–1009.

Tjoa, J. N. K. Y., Mueller, M., and van der Tak, F. F. S. (2020). The subsurface habitability of small, icy exomoons. *Astron. Astrophys.,* 636:A50.

Todd, Z. R. and Öberg, K. I. (2020). Cometary delivery of hydrogen cyanide to the early Earth. *Astrobiology,* 20(9):1109–1120.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Tomasello, M. (2008). *Origins of Human Communication*. The MIT Press, Cambridge, MA.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press, Cambridge, MA.

Toner, J. D. and Catling, D. C. (2019). Alkaline lake settings for concentrated prebiotic cyanide and the origin of life. *Geochim. Cosmochim. Acta,* 260:124–132.

Toner, J. D. and Catling, D. C. (2020). A carbonate-rich lake solution to the phosphate problem of the origin of life. *Proc. Natl. Acad. Sci. USA,* 117(2): 883–888.

Tononi, G. and Koch, C. (2015). Consciousness: Here, there and everywhere? *Phil. Trans. R. Soc. B,* 370(1668):20140167.

Toparlak, O. D. and Mansy, S. S. (2019). Progress in synthesizing protocells. *Exp. Biol. Med.,* 244(4):304–313.

Townes, C. H. (1983). At what wavelengths should we search for signals from extraterrestrial intelligence? *Proc. Natl. Acad. Sci. USA,* 80(4):1147–1151.

Trainer, M. G., Pavlov, A. A., Dewitt, H. L., Jimenez, J. L., McKay, C. P., Toon, O. B., and Tolbert, M. A. (2006). Organic haze on Titan and the early Earth. *Proc. Natl. Acad. Sci. USA,* 103(48):18035–18042.

Traub, W. A. and Oppenheimer, B. R. (2010). Direct imaging of exoplanets. In Seager, S., editor, *Exoplanets,* pages 111–156. The University of Arizona Press, Tucson, AZ.

Treumann, R. A. (2006). The electron-cyclotron maser for astrophysical application. *Astron. Astrophys. Rev.,* 13(4):229–315.

Trevors, J. T. and Pollack, G. H. (2005). Hypothesis: The origin of life in a hydrogel environment. *Prog. Biophys. Mol. Biol.,* 89(1):1–8.

Trewavas, A. (2014). *Plant Behaviour and Intelligence*. Oxford University Press, Oxford, UK.

Trewavas, A. (2017). The foundations of plant intelligence. *Interface Focus,* 7(3): 20160098.

Trifonov, E. N. (2011). Vocabulary of definitions of life suggests a definition. *J. Biomol. Struct. Dyn.,* 29(2):259–266.

Trinks, H., Schröder, W., and Biebricher, C. K. (2005). Ice and the origin of life. *Orig. Life Evol. Biosph.,* 35(5):429–445.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.,* 46(1): 35–57.

Tsander, F. A. (1924). Flights to other planets. *Tekhnika i Zhizn',* 13:15–16.

Tsander, F. A. (1964). *Problems of Flight by Jet Propulsion: Interplanetary Flights*. Translated by IPST Staff. Israel Program for Scientific Translations, Ltd., Jerusalem, Israel.

Tsapras, Y. (2018). Microlensing searches for exoplanets. *Geosciences,* 8(10):365.

Tsiolkovsky, K. E. (1903). The investigation of space by means of reactive devices. *Nauchnoe Obozrenie,* 5:3.

Tsokolov, S. A. (2009). Why is the definition of life so elusive? Epistemological considerations. *Astrobiology,* 9(4):401–412.

Tsokolov, S. (2010). A theory of circular organization and negative feedback: Defining life in a cybernetic context. *Astrobiology,* 10(10):1031–1042.

Tsuda, Y., Mori, O., Funase, R., Sawada, H., Yamamoto, T., Saiki, T., Endo, T., and Kawaguchi, J. (2011). Flight status of IKAROS deep space solar sail demonstrator. *Acta Astronaut.,* 69:833–840.

Tuck, A. (2002). The role of atmospheric aerosols in the origin of life. *Surv. Geophys.,* 23(5):379–409.

Turbet, M., Bolmont, E., Bourrier, V., Demory, B.-O., Leconte, J., Owen, J. and Wolf, E. T. (2020). A review of possible planetary atmospheres in the TRAPPIST-1 system. *Space Sci. Rev.,* 216(5):100.

Turner, S., Wilde, S., Wörner, G., Schaefer, B., and Lai, Y.-J. (2020). An andesitic source for Jack Hills zircon supports onset of plate tectonics in the Hadean. *Nat. Commun.,* 11:1241.

Turyshev, S. G., Klupar, P., Loeb, A., Manchester, Z., Parkin, K., Witten, E., and Worden, S. P. (2020). Exploration of the outer solar system with fast and small sailcraft. arXiv e-prints. arXiv:2005.12336.

Tuttle, R. H. (2014). *Apes and Human Evolution*. Harvard University Press, Cambridge, MA.

Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature,* 400(6744):525–531.

Tyrrell, T. (2013). *On Gaia: A Critical Investigation of the Relationship between Life and Earth*. Princeton University Press, Princeton, NJ.

Udry, S. and Santos, N. C. (2007). Statistical properties of exoplanets. *Annu. Rev. Astron. Astrophys.,* 45:397–439.

Unterborn, C. T., Johnson, J. A., and Panero, W. R. (2015). Thorium abundances in solar twins and analogs: Implications for the habitability of extrasolar planetary systems. *Astrophys. J.,* 806(1):139.

Unterborn, C. T., Desch, S. J., Hinkel, N. R., and Lorenzo, A. (2018). Inward migration of the TRAPPIST-1 planets as inferred from their water-rich compositions. *Nat. Astron.,* 2:297–302.

Unterborn, C. T., Hinkel, N. R., and Desch, S. J. (2018). Updated compositional models of the TRAPPIST-1 planets. *Res. Notes AAS,* 2(3):116.

Urey, H. C. (1952). On the early chemical history of the Earth and the origin of life. *Proc. Natl. Acad. Sci. USA,* 38(4):351–363.

Usoskin, I. G. (2017). A history of solar activity over millennia. *Living Rev. Sol. Phys.,* 14:3.

