

ASTROBIOLOGY



*Debra Fischer, Allyson Sheffield, Joshua Tan, and Lily
Ling Zhao*

Astrobiology

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 10/16/2026

TABLE OF CONTENTS

Licensing

1: Understanding the Cosmos

- 1.1: Sizing Up the Cosmos
- 1.2: Energy and Light
- 1.3: Spectroscopy and Stars
- 1.4: Laws of Motion and Gravity

2: The Chemistry of Life

- 2.1: Introduction to Atoms and Elements
- 2.2: Basic Chemistry of Life
- 2.3: Key Biochemicals- Amino Acids, Nucleic Acids, and the Genetic Code
- 2.4: Alternative Biochemistries- Speculations about Strange Possibilities

3: Origins of Life

- 3.1: Definition of Life
- 3.2: From Biochemistry to Biology
- 3.3: Emergence of Life on Earth
- 3.4: Development of Life on Earth
- 3.5: Extremophiles

4: The Search for Life in Our Solar System

- 4.1: Solar System Formation
- 4.2: Earth Characteristics
- 4.3: Searching for Life In Our Solar System
- 4.4: The Search for Life on Mars
- 4.5: The Jovian Moons

5: The Search for Life Beyond Our Solar System

- 5.1: The Habitable Zone
- 5.2: Exoplanets- Early History and Direct Imaging and Doppler Detection Methods
- 5.3: Transit, Astrometry and Gravitational Microlensing Techniques
- 5.4: Exoplanets- Statistics and Modern Discoveries

6: Searches for Technological and Intelligent Life in the Universe

- 6.1: The Challenges of Space Flight
- 6.2: The Search for Extraterrestrial Intelligence (SETI)
- 6.3: Biosignatures and Technosignatures
- 6.4: The Drake Equation and the Fermi Paradox
- 6.5: Speculations and Skepticism

Index

Glossary

Detailed Licensing

This page titled [Astrobiology \(Fischer, Sheffield, Tan, and Zhao\)](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by Debra Fischer, Allyson Sheffield, Joshua Tan, and Lily Ling Zhao via [source content](#) that was edited to the style and standards of the LibreTexts platform.

Licensing

A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.

This page titled [Licensing](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

1: UNDERSTANDING THE COSMOS

1.1: Sizing Up the Cosmos

1.2: Energy and Light

1.3: Spectroscopy and Stars

1.4: Laws of Motion and Gravity

1.1: Sizing Up the Cosmos

To approach the subject of astrobiology – specifically, of exploring possibilities that other life exists beyond the Earth – it is worthwhile to ponder the size and age of the cosmos. In this grand scheme, when and where did life on Earth first appear? The moment when the first life appeared is not precisely known, but several lines of evidence provide compelling clues for the timing of the rise of life on Earth.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain the idea of the “cosmic calendar” and the timing of key events for life
- Express small and large numbers using powers of ten
- Describe the range of sizes for objects in the universe
- Explain light units such as light years and light seconds
- Describe how parallax can be used to estimate the distance to an object

Our Place in Time: The Cosmic Calendar

The universe emerged at the moment of the Big Bang, 13.8 billion years ago. This is our starting point for time. Key events since this moment, including the formation of our galaxy and our solar system, can be identified in a memorable way by compressing all of time into one calendar year. This idea for the “cosmic calendar” was first articulated by astronomer Carl Sagan. On this cosmic calendar (Figure 1), each month represents a little over a billion years. Each day covers a span of around 40 million years. The very first stars and galaxies form in late January, and our home galaxy, the Milky Way, forms in early March. Our solar system comes on the scene in early September, and the first life on Earth probably formed by mid- to late-September.

Life starts as very simple cells and gradually evolves into more, with the first multicellular life appearing in November, and the first plants on December 20th. Primates, including humans, appear on December 31. In fact, modern humans show up on December 31, at 11:52 PM, which is on the order of the real-time equivalent of hundreds of thousands of years ago. In the bottom panel of Figure 1, which zooms in on the last 60 seconds on December 31, we see that all of modern science, including the development of both the telescope and microscope in the 17th century, happen in the last second at 11:59 PM.



Figure 1 – The Cosmic Calendar. If we scale the history of our universe to the length of one Earth year, we can get a sense that living beings similar to us have been in existence for a very small fraction of the total amount of time. **When did the dinosaurs go extinct on the Cosmic Calendar?** Credit: [Wikipedia](#), CC BY-SA

There are a few ways to think about this time frame. On one hand, technologically advanced humans on Earth are newcomers. Humans have been sending out radio waves for less than one cosmic second. However, considering that the first life to appear on Earth took only a few cosmic weeks to form, perhaps the initial formation of life on other worlds will be a relatively speedy process. When considering just how many planets there are in the cosmos, it is tempting to think that the odds of life existing on at least one other planet are good.