Vaesen, K. (2012). The cognitive bases of human tool use. *Behav. Brain Sci.,* 35(4):203–218.

Vaesen, K., Collard, M., Cosgrove, R., and Roebroeks, W. (2016). Population size does not explain past changes in cultural complexity. *Proc. Natl. Acad. Sci. USA,* 113(16):E2241–E2247.

Vaidya, N., Manapat, M. L., Chen, I. A., Xulvi-Brunet, R., Hayden, E. J., and Lehman, N. (2012). Spontaneous network formation among cooperative RNA replicators. *Nature,* 491(7422):72–77.

Vakoch, D. A. (Ed.) (2013). *Astrobiology, History, and Society: Life beyond Earth and the Impact of Discovery*. Springer-Verlag, Berlin, Germany.

Vakoch, D. A. and Dowd, M. F. (Eds.) (2015). *The Drake Equation*. Cambridge University Press, Cambridge, UK.

Valencia, D. and O'Connell, R. J. (2009). Convection scaling and subduction on Earth and super-Earths. *Earth Planet. Sci. Lett.,* 286:492–502.

Valencia, D., Tan, V. Y. Y., and Zajac, Z. (2018). Habitability from tidally induced tectonics. *Astrophys. J.,* 857(2):106.

Valentine, J. W. and Moores, E. M. (1970). Plate-tectonic regulation of faunal diversity and sea level: A model. *Nature,* 228(5272):657–659.

Vallverdú, J., Castro, O., Mayne, R., T., Levin, M., Baluška, F., Gunji, Y., Dussutour, A., Zenil, H., and Adamatzky, A. (2018). Slime mould: The fundamental mechanisms of biological cognition. *Biosystems,* 165:57–70.

Valtonen, M., Nurmi, P., Zheng, J.-Q., Cucinotta, F. A., Wilson, J. W., Horneck, G., Lindegren, L., Melosh, J., Rickman, H., and Mileikowsky, C. (2009). Natural transfer of viable microbes in space from planets in extra-Solar systems to a planet in our Solar system and vice versa. *Astrophys. J.,* 690(1): 210–215.

Vance, S., Harnmeijer, J., Kimura, J., Hussmann, H., deMartin, B., and Brown, J. M. (2007). Hydrothermal Systems in Small Ocean Planets. *Astrobiology,* 7(6):987–1005.

Vance, S. D., Hand, K. P., and Pappalardo, R. T. (2016). Geophysical controls of chemical disequilibria in Europa. *Geophys. Res. Lett.,* 43(10):4871–4879.

Vance, S. D., Panning, M. P., Stähler, S., Cammarano, F., Bills, B. G., Tobie, G., Kamata, S., Kedar, S., Sotin, C., Pike, W. T., Lorenz, R., Huang, H.-H., Jackson, J. M., and Banerdt, B. (2018). Geophysical investigations of habitability in ice-covered ocean worlds. *J. Geophys. Res. Planets,* 123(1):180–205.

van Gestel, J. and Tarnita, C. E. (2017). On the origin of biological construction, with a focus on multicellularity. *Proc. Natl. Acad. Sci. USA,* 114(42):11018–11026.

Van Lawick-Goodall, J. (1971). Tool-using in primates and other vertebrates. *Adv. Study Behav.,* 3:195–249.

Van Schaik, C. P. (2016). *The Primate Origins of Human Nature*. Wiley-Blackwell, Hoboken, NJ.

Van Valen, L. (1973). A new evolutionary law. *Evol. Theory,* 1:1–30.

Varma, S. J., Muchowska, K. B., Chatelain, P., and Moran, J. (2018). Native iron reduces CO_2 to intermediates and end-products of the acetyl-CoA pathway. *Nat. Ecol. Evol.,* 2:1019–1024.

Vasas, V., Fernando, C., Santos, M., Kauffman, S., and Szathmáry, E. (2012). Evolution before genes. *Biol. Direct,* 7:1.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behav. Brain Sci.,* 43(e90):1–75.

Veras, D., Armstrong, D. J., Blake, J. A., Gutiérrez-Marcos, J. F., Jackson, A. P., and Schäefer, H. (2018). Dynamical and biological panspermia constraints within multiplanet exosystems. *Astrobiology,* 18(9):1106–1122.

Vermeij, G. J. (2006). Historical contingency and the purported uniqueness of evolutionary innovations. *Proc. Natl. Acad. Sci. USA,* 103(6):1804–1809.

Vermeij, G. J. (2017). How the land became the locus of major evolutionary innovations. *Curr. Biol.,* 27(20):3178–3182.

Vernadsky, V. I. (1926). *The Biosphere*. Scientific Chemico-Technical Publishing, Leningrad, USSR.

Vernadsky, W. I. (1945). The biosphere and the noösphere. *Am. Sci.,* 33(1): 1–12.

Vida, K., Kővári, Z., Pál, A., Oláh, K., and Kriskovics, L. (2017). Frequent flaring in the TRAPPIST-1 system–unsuited for life? *Astrophys. J.,* 841(2):124.

Vida, K., Leitzinger, M., Kriskovics, L., Seli, B., Odert, P., Kovács, O. E., Korhonen, H., and van Driel-Gesztelyi, L. (2019). The quest for stellar coronal mass ejections in late-type stars. I. Investigating Balmer-line asymmetries of single stars in Virtual Observatory data. *Astron. Astrophys.,* 623:A49.

Vida, K., Oláh, K., Kővári, Z., van Driel-Gesztelyi, L., Moór, A., and Pál, A. (2019). Flaring activity of Proxima Centauri from *TESS* observations: Quasiperiodic oscillations during flare decay and inferences on the habitability of Proxima b. *Astrophys. J.,* 884(2):160.

Vidal, C. (2014). *The Beginning and the End: The Meaning of Life in a Cosmological Perspective*. Springer, New York, NY.

Vidotto, A. A., Jardine, M., Morin, J., Donati, J.-F., Lang, P., and Russell, A. B. (2013). Effects of M dwarf magnetic fields on potentially habitable planets. *Astron. Astrophys.,* 557:A67.

Vidotto, A. A. and Donati, J.-F. (2017). Predicting radio emission from the newborn hot Jupiter V830 Tauri b and its host star. *Astron. Astrophys.,* 602:A39.

Viedma, C. (2005). Chiral symmetry breaking during crystallization: Complete chiral purity induced by nonlinear autocatalysis and recycling. *Phys. Rev. Lett.,* 94(6):065504.

Viewing, D. (1975). Directly interacting extra-terrestrial technological communities. *J. Br. Interplanet. Soc.,* 28:735–744.

Viewing, D. R. J., Horswell, C. J., and Palmer, E. W. (1977). Detection of starships. *J. Br. Interplanet. Soc.,* 30:99–104.

Villarroel, B., Imaz, I., and Bergstedt, J. (2016). Our sky now and then: Searches for lost stars and impossible effects as probes of advanced extraterrestrial civilizations. *Astron. J.,* 152(3):76.

Villarroel, B., Soodla, J., Comerón, S., Mattsson, L., Pelckmans, K., López-Corredoira, M., Krisciunas, K., Guerras, E., Kochukhov, O., Bergstedt, J., Buelens, B., Bär, R. E., Cubo, R. Enriquez, J. E., Gupta, A. C., Imaz, I., Karlsson, T., Prieto, M. A., Shlyapnikov, . . . Ward, M. J. (2020). The vanishing and appearing sources during a century of observations project. I. Usno objects missing in modern sky surveys and follow-up observations of a "missing star." *Astron. J.,* 159(1):8.

Virgil. (1875). *P. Vergili Maronis Opera: The Works of Virgil,* vol. 3. Edited by G. Long. Whittaker & Co., London, UK (2nd edition).

Visscher, C., Lodders, K., and Fegley, B. (2006). Atmospheric chemistry in giant planets, brown dwarfs, and low-mass dwarf stars. II. Sulfur and phosphorus. *Astrophys. J.,* 648(2):1181–1195.

Vitas, M. and Dobovišek, A. (2019). Towards a general definition of life. *Orig. Life Evol. Biosph.,* 49:77–88.

Vladilo, G. and Hassanali, A. (2018). Hydrogen bonds and life in the Universe. *Life,* 8(1):1.

von Braun, W. and Ordway, F. I. (1975). *History of Rocketry & Space Travel*. Thomas Y. Crowell Co., New York, NY (3rd edition).

von Helmholtz, H. (1893). On the origin of the planetary system. In *Popular Lectures on Scientific Subjects,* Vol. II, pp. 139–197. Translated by E. Atkinson. Longmans, Green & Co., London, UK.

von Hoerner, S. (1961). The search for signals from other civilizations. *Science,* 134(3493):1839–1843.

von Hoerner, S. (1962). The general limits of space travel. *Science,* 137(3523):18–23.

von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press, Urbana, IL.

Vreeland, R. H., Rosenzweig, W. D., and Powers, D. W. (2000). Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature,* 407(6806):897–900.

Vukotić, B. and Ćirković, M. M. (2012). Astrobiological complexity with probabilistic cellular automata. *Orig. Life Evol. Biosph.,* 42(4):347–371.

Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J., and Brasier, M. D. (2011). Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.,* 4(10):698–702.

Wachowius, F., Attwater, J., and Holliger, P. (2017). Nucleic acids: Function and potential for abiogenesis. *Q. Rev. Biophys.,* 50:e4.

Wächtershäuser, G. (1988). Before enzymes and templates: Theory of surface metabolism. *Microbiol. Rev.,* 52(4):452–484.

Wächtershäuser, G. (1990). Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. USA,* 87(1):200–204.

Wächtershäuser, G. (2007). On the chemistry and evolution of the pioneer organism. *Chem. Biodivers.,* 4(4):584–602.

Wagner, A. J., Zubarev, D. Y., Aspuru-Guzik, A., and Blackmond, D. G. (2017). Chiral sugars drive enantioenrichment in prebiotic amino acid synthesis. *ACS Cent. Sci.,* 3(4):322–328.

Waite, J. H., Glein, C. R., Perryman, R. S., Teolis, B. D., Magee, B. A., Miller, G., Grimes, J., Perry, M. E., Miller, K. E., Bouquet, A., Lunine, J. I., Brockwell, T., and Bolton, S. J. (2017). Cassini finds molecular hydrogen in the Enceladus plume: Evidence for hydrothermal processes. *Science,* 356(6334): 155–159.

Walker, J. C. G., Hays, P. B., and Kasting, J. F. (1981). A negative feedback mechanism for the long-term stabilization of the Earth's surface temperature. *J. Geophys. Res.,* 86(C10):9776–9782.

Walker, S. I. (2017). Origins of life: A problem for physics, a key issues review. *Rep. Prog. Phys.,* 80(9):092601.

Walker, S. I., Bains, W., Cronin, L., DasSarma, S., Danielache, S., Domagal-Goldman, S., Kacar, B., Kiang, N. Y., Lenardic, A., Reinhard, C. T., Moore, W., Schwieterman, E. W., Shkolnik, E. L. and Smith, H. B. (2018). Exoplanet biosignatures: Future directions. *Astrobiology,* 18(6):779–824.

Wallis, M. K. and Wickramasinghe, N. C. (2004). Interstellar transfer of planetary microbiota. *Mon. Not. R. Astron. Soc.,* 348(1):52–61.

Wallmann, K. (2010). Phosphorus imbalance in the global ocean? *Global Biogeochem. Cy.,* 24:GB4030.

Walter, K. U., Vamvaca, K., and Hilvert, D. (2005). An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.,* 280(45):37742–37746.

Walters, C., Hoover, R. A. and Kotra, R. K. (1980). Interstellar colonization—A new parameter for the Drake equation. *Icarus,* 41:193–197.

Waltham, D. (2019). Is Earth special? *Earth-Sci. Rev.,* 192:445–470.

Wang, J., Mawet, D., Ruane, G., Hu, R., and Benneke, B. (2017). Observing exoplanets with high dispersion coronagraphy. I. The scientific potential of current and next-generation large ground and space telescopes. *Astron. J.,* 153(4):183.

Wang, Z. and Cuntz, M. (2019). S-type and P-type habitability in stellar binary systems: A comprehensive approach. III. Results for Mars, Earth, and super-Earth planets. *Astrophys. J.,* 873(2):113.