Our Place in Space

Space also emerged at the moment of the Big Bang, and it has been expanding ever since. Contemplating the size of the cosmos can seem mind-bending but we can ground ourselves by starting with some size comparisons of more familiar objects. When comparing the sizes of two objects, a good first step is to be clear as to what is being compared, as the word size can mean different things — length and volume, for example, are do not measure the same thing. Here, let's compare the length of some different objects, where the length is the distance from one end to the other end. Of course not all objects are symmetric so you may need to be more specific about how you are defining length, depending on what is being compared. For example, bacteria come in different shapes, such as rods, spheres, and spirals. In specifying a bacterium's size, you would need to be clear about the start and end points of the measurement. In astronomy, galaxies also come in different shapes – ellipsoids, spirals, and irregular shapes – so it is important to know which dimension (say left to right or top to bottom) the length is referring to.



Figure 2 – The micro-world of organisms smaller than the human eye can see requires microscopy to detect. **Can you estimate how many times longer the bacterial cell is compared to the penicillin (the smaller purple spheres)?** Credit: [wikimedia](#), CC BY-SA

Units of length are important to consider. Lengths and distances are one-dimensional measures and are expressed in units such as meters or miles. Of course, the length of any object can be expressed in *any* units: what's important is that the objects being compared have the same units. Let's compare the size of a grapefruit and a bacterium. The length of a typical grapefruit is around 10 inches, which is the same as 25.4 centimeters (cm), 0.254 meters (m), 0.000254 kilometers (km) or 0.000158 miles. These are all equivalent measures, and it is a matter of convenience when choosing units. When directly comparing two measurements, all that matters is that the units chosen for the grapefruit are the same as those for a bacterium. We know that bacteria are very small, and a typical size is reported in units of micrometers (μm), where the prefix micro indicates 0.000001 meters. So, we can compare a grapefruit and a bacterium using meters (m): a grapefruit is 0.254 m and a bacterium is 0.000001 m. Clearly, the grapefruit is larger than the bacterium, and we can find exactly how much larger by dividing the two sizes: $(0.254 \text{ m}) / (0.000001 \text{ m}) = 254,000$. Dividing the two sizes (or taking the ratio) shows that the grapefruit is 254,000 times larger than the bacterium. We could still use a way to simplify our notation, as it's easy to accidentally leave off one of the zeroes in the size of the bacterium. That's where "powers of ten" notation becomes useful.

Powers of Ten

Scaling to the largest sizes we will consider, a [galaxy](#) is a vast collection of stars that are held together by gravity. The length of a typical spiral galaxy, like the Milky Way, is about 1,000,000,000,000,000,000,000 meters (m), or a billion trillion m. This is a

huge number, with 21 zeros after the one (this length is called a sextillion meters). The number may seem more intuitive when expressed as one billion trillion, as billion and trillion are familiar values that show up in our daily life (for example, billionaires are often in the news, and although no trillionaires yet exist, we intuitively know how much money they would have). When writing these numbers the prefixes M and G are sometimes used, where M stands for a million or the prefix “Mega” and G for a billion or the prefix “Giga”.

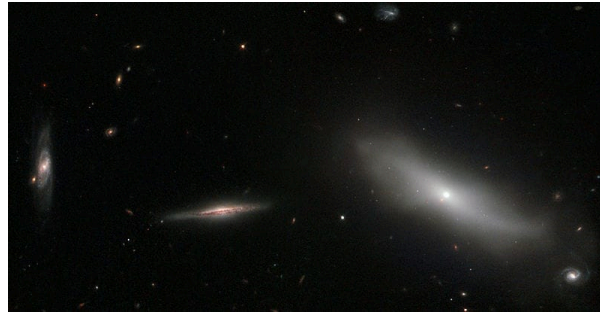


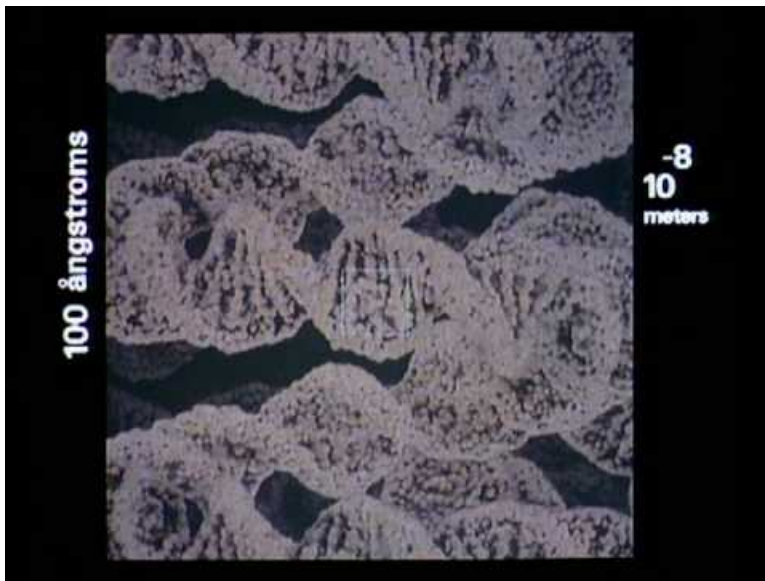
Figure 3 – Galaxies are immense in scale, but because they are so far away, their faintness requires telescopes to be able to detect them.
Credit: [ESA/Hubble & NASA](#), public domain