Ward, L. M., Kirschvink, J. L., and Fischer, W. W. (2016). Timescales of oxygenation following the evolution of oxygenic photosynthesis. *Orig. Life Evol. Biosph.,* 46(1):51–65.

Ward, L. M., Rasmussen, B., and Fischer, W. W. (2019). Primary productivity was limited by electron donors prior to the advent of oxygenic photosynthesis. *J. Geophys. Res. Biogeosci.,* 124(2):211–226.

Ward, P. and Brownlee, D. (2000). *Rare Earth: Why Complex Life Is Uncommon in the Universe*. Copernicus Books, New York, NY.

Wargelin, B. J. and Drake, J. J. (2002). Stringent X-ray constraints on mass loss from Proxima Centauri. *Astrophys. J.,* 578(1):503–514.

Warke, M. R., Di Rocco, T., Zerkle, A. L., Lepland, A., Prave, A. R., Martin, A. P., Ueno, Y., Condon, D. J., and Claire, M. W. (2020). The Great Oxidation Event preceded a Paleoproterozoic "snowball Earth." *Proc. Natl. Acad. Sci. USA,* 117(24):13314–13320.

Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, A. D., Poirier, C., Gałuszka, A., Cearreta, A., Edgeworth, M., Ellis, E. C., Ellis, M., Jeandel, C., Leinfelder, R., McNeill, J. R., Richter, D. deB., Steffen, W., Syvitski, J., Vidas, D., Wagreich, M., Williams, M., . . . and Wolfe, A. P. (2016). The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science,* 351(6269):aad2622.

Watson, A. J. (2008). Implications of an anthropic model of evolution for emergence of complex life and intelligence. *Astrobiology,* 8(1):175–185.

Watson, R. A. and Szathmáry, E. (2016). How can evolution learn? *Trends Ecol. Evol.,* 31(2):147–157.

Webb, D. F. and Howard, T. A. (2012). Coronal Mass Ejections: Observations. *Living Rev. Sol. Phys.,* 9:3.

Webb, S. (2015). *If the Universe Is Teeming with Aliens . . . Where Is Everybody?* Springer, New York, NY.

Weber, A. L. and Miller, S. L. (1981). Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.,* 17(5):273–284.

Weintraub, D. A. (2014). *Religions and Extraterrestrial Life: How Will We Deal with It?* Springer, Cham, Switzerland.

Weiss, B. P., Kirschvink, J. L., Baudenbacher, F. J., Vali, H., Peters, N. T., Macdonald, F. A., and Wikswo, J. P. (2000). A low temperature transfer of ALH84001 from Mars to Earth. *Science,* 290(5492):791–795.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. (2016). The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.,* 1:16116.

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., and Martin, W. F. (2018). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet.,* 14(8):e1007518.

Weller, M. B. and Kiefer, W. S. (2020). The physics of changing tectonic regimes: Implications for the temporal evolution of mantle convection and the thermal history of Venus. *J. Geophys. Res. Planets,* 125(1):e2019JE005960.

Wells, L. E., Armstrong, J. C., and Gonzalez, G. (2003). Reseeding of early Earth by impacts of returning ejecta during the late heavy bombardment. *Icarus,* 162(1):38–46.

Wertz, J. R. (1976). The human analogy and the evolution of extraterrestrial civilisations. *J. Br. Interplanet. Soc.,* 29:445–464.

Wesson, P. S. (2010). Panspermia, past and present: Astrophysical and biophysical conditions for the dissemination of life in space. *Space Sci. Rev.,* 156: 239–252.

West, G. B., Brown, J. H., and Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. *Science,* 276(5309):122–126.

West, S. A., Fisher, R. M., Gardner, A., and Kiers, E. T. (2015). Major evolutionary transitions in individuality. *Proc. Natl. Acad. Sci. USA,* 112(33):10112–10119.

West, S. A. and Cooper, G. A. (2016). Division of labour in microorganisms: An evolutionary perspective. *Nat. Rev. Microbiol.,* 14(11):716–723.

Westall, F. and Brack, A. (2018). The importance of water for life. *Space Sci. Rev.,* 214(2):50.

Westall, F., Hickman-Lewis, K., Hinman, N., Gautret, P., Campbell, K. A., Bréhéret, J. G., Foucher, F., Hubert, A., Sorieul, S., Dass, A. V., Kee, T. P., Georgelin, T., and Brack, A. (2018). A hydrothermal-sedimentary context for the origin of life. *Astrobiology,* 18(3):259–293.

Westby, T. and Conselice, C. J. (2020). The astrobiological Copernican weak and strong limits for intelligent life. *Astrophys. J.,* 896(1):58.

Westerhoff, H. V., Brooks, A. N., Simeonidis, E., García-Contreras, R., He, F., Boogerd, F. C., Jackson, V. J., Goncharuk, V., and Kolodkin, A. (2014). Macromolecular networks and intelligence in microorganisms. *Front Microbiol.,* 5:379.

Westheimer, F. H. (1987). Why nature chose phosphates. *Science,* 235(4793): 1173–1178.

Westmoreland, S. (2010). A note on relativistic rocketry. *Acta Astronaut.,* 67:1248–1251.

Wheat, C. G., Feely, R. A., and Mottl, M. J. (1996). Phosphate removal by oceanic hydrothermal processes: An update of the phosphorus budget in the oceans. *Geochim. Cosmochim. Acta,* 60(19):3593–3608.

Wheat, C. G., McManus, J., Mottl, M. J., and Giambalvo, E. (2003). Oceanic phosphorus imbalance: Magnitude of the mid-ocean ridge flank hydrothermal sink. *Geophys. Res. Lett.,* 30(17):1895.

Whicher, A., Camprubi, E., Pinna, S., Herschy, B., and Lane, N. (2018). Acetyl phosphate as a primordial energy currency at the origin of life. *Orig. Life Evol. Biosph.,* 48(2):159–179.

Whipple, F. L. (1938). Photographic meteor studies, I. *Proc. Am. Philos. Soc.,* 79(4): 499–548.

White, T. D., Lovejoy, C. O., Asfaw, B., Carlson, J. P., and Suwa, G. (2015). Neither chimpanzee nor human, Ardipithecus reveals the surprising ancestry of both. *Proc. Natl. Acad. Sci. USA,* 112(16):4877–4884.

Whitehead, H. and Rendell, L. (2015). *The Cultural Lives of Whales and Dolphins*. The University of Chicago Press, Chicago, IL.

Whitehead, H. (2017). Gene-culture coevolution in whales and dolphins. *Proc. Natl. Acad. Sci. USA,* 114(30):7814–7821.

Whitehead, H., Laland, K. N., Rendell, L., Thorogood, R., and Whiten, A. (2019). The reach of gene-culture coevolution in animals. *Nat. Commun.,* 10: 2405.

Whiten, A. and Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Phil. Trans. R. Soc. B,* 367(1599):2119–2129.

Whiten, A. (2017). Culture extends the scope of evolutionary biology in the great apes. *Proc. Natl. Acad. Sci. USA,* 114(30):7790–7797.

Whiten, A. (2019). Cultural evolution in animals. *Annu. Rev. Ecol. Evol. Syst.,* 50: 27–48.

Whitmire, D. P. (1975). Relativistic spaceflight and the catalytic nuclear ramjet. *Acta Astronaut.,* 2:497–509.

Whitmire, D. P. and Wright, D. P. (1980). Nuclear waste spectrum as evidence of technological extraterrestrial civilizations. *Icarus,* 42(1):149–156.

Wickramasinghe, C. (2010). The astrobiological case for our cosmic ancestry. *Int. J. Astrobiol.,* 9(2):119–129.

Wickramasinghe, N. C., Wickramasinghe, D. T., Tout, C. A., Lattanzio, J. C., and Steele, E. J. (2019). Cosmic biology in perspective. *Astrophys. Space Sci.,* 364(11):205.

Wierzchos, J., Casero, M. C., Artieda, O., and Ascaso, C. (2018). Endolithic microbial habitats as refuges for life in polyextreme environment of the Atacama Desert. *Curr. Opin. Microbiol.,* 43:124–131.

Wilde, S. A., Valley, J. W., Peck, W. H., and Graham, C. M. (2001). Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature,* 409(6817):175–178.

Williams, D. M. and Gaidos, E. (2008). Detecting the glint of starlight on the oceans of distant planets. *Icarus,* 195(2):927–937.

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., and Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.,* 4(1):138–147.

Wilson, E. O. (2012). *The Social Conquest of Earth*. W. W. Norton & Co., New York, NY.

Wilson, E. O. (2016). *Half-Earth: Our Planet's Fight for Life*. Liveright Publishing Corporation, New York, NY.

Wilson, F. R. (1998). *The Hand: How Its Use Shapes the Brain, Language, and Human Culture*. Pantheon Books, New York, NY.

Winn, J. N. (2010). Exoplanet transits and occultations. In Seager, S., editor, *Exoplanets,* pages 55–77. The University of Arizona Press, Tucson, AZ.

Winn, J. N. and Fabrycky, D. C. (2015). The occurrence and architecture of exoplanetary systems. *Annu. Rev. Astron. Astrophys.,* 53:409–447.

Wissner-Gross, A. D. and Freer, C. E. (2013). Causal entropic forces. *Phys. Rev. Lett.,* 110(16):168702.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA,* 74(11):5088–5090.

Woese, C. R. (1979). A proposal concerning the origin of life on the planet Earth. *J. Mol. Evol.,* 13(2):95–101.

Wogan, N. F. and Catling, D. C. (2020). When is chemical disequilibrium in Earth-like planetary atmospheres a biosignature versus an anti-biosignature? Disequilibria from dead to living worlds. *Astrophys. J.,* 892(2):127.

Wolf, E. T. and Toon, O. B. (2010). Fractal organic hazes provided an ultraviolet shield for early Earth. *Science,* 328(5983):1266–1268.

Wolf, E. T. and Toon, O. B. (2015). The evolution of habitable climates under the brightening Sun. *J. Geophys. Res. Atmos.,* 120(12):5775–5794.

Wolfe-Simon, F., Blum, J. S., Kulp, T. R., Gordon, G. W., Hoeft, S. E., Pett-Ridge, J., Stolz, J. F., Webb, S. M., Weber, P. K., Davies, P. C. W., Anbar, A. D., and Oremland, R. S. (2011). A bacterium that can grow by using arsenic instead of phosphorus. *Science,* 332(6034):1163–1166.

Wolstencroft, R. D. and Raven, J. A. (2002). Photosynthesis: Likelihood of occurrence and possibility of detection on Earth-like planets. *Icarus,* 157(2):535–548.

Wong, M. L., Charnay, B. D., Gao, P., Yung, Y. L., and Russell, M. J. (2017). Nitrogen oxides in early earth's atmosphere as electron acceptors for life's emergence. *Astrobiology,* 17(10):975–983.

Wood, R., Liu, A. G., Bowyer, F., Wilby, P. R., Dunn, F. S., Kenchington, C. G., Hoyal Cuthill, J. F., Mitchell, E. G., and Penny, A. (2019). Integrated records of environmental change and evolution challenge the Cambrian explosion. *Nat. Ecol. Evol.,* 3:528–538.

Woolf, N. and Angel, J. R. (1998). Astronomical searches for Earth-like planets and signs of life. *Annu. Rev. Astron. Astrophys.,* 36:507–538.

Worden, S. P., Drew, J., Siemion, A., Werthimer, D., DeBoer, D., Croft, S., MacMahon, D., Lebofsky, M., Isaacson, H., Hickish, J., Price, D., Gajjar, V. and Wright, J. T. (2017). Breakthrough Listen—A new search for life in the Universe. *Acta Astronaut.,* 139:98–101.

Wordsworth, R. (2015). Atmospheric heat redistribution and collapse on tidally locked rocky planets. *Astrophys. J.,* 806(2):180.

Wordsworth, R. D., Schaefer, L. K., and Fischer, R. A. (2018). Redox evolution via gravitational differentiation on low-mass planets: Implications for abiotic oxygen, water loss, and habitability. *Astron. J.,* 155(5):25.