One additional way to express these number is by using powers of ten. A million expressed in powers of ten is 10^6 , since there are 6 zeros. Similarly, the Milky Way's size can be written as 10^{21} m. There is also a prefix for 10^{21} , called zetta and abbreviated by a Z. The Milky Way's size can be expressed as 10^{21} m, 1 Zm or 1 sextillion m. [This link](#) to the National Institute of Standards and Technology lists the common prefixes that we use. In this table, we see that a yoctosecond is a very small measure of time – it is a mere septillionth of a second or 10^{-24} s. We will make regular use of powers of ten notation in this course, as we will be dealing with immensely large and small numbers.

In the context of comparing such a wide range of phenomena as this subject warrants, a good first step is to estimate the order of magnitude of the scale of an object. The easiest way to do this is to write the measurement in [scientific notation](#) and consider only the exponent. Notice that we need to identify the units of measurement as well as the order of magnitude to avoid ambiguity.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

To get a better feel for very large and very small numbers in nature, the following film, created by Charles and Ray Eames in the 1970s, explores the very largest and very smallest numbers using powers of ten. Note that since the creation of this film, scientific advances have aided in the more complete understanding of the largest and smallest scales so that we now have more exacting probes of the largest scales and the smallest scales shown in the film. Our maps of galaxy clusters and investigations to the observable limits of our universe have improved to the extent that we can accurately identify the distribution of millions of galaxies while our understanding of the standard model of particle physics has been confirmed down to the centers of atomic nuclei.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=40>

Copyright © Eames Office, LLC. All rights reserved.

The smallest measurable sizes bring us inside the nucleus of an atom, where subatomic particles called protons, with a size of just 10^{-15} m, are found. On the largest scale, the size of the observable universe is estimated to be about 10^{26} m. These are such vastly different sizes it is hard to even imagine comparing them, but it can be done very simply. The question of “how many times larger” one object is compared to another object can be found by dividing the two numbers: $10^{26} \text{ m} / 10^{-15} \text{ m} = 10^{41}$ (for a review on working with powers of ten and scientific notation, please see [Appendix B](#)). The answer tells us that the observable universe is 10^{41} times larger, or “41 powers of ten,” larger than a proton. This can also be stated by saying that the universe is 41 orders of magnitude larger than a proton.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

This type of analysis is powerful. Compared to the entire size of an atom, the nucleus is about 10,000 times smaller (10^4 , or 4 orders of magnitude). This tells us that an atom – the fundamental building block of matter – is mostly empty space. The universe, it turns out, is also filled with huge voids of empty space.

Concept Check: Exponential sizes

Storage devices come in different sizes, and a small flash drive might have 64 MB of space and a large drive could have 500 GB. The small flash drive has 64 million bytes of space, where a million is 1,000,000 – one with six zeros. The “M” in MB stands for the prefix Mega, meaning million, and “B” is the abbreviation for bytes. Note that G is the prefix Giga, meaning billion.

To compare the size of a 512 MB flash drive to a 64 MB flash drive, this can be answered in one step by dividing the two sizes: $512 \text{ MB} / 64 \text{ MB} = 8$. The 512 MB is eight times larger than the 64 MB drive. To state it another way, you would need eight 64 MB drives to equal one 512 MB drive. Notice that the answer, eight, has no units (or, we say is “dimensionless”). This makes sense – since both sizes are in MB, we compared “apples to apples”.

If the question asked how many times larger a 512 GB drive is compared to a 64 MB drive, you would need to make the units agree before dividing the numbers. Do this!

Show Solution

There are a few different ways to approach this: you could convert from GB to MB, from MB to GB, or convert both units into B. Let's convert 512 GB into MB, so that both quantities will have the same units:

$$512 \text{ GB} \times (10^3 \text{ MB}/1 \text{ GB}) = 512 \times 10^3 \text{ MB}$$

Now, we can divide the two quantities since they are both in MB: $(512,000 \text{ MB}/64 \text{ MB}) = 8000$

It takes eight thousand 64 MB flash drives to equal one 512 GB drive.