Worth, R. J., Sigurdsson, S., and House, C. H. (2013). Seeding life on the moons of the outer planets via lithopanspermia. *Astrobiology,* 13(12):1155–1165.

Wrangham, R. (2009). *Catching Fire: How Cooking Made Us Human*. Basic Books, New York, NY.

Wrangham, R. (2017). Control of fire in the Paleolithic: Evaluating the cooking hypothesis. *Curr. Anthropol.,* 58(S16):S303–S313.

Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship between Virtue and Violence in Human Evolution*. Vintage Books, New York, NY.

Wright, J. S., Fu, R., Worden, J. R., Chakraborty, S., Clinton, N. E., Risi, C., Sun, Y., and Yin, L. (2017). Rainforest-initiated wet season onset over the southern Amazon. *Proc. Natl. Acad. Sci. USA,* 114(32):8481–8486.

Wright, J. T., Cartier, K. M. S., Zhao, M., Jontof-Hutter, D., and Ford, E. B. (2016). The search for extraterrestrial civilizations with large energy supplies. IV. The signatures and information content of transiting megastructures. *Astrophys. J.,* 816(1):17.

Wright, J. T., Mullan, B., Sigurdsson, S. and Povich, M. S. (2014). The $\hat{\text{G}}$ infrared search for extraterrestrial civilizations with large energy supplies. I. Background and justification. *Astrophys. J.,* 792(1):26.

Wright, J. T., Griffith, R. L., Sigurdsson, S., Povich, M. S., and Mullan, B. (2014). The $\hat{\text{G}}$ infrared search for extraterrestrial civilizations with large energy supplies. II. Framework, strategy, and first result. *Astrophys. J.,* 792(1):27.

Wright, J. T. (2018a). Exoplanets and SETI. In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets,* pages 3405–3412. Springer, Cham, Switzerland.

Wright, J. T. (2018b). Prior indigenous technological species. *Int. J. Astrobiol.,* 17(1):96–100.

Wright, J. T. (2018c). Radial velocities as an exoplanet discovery method. In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets,* pages 619–631. Springer, Cham, Switzerland.

Wright, J. T. and Oman-Reagan, M. P. (2018). Visions of human futures in space and SETI. *Int. J. Astrobiol.,* 17(2):177–188.

Wright, J. T., Kanodia, S., and Lubar, E. (2018). How much SETI has been done? Finding needles in the n-dimensional cosmic haystack. *Astron. J.,* 156(6):260.

Wright, J. T. (2020). Dyson spheres. *Serb. Astron. J.,* 200:1–18.

Wright, S. A., Horowitz, P., Maire, J., Werthimer, D., Antonio, F., Aronson, M., Chaim-Weismann, S., Cosens, M., Drake, F. D., Howard, A. W., Marcy, G. W., Raffanti, R., Siemion, A. P. V., Stone, R. P. S., Treffers, R. R. and Uttamchandani, A. (2018). Panoramic optical and near-infrared SETI instrument: Overall specifications and science program. *Proceedings Volume 10702, Ground-based and Airborne Instrumentation for Astronomy VII,* 107025I.

Wu, L.-F. and Sutherland, J. D. (2019). Provisioning the origin and early evolution of life. *Emerg. Top. Life. Sci.,* 3(5):459–468.

Xavier, J. C., Hordijk, W., Kauffman, S., Steel, M., and Martin, W. F. (2020). Autocatalytic chemical networks at the origin of metabolism. *Proc. R. Soc. B,* 287(1922):20192377.

Xu, J., Ritson, D. J., Ranjan, S., Todd, Z. R., Sasselov, D. D., and Sutherland, J. D. (2018). Photochemical reductive homologation of hydrogen cyanide using sulfite and ferrocyanide. *Chem. Commun.,* 54(44):5566–5569.

Xu, J., Green, N. J., Gibard, C., Krishnamurthy, R., and Sutherland, J. D. (2019). Prebiotic phosphorylation of 2-thiouridine provides either nucleotides or DNA building blocks via photoreduction. *Nat. Chem.,* 11(5):457–462.

Xu, J., Chmela, V., Green, N. J., Russell, D. A., Janicki, M. J., Góra, R. W., Szabla, R., Bond, A. D., and Sutherland, J. D. (2020). Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature,* 582(7810): 60–66.

Xue, M., Black, R., Cornell, C. E., Drobny, G. P., and Keller, S. L. (2020). A step toward molecular evolution of RNA: Ribose binds to prebiotic fatty acid membranes, and nucleosides bind better than individual bases do. *ChemBioChem,* 21(19):2764–2767.

Yadav, M., Kumar, R., and Krishnamurthy, R. (2020). Chemistry of abiotic nucleotide synthesis. *Chem. Rev.,* 120(11):4766–4805.

Yahalomi, D., Atkinson, S. D., Neuhof, M., Chang, E. S., Philippe, H., Cartwright, P., Bartholomew, J. L., and Huchon, D. (2020). A cnidarian parasite of salmon (Myxozoa: *Henneguya*) lacks a mitochondrial genome. *Proc. Natl. Acad. Sci. USA,* 117(10):5358–5363.

Yamaguchi, M., Mori, Y., Kozuka, Y., Okada, H., Uematsu, K., Tame, A., Furukawa, H., Maruyama, T., O'Driscoll Worman, C., and Yokoyama, K.

(2012). Prokaryote or eukaryote? A unique microorganism from the deep sea. *J. Electron Microsc.,* 61(6):423–431.

Yamashiki, Y. A., Maehara, H., Airapetian, V., Notsu, Y., Sato, T., Notsu, S., Kuroki, R., Murashima, K., Sato, H., Namekata, K., Sasaki, T., Scott, T. B., Bando, H., Nashimoto, S., Takagi, F., Ling, C., Nogami, D., and Shibata, K. (2019). Impact of stellar superflares on planetary habitability. *Astrophys. J.,* 881(2):114.

Yang, J., Boué, G., Fabrycky, D. C., and Abbot, D. S. (2014). Strong dependence of the inner edge of the habitable zone on planetary rotation rate. *Astrophys. J. Lett.,* 787(1):L2.