If the conversion directly from GB to MB was not clear, we can do this in two steps:

$$512 \text{ GB} \times (10^9 \text{ B}/1 \text{ GB}) \times (1 \text{ MB}/10^6 \text{ B}) = 512 \times 10^9 \times 10^{-6} \text{ B} = 512 \times 10^3 \text{ MB} = 512,000 \text{ MB}$$

Stars

The power of a star can be measured in Watts (W), just like light bulbs. The Sun's power output is $3.8 \times 10^{26} \text{ W}$ and the power of the star Arcturus is $6.5 \times 10^{28} \text{ W}$.

(a) How many times more powerful is Arcturus than the Sun? (b) How many times less powerful is the Sun than Arcturus?

Show Answer

(a) We can simply divide the power of Arcturus by the power of the Sun to find this:

$$(6.5 \times 10^{28} \text{ W}) / (3.8 \times 10^{26} \text{ W}) = 171$$

(b) 0.0058 or (1/171)

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Density

Another important measure in the context of life and other worlds is density. The density of a planet, for example, will be important in assessing whether that planet has a rocky surface, like Earth, or an entirely gassy composition, like Jupiter.

The density of an object tells us how much stuff fits into its available space. Or, to be more specific, how much mass fits into its volume. You cannot tell the density of an object just by its mass or just by its volume – you need to consider both quantities together. For example, two planets could have the exact same volume and have the same size and shape in a picture, but their densities depend on what kind of matter is inside. A tennis ball and a billiard ball have roughly the same volume, but the tennis ball is filled with air and is much less dense than a billiard ball of the same size. By definition, the density (ρ) of an object is found by dividing its mass by its volume and is often reported in units of g/cm^3 (since mass can be measured in grams and volume is three-dimensional and is measured in cubic units, such as cm^3):

$$\rho = \frac{M}{V} \quad (1.1.1)$$

Table 1 lists some common densities:

Table 1: Density of some common materials (left) and celestial objects (right). Note that the density of the Universe is many orders of magnitude smaller than the density of air, confirming that the Universe is mostly filled with empty space.

Substance	Density	Astronomical Object	Density
hydrogen	$0.00009 \text{ g}/\text{cm}^3$	Earth	$5.5 \text{ g}/\text{cm}^3$
air	$0.001 \text{ g}/\text{cm}^3 = 1 \text{ kg}/\text{m}^3$	Sun	$1.4 \text{ g}/\text{cm}^3$

water	1 g/cm ³	Saturn	0.7 g/cm ³
iron	8 g/cm ³	white dwarf	700,000 g/cm ³
lead	11 g/cm ³	neutron star	7×10^{14} g/cm ³
gold	19 g/cm ³	Universe	9×10^{-30} g/cm ³

Concept Check: Density

- (a) Looking at Table 1, how many times more dense is a neutron star compared to the planet Saturn?
- (b) Of the elements shown in Table 1, gold has the highest density. Is gold the heaviest element? Look this up!
- (c) Which object has a higher density (or do they have the same density): a 1 milligram piece of gold or a 50 milligram piece of gold?

Show Answer

- (a) 10^{15} or 1 quadrillion

To find how many more times dense a neutron star is compared to Saturn, take the ratio of the two densities:

$$(\text{density of neutron star})/(\text{density of Saturn}) = (7 \times 10^{14} \text{ g/cm}^3)/(0.7 \text{ g/cm}^3) = 10^{15}$$

- (b) osmium is the densest element

- (c) They have the same density! Density is an intrinsic property of an element and doesn't change with the size of the piece.

Light Units and Speed

To explore the chemistry and possible biology of other worlds, ideally we would send humans there to run experiments. Humans have been to the Moon to collect rocks, and there are currently several rovers that are controlled from Earth analyzing rocks and soil on the surface of Mars. There are plans to send humans to Mars in the next decade or two. Why not send humans sooner? There turn out to be many challenges to physically traveling beyond the Earth, but the distances involved are foremost.

We are familiar with long distances on Earth. The distance from NYC to Sydney, Australia is about 10,000 miles (16,000 km) and a typical jet flies at a speed of 575 mph, so this flight takes over 17 hours. How about the distance from Earth to the Moon or Mars? The Moon is around 240,000 miles (384,000 km) away and Mars is over 200 million miles from Earth. In 2020, the *Perseverance* spacecraft travelled at 24,600 mph to Mars and still took about 7 months to get there. The amount of time it takes to travel somewhere is the distance divided by the speed you are traveling, or:

$$\text{time} = (\text{distance})/(\text{speed})$$

If you are traveling to a place 30 miles away at a steady speed of 60 mph, then it takes you $(30 \text{ miles})/(60 \text{ mph}) = 0.5$ hour to get there.

Concept Check: Distances, speed and time

If it takes 18 days for a rover moving at a steady speed to travel a distance of 6 km, how fast was the rover moving?

Show Answer

$$\text{speed} = (\text{distance})/(\text{time})$$

$$\text{speed} = (6 \text{ km})/(18 \text{ hours}) = 1/3 \text{ km/hr} = 0.33 \text{ km/hr}$$

In fact, this is the maximum speed that the Perseverance rover can reach.

Perhaps you can just travel at faster speeds to get someplace far, like Mars, more quickly? While this idea works well for speeds on Earth (and hence we have supersonic jets and bullet trains), there is a limit to how fast anything can travel. This cosmic speed limit is called the speed of light, and nothing can travel faster. In fact, only light, which has no mass, can travel at the speed of light; any object with mass can never reach this speed due to the effects of Einstein's theory of special relativity. So, the fastest possible way to reach a distant world is to send light. This may not seem very helpful at first, but light can contain information – think about radio or TV signals – so this is a potential way to send communications to distant worlds.

How long would light take to reach Mars? The speed of light, which is denoted by the symbol c , can be expressed in any units, such as km/hr, mph, or feet/s. Two common choices are m/s and km/s:

$$c = 300,000 \text{ km/s} = 3 \times 10^5 \text{ km/s} = 3 \times 10^8 \text{ m/s}$$

(For comparison, a variety of speeds that we may encounter in astrobiology are listed in Table 2.) The distance to Mars when it is closest to the Earth is 55 million km. Using the simple relationship for time, distance, and speed, we arrive at an answer of 183 seconds. That's fast – only about 3 minutes! And definitely faster than the 7 months the *Perseverance* spacecraft took. This way of thinking about distances, times, and the speed of light is so useful, in fact, that astronomers often use light units to express distances. The Sun is 93 million miles away from Earth. If we do the same calculation for how long light from the Sun takes to reach us on Earth, it is 8 minutes. That's about three times as long as light takes to get to and from Mars. We can now express the distances d to the Sun and Mars, in light units: $d_{\text{Sun}} = 8$ light minutes and $d_{\text{Mars}} = 3$ light minutes

Table 2: Speeds that we may encounter in various contexts in the study of astrobiology.

Scenario	Speed	Scenario	Speed
Drift speed of the continents	$3 \times 10^{-9} \text{ m/s}$	Orbital speed of the Earth around the Sun	30 km/s
Growth speed of human hair	$5 \times 10^{-9} \text{ m/s}$	Escape speed from the Solar System (at the distance of Earth)	40 km/s
Automobile	100 km/hour = 28 m/s	Fastest human-made object	110 km/s
Rotational speed of the Earth at the Equator	1700 km/hour = 480 m/s	Orbital speed of the Sun around the Galaxy	220 km/s
Orbital speed in low-Earth orbit	8 km/s	Escape speed from the Galaxy (at the location of the Sun)	310 km/s
Escape speed at Earth's surface	11 km/s	Speed of light = c	$3 \times 10^5 \text{ km/s}$

Worked Example: Light Units

Show that 1 light year is equal to 9.46 trillion km.

Recall that a light year is the distance that light travels in 1 year. We can use the relationship between time, distance and speed, along with unit conversions, to show this.

$$\text{distance} = (\text{speed}) \times (\text{time})$$

Here, we know the speed—the speed of light—and can express it in any units...let's use km/s, since the answer is in units of km. In the equation below, we use $3 \times 10^5 \text{ km/s}$ for the speed. For the time, we use 1 year and convert this to seconds.

$$\text{distance} = (3 \times 10^5 \text{ km/s}) \times (1 \text{ year})$$

$$\text{distance} = (3 \times 10^5 \text{ km/s}) \times (1 \text{ year}) \times (365 \text{ days/1 yr}) \times (24 \text{ hours/1 day}) \times (60 \text{ min/1 hr}) \times (60 \text{ sec/1 min}) = 9.46 \times 10^{12} \text{ km}$$

Recall that a light year is the distance that light travels in 1 year. We can use the relationship between time, distance and speed, along with unit conversions, to show this.

$$\text{distance} = (\text{speed}) \times (\text{time})$$

Here, we know the speed—the speed of light—and can express it in any units...let's use km/s, since the answer is in units of km. In the equation below, we use 3×10^5 km/s for the speed. For the time, we use 1 year and convert this to seconds.

$$\text{distance} = (3 \times 10^5 \text{ km/s}) \times (1 \text{ year})$$

$$\text{distance} = (3 \times 10^5 \text{ km/s}) \times (1 \text{ year}) \times (365 \text{ days/1 yr}) \times (24 \text{ hours/1 day}) \times (60 \text{ min/1 hr}) \times (60 \text{ sec/1 min}) = 9.46 \times 10^{12} \text{ km}$$

This is equal to 9.46 trillion km.

How long does it take light from the Sun to reach Saturn, in minutes?

Start with the average distance from the Sun to Saturn in km, which is 1.4 billion km.