Yates, J. S., Palmer, P. I., Biller, B., and Cockell, C. S. (2017). Atmospheric habitable zones in Y dwarf atmospheres. *Astrophys. J.,* 836(2):184.

Yi, R., Tran, Q. P., Ali, S., Yoda, I., Adam, Z. R., Cleaves, H. J., and Fahrenbach, A. C. (2020). A continuous reaction network that produces RNA precursors. *Proc. Natl. Acad. Sci. USA,* 117(24):13267–13274.

Yockey, H. P. (2005). *Information Theory, Evolution, and the Origin of Life*. Cambridge University Press, Cambridge, UK.

Youngblood, A., France, K., Loyd, R. O. P., Brown, A., Mason, J. P., Schneider, P. C., Tilley, M. A., Berta-Thompson, Z. K., Buccino, A., Froning, C. S., Hawley, S. L., Linsky, J., Mauas, P. J. D., Redfield, S., Kowalski, A., Miguel, Y., Newton, E. R., Rugheimer, S., Segura, A., Roberge, A., and Vieytes, M. (2017). The MUSCLES Treasury Survey. IV. Scaling relations for ultraviolet, Ca II K, and energetic particle fluxes from M dwarfs. *Astrophys. J.,* 843(1):31.

Yung, Y. L., Chen, P., Nealson, K., Atreya, S., Beckett, P., Blank, J. G., Ehlmann, B., Eiler, J., Etiope, G., Ferry, J. G., Forget, F., Gao, P., Hu, R., Kleinböhl, A., Klusman, R., Lefèvre, F., Miller, C., Mischna, M., Mumma, M., . . . Worden, J. (2018). Methane on Mars and habitability: Challenges and responses. *Astrobiology,* 18(10):1221–1242.

Yurtsever, U. and Wilkinson, S. (2018). Limits and signatures of relativistic spaceflight. *Acta Astronaut.,* 142:37–44.

Zachar, I. and Szathmáry, E. (2017). Breath-giving cooperation: critical review of origin of mitochondria hypotheses. *Biol. Direct,* 12:19.

Zackrisson, E., Calissendorff, P., Asadi, S., and Nyholm, A. (2015). Extragalactic SETI: The Tully-Fisher relation as a probe of Dysonian astroengineering in disk galaxies. *Astrophys. J.,* 810(1):23.

Zackrisson, E., Calissendorff, P., González, J., Benson, A., Johansen, A. and Janson, M. (2016). Terrestrial planets across space and time. *Astrophys. J.,* 833(2):214.

Zackrisson, E., Korn, A. J., Wehrhahn, A., and Reiter, J. (2018). SETI with Gaia: The observational signatures of nearly complete Dyson spheres. *Astrophys. J.,* 862(1):21.

Zaffos, A., Finnegan, S., and Peters, S. E. (2017). Plate tectonic regulation of global marine animal diversity. *Proc. Natl. Acad. Sci. USA,* 114(22):5653–5658.

Zahnle, K., Arndt, N., Cockell, C., Halliday, A., Nisbet, E., Selsis, F., and Sleep, N. H. (2007). Emergence of a habitable planet. *Space Sci. Rev.,* 129(1):35–78.

Zahnle, K., Schaefer, L., and Fegley, B. (2010). Earth's earliest atmospheres. *Cold Spring Harb. Perspect. Biol.,* 2(10):a004895.

Zahnle, K. J., Lupu, R., and Catling, D. C. (2020). Creation and evolution of impact-generated reduced atmospheres of early Earth. *Planet. Sci. J.,* 1(1):11.

Zaia, D. A. M., Zaia, C. T. B. V., and de Santana, H. (2008). Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.,* 38(6):469–488.

Zalasiewicz, J., Williams, M., Waters, C. N., Barnosky, A. D., Palmesino, J., Rönnskog, A.-S., Edgeworth, M., Neal, C., Cearreta, A., Ellis, E. C., Grinevald, J., Haff, P., Ivar do Sul, J. A., Jeandel, C., Leinfelder, R., McNeill, J. R., Odada, E., Oreskes, N., Price, S. J., . . . Wolfe, A. P. (2017). Scale and diversity of the physical technosphere: A geological perspective. *Anthropocene Rev.,* 4(1):9–22.

Zanazzi, J. J. and Triaud, A. H. M. J. (2019). The ability of significant tidal stress to initiate plate tectonics. *Icarus,* 325:55–66.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., and Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature,* 541(7637):353–358.

Zarka, P. (2007). Plasma interactions of exoplanets with their parent star and associated radio emissions. *Planet. Space Sci.,* 55(5):598–617.

Zasova, L. V., Krasnopolskii, V. A., and Moroz, V. I. (1981). Vertical distribution of SO_2 in upper cloud layer of Venus and origin of U.V.-absorption. *Adv. Space Res.,* 1(9):13–16.

Zeder, M. A. (2011). The origins of agriculture in the Near East. *Curr. Anthropol.,* 52(S4):S221–S235.

Zendejas, J., Segura, A., and Raga, A. C. (2010). Atmospheric mass loss by stellar wind from planets around main sequence M stars. *Icarus,* 210(2):539–544.

Zeng, L., Sasselov, D. D., and Jacobsen, S. B. (2016). Mass-radius relation for rocky planets based on PREM. *Astrophys. J.,* 819(2):127.

Zeng, L., Jacobsen, S. B., Sasselov, D. D., Petaev, M. I., Vanderburg, A., Lopez-Morales, M., Perez-Mercader, J., Mattsson, T. R., Li, G., Heising, M. Z., Bonomo, A. S., Damasso, M., Berger, T. A., Cao, H., Levi, A., and Wordsworth, R. D. (2019). Growth model interpretation of planet size distribution. *Proc. Natl. Acad. Sci. USA,* 116(20):9723–9728.

Zhang, B. (2020). A quantitative assessment of communicating extra-terrestrial intelligent civilizations in the galaxy and the case of FRB-like signals. *Front. Phys.,* 15(5):54502.

Zhang, X. (2020). Atmospheric regimes and trends on exoplanets and brown dwarfs. *Res. Astron. Astrophys,* 20(7):99.

Zhu, T. F. and Szostak, J. W. (2009). Coupled growth and division of model protocell membranes. *J. Am. Chem. Soc.,* 131(15):5705–5713.