Show Answer

Light from the Sun is traveling at the speed of light, or 3×10^5 km/s.

$$\text{time} = (\text{distance})/(\text{speed}) = (1.4 \times 10^9 \text{ km})/(3 \times 10^5 \text{ km/s}) = 4,667 \text{ seconds}$$

$$(4667 \text{ sec})(1 \text{ min}/60 \text{ sec}) = 78 \text{ minutes}$$

We can thus say that Saturn is 78 light minutes away from the Sun.

How about the closest star to the Sun? That's a red star named Proxima Centauri, and it is staggering 40,208,000,000,000 km (40 trillion km or 4×10^{13} km) away. In light units, however, this distance is 4 light years. It takes light 4 years to reach Proxima Centauri. The distances in light units to some nearby astronomical objects are shown in Figure 4. Light units are very convenient for expressing distances to stars that have intriguing planets orbiting them, as this tells us how long a radio signal would take to arrive at the star system. Also, light units remind us that we have a way of looking into the past! Notice in Figure 4 that the Andromeda galaxy is 2.5 million light years away. This means that light from Andromeda takes 2.5 million years to reach us, and that consequently we are seeing Andromeda as it looked 2.5 million years ago. This gives astronomers a way to see how the oldest galaxies in the universe, that are more than 13 billion light years away, looked when they were very young!

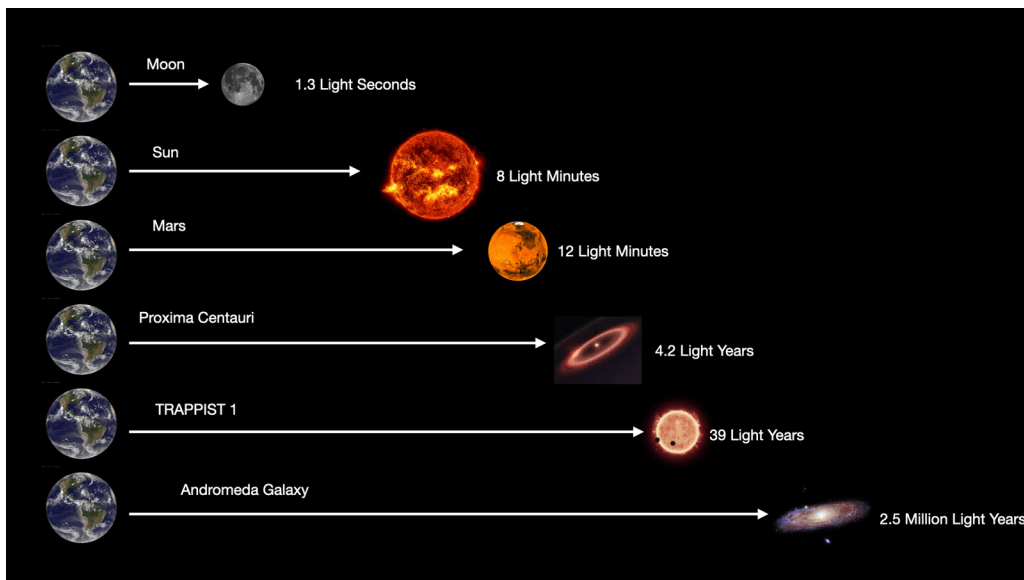


Figure 4 – The distances between the Earth and various locations in the Universe given in light units. **Can you think of an object in space that is even closer to the Earth than 1 light second?** Credit: Josh Tan

Concept Check: Distances

In Figure 4, the distance of 12 light minutes from Earth to Mars is actually the average, as the distance from Earth to Mars varies from 3 light minutes to 22 light minutes (at the furthest distance, it is on the opposite side of the Sun compared to Earth). When Mars and the Earth are at their closest distance, how long does it take light from the Sun to reach Mars?

It may help to make a sketch of the alignment of the Sun, Mars and Earth when Mars and Earth are closest to each other.

Show Sketch and Answer



Figure 5. The alignment of the Sun, Earth and Mars at closest approach, called opposition, (left) and farthest distance, called conjunction (right). Credit: NASA, public domain

The left panel shows the alignment of the Sun, Earth, and Mars when Mars is closest to the Earth. The time for sunlight to reach the Earth is 8 light minutes and the time for light to travel from Earth to Mars is 3 light minutes at closest approach. Thus, the total time is $8 \text{ light min} + 3 \text{ light min} = 11 \text{ light min}$

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Other units of distance in astronomy

If you look up the distance to an astronomical object, the units will vary. For example, the distance to the star Vega might be

reported in light years or parsecs — it is 25 light years away which is the same as 7.8 parsecs. The Oort cloud surrounding our solar system is 50,000 AU away. Let's introduce these units as well.

Astronomical Units

When reporting distances within solar systems, astronomical units (AU) are typically used. One AU is defined to be the average distance from the Earth to the Sun. Mercury and Venus are closer to the Sun than the Earth and thus have distances in AU that are less than 1 AU. Saturn is at about 10 AU from the Sun and Neptune is at 30 AU.

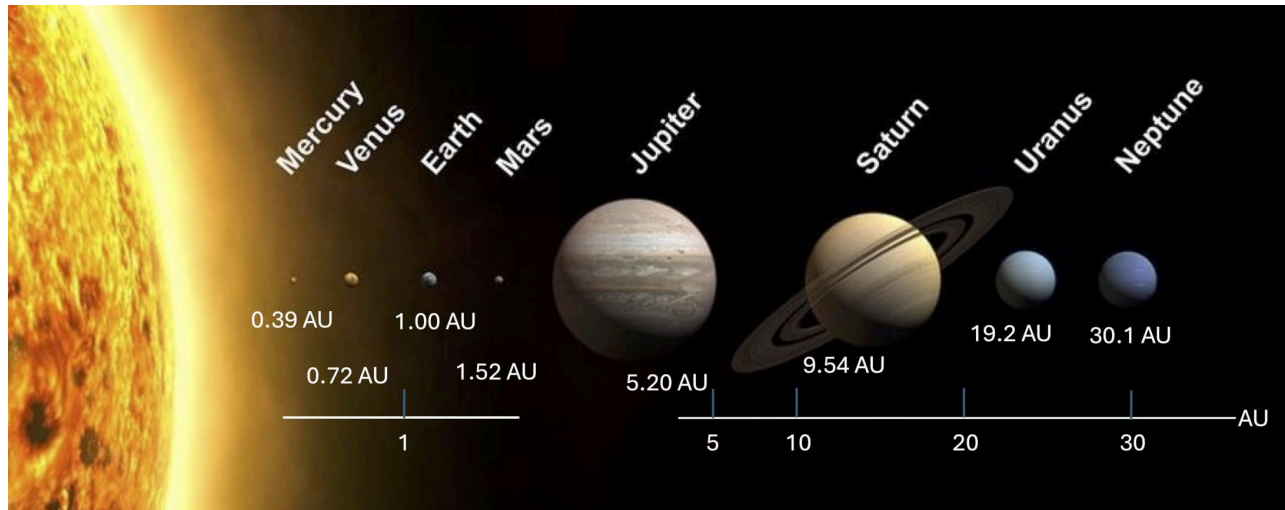


Figure 6 – The planets in our solar system arranged by their distances from the Sun. Units of AU (astronomical units) are used for the planetary distances. **The Asteroid Belt in our solar system is found at distances between 2.2 and 3.2 AU. Which two planets is the Asteroid Planet located in between?** Credit: Wikipedia (modified), CC BY-SA

When you read about the distance of an exoplanet from its host star, these are also reported in AU and have the exact same meaning as in our solar system. The exoplanet Proxima b orbits the red star Proxima Centauri, and Proxima b is an average distance of 0.0485 AU away from Proxima Centauri. This means Proxima b is very close to its star: it is nearly 2 orders of magnitude closer to Proxima Centauri than the Earth is from the Sun.

Parsecs

Parsecs (abbreviated pc) and light years have similar values:

$$1 \text{ pc} = 3.26 \text{ light years}$$

It is usually a matter of context in deciding which units to use although it is usually more intuitive for students to use light units.

The word parsec is a portmanteau of the words “parallax” and “arcsecond” and is defined as “the distance at which we have a *parallax* of one second.” Parallax is an angular measure, meaning it has units of degrees or radians (or arcminutes and arcseconds). The apparent shift in the direction of an object as a result of the motion of the observer is called parallax. We call the shift in the apparent direction of a star due to Earth's orbital motion *stellar parallax*. The Greeks made dedicated efforts to observe stellar parallax, even enlisting the aid of Greek soldiers with the clearest vision, but to no avail. The brighter (and presumably nearer) stars just did not seem to shift as the Greeks observed them in the spring and then again in the fall (when Earth is on the opposite side of the Sun). In reality, stellar parallax angles were too small to measure with scientific instruments until the 19th century.

Concept Check: Angular Units

The basis for angular units is that there are 360° or 2π radians in a full circle. Arcseconds and arcminutes are simply divisions of a degree into smaller units, where 1 degree contains 60 arcminutes (abbreviated as 60'') and 1 arcminute contains 60 arcseconds (abbreviated as 60''').