Zink, J. K. and Hansen, B. M. S. (2019). Accounting for multiplicity in calculating eta Earth. *Mon. Not. R. Astron. Soc.,* 487(1):246–252.

Zink, K. D. and Lieberman, D. E. (2016). Impact of meat and Lower Palaeolithic food processing techniques on chewing in humans. *Nature,* 531(7595):500–503.

Ziolkowski, L. A., Wierzchos, J., Davila, A. F., and Slater, G. F. (2013). Radiocarbon evidence of active endolithic microbial communities in the hyperarid core of the Atacama Desert. *Astrobiology,* 13(7):607–616.

Zozulia, O., Dolan, M. A., and Korendovych, I. V. (2018). Catalytic peptide assemblies. *Chem. Soc. Rev.,* 47(10):3621–3639.

Zubrin, R. M. and Andrews, D. G. (1991). Magnetic sails and interplanetary travel. *J. Spacecraft Rockets,* 28(2):197–203.

Zwicker, D., Seyboldt, R., Weber, C. A., Hyman, A. A., and Jülicher, F. (2017). Growth and division of active droplets provides a model for protocells. *Nat. Phys.,* 13:408–413.

ACKNOWLEDGMENTS

Neither scientists nor fairy-tale characters reside or work in solitude and stillness in cloistered ivory towers, contrary to some banal pop culture stereotypes. Hence, it behooves us to acknowledge and thank the many people whose contributions played an immeasurable role during our journey. Without the boundless support, patience, and affection bestowed by our families, writing this book would not have been possible. Our heartfelt sentiments, especially when it comes to what we owe them, are perfectly encapsulated by the following graceful lines from "Goblin Market" by Christina Rossetti (1865, p. 30):

> To cheer one on the tedious way,
> To fetch one if one goes astray,
> To lift one if one totters down,
> To strengthen whilst one stands.

We consider ourselves privileged to have received stimulating and insightful comments or personal support (which is equally indispensable) from numerous scholars in the course of authoring publications as well as at conferences and in conversations. In this context, we thank Fred Adams, William Bains, Amedeo Balbi, Steve Balbus, John Barrow, Dan Batcheldor, Jim Benford, Shmuel Bialy, Saida Caballero-Nieves, Erica Cartmill, David Catling, Sandeep Choubey, Charles Cockell, Luca Comisso, Charlie Conroy, Dipjyoti Das, Chuanfei Dong, Freeman Dyson, John Forbes, Jacob Foster, Adam Frank, Idan Ginsburg, James Guillochon, Jacob Haqq-Misra, David Harris,

Andreas Hein, Michael Hippke, Nia Imara, David Jewitt, James Kasting, Jonathan Katz, Hridesh Kedia, Giannis Keramidas, Edwin Kite, Ravi Kopparapu, Andy Knoll, David Krakauer, Laura Kreidberg, Sebastiaan Krijt, Russell Kulsrud, Richard Levins, Jeffrey Linsky, Swadesh Mahajan, Zac Manchester, Rocco Mancinelli, Dani Maoz, Chris McKay, Gary Melnick, Phil Morrison, Stephanie Olson, Csaba Palotai, Adina Paytan, Roger Penrose, Eric Perlman, Ramses Ramirez, Sukrit Ranjan, John Raymond, Jeremy Riousset, Michael Russell, Jean Schneider, Andrew Siemion, Amir Siraj, Norm Sleep, Hector Socas-Navarro, Jill Tarter, Francesco Tombesi, Amaury Triaud, Ed Turner, David Wolpert, Richard Wrangham, and Jason Wright.

We are grateful to Zachary Adam, Lewis Alcott, Stuart Armstrong, Giada Arney, Luc Arnold, Dimitra Atri, William Bains, Fernando Ballesteros, Steven Benner, David Catling, Charles Cockell, Edward DeLong, Albert Fahrenbach, Paul Falkowski, Tom Fenchel, Woodward Fischer, Adam Frank, Yuka Fujii, Eric Gaidos, Jacob Haqq-Misra, Christopher Hawkesworth, Ákos Kereszturi, Daniel Koll, Geoffrey Landis, Sanjay Limaye, Maria Lugaro, Timothy Lyons, Ben Pearce, Ramses Ramirez, Paul Rimmer, Anders Sandberg, Edward Schwieterman, Sara Seager, Andrew Siemion, Steinn Sigurdsson, Hector Socas-Navarro, Eva Stüeken, and Joshua Winn for their generous support in connection with the images in the book, and to the various publishers for granting us the appropriate permissions to reuse illustrations, plots, and tables herein.

It is a pleasure to thank Nina Zonnevylle, Mark Palmer, Uma Mirani, Leslie Smith, Kathy Alger, and Heidi Kendig for ensuring the smooth functioning of our primary workplaces—the Institute for Theory and Computation at Harvard University and the Department of Aerospace, Physics and Space Science at the Florida Institute of Technology—and for the enjoyable conversations with regard to sundry matters. We wish to express our gratitude to the reviewers, whose perspicacious and constructive comments greatly aided in enhancing the quality of our book. We are much obliged to our editor, Janice Audet of Harvard University Press, for her meticulous work and inputs. On the same note, we are grateful to Emeralde Jensen-Roberts, Stephanie Vyce, and other employees of Harvard University Press for their diligent support and assistance over the span of publishing this tome. Continuing in this vein, we appreciate the hard work, timeliness, and professionalism displayed by the production editor Sherry Gerstein, the copyeditor Elizabeth Asborno, and the associated personnel at Westchester

Publishing Services during the production process. If we have excluded any crucial individuals unwittingly, we apologize for such an oversight.

In addition to the aforementioned people, countless others from all walks of life have rendered invisible, yet indisputably invaluable, aid over the years. And last, but by no means the least, the Earth, the Sun, and the Universe as a whole have proven amenable to the evolution of our complex and intricate biosphere, and the advent of *Homo sapiens*. To these entities along with everyone else highlighted earlier, we have nothing but the deepest gratitude:

> I can no other answer make but thanks,
> And thanks; and ever thanks.
>
> —William Shakespeare, *Twelfth Night, or What You Will*