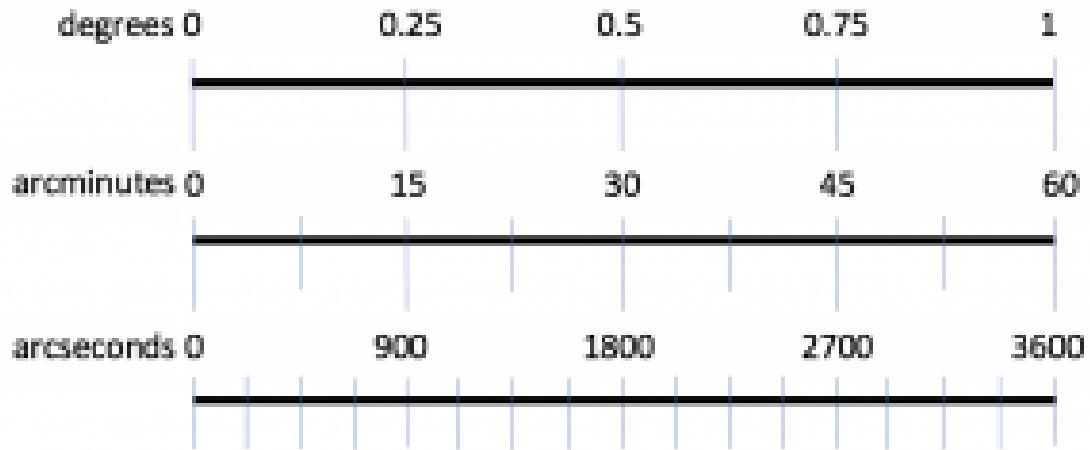


Figure 7 – Angular units: degrees, arcminutes and arcseconds. One degree is divided into 60 arcminutes, similar to how one foot is divided into 12 inches. One degree is further into 3600 arcseconds. From this scale, we see that 0.5 degrees, 30 arcminutes, and 1800 arcseconds are all equivalent measures. **Can you guess why the word “arc” was used for these angular measures?** Credit: Debra Fischer

The best way to visualize parallax angles is to hold up something like a pencil at a distance of just a few inches from your nose. Now, hold up another pencil at the greatest distance that you can reach (as demonstrated below). If you wink your eyes (left then right), you’ll see that the pencil that is closest to your face appears to have a much larger shift (right to left) than the distant pencil.



Figure 8 – Demonstrating parallax shift. Hold an object close to your face and a similar object farther away. **When you wink your eyes from one side to the other, which of these two objects moves the most with respect to very distant objects?** Credit: Debra Fischer

We use the same principle to measure distances to stars. Stars that are close to us will have larger parallaxes than more distant stars. Winking our eyes won’t work for this measurement because the baseline (the distance between our eyes) is too small. Instead, we use the baseline of the Earth’s orbit to see apparent position of the foreground stars jump relative to the background stars.

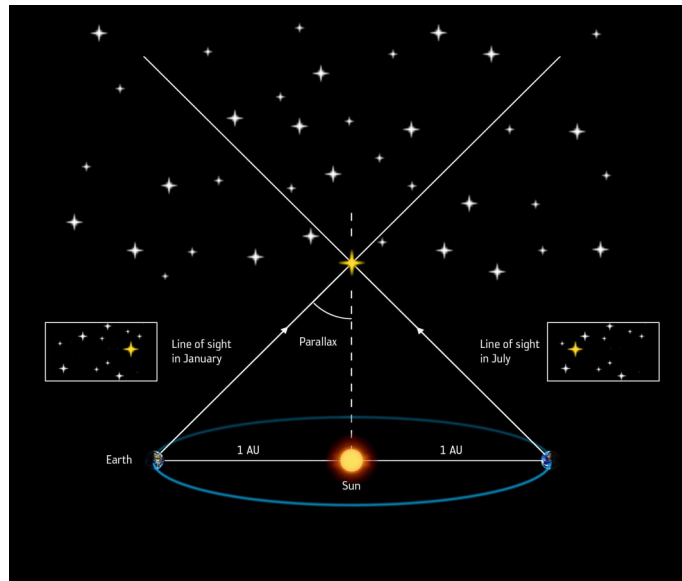


Figure 9 – Stellar parallax. If we observe a star field in January and again in July, stars that are closer will appear to move with respect to background stars. This parallax effect allows us to determine the distance to nearby stars. **Is the parallax angle larger or smaller for closer stars?** Note that the baseline is 1 AU, the average distance from the Earth to the Sun, whereas the baseline in Fig. 7 is the distance between the human's eyes. Credit: [ESA](#), CC BY-NC-SA

The parallax angle is abbreviated as p . It can be shown using simple trigonometry that the distance to a star is inversely proportional to its parallax angle:

$$d = \frac{1}{p} \quad (1.1.2)$$

where p is reported in arcseconds and d is given in parsecs.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Key Concepts and Summary

The scale of space and time in the universe is so enormous that we need strategies to organize a basic understanding of where and when. The cosmic calendar re-scales the 13.8 billion year age of the universe into a “cosmic year”, effectively capturing the timing for the rise of life and the very recent appearance of modern humans. The exponential spatial scale of the universe is captured as powers of ten, to compare the smallest and largest sizes and distances. A common unit for distances is light travel time: light could travel the circumference of the Earth seven times in just one second. It takes light 8 minutes to travel from the Sun to the Earth; 4.3 years to travel from Earth to the nearest star system, alpha Centauri, and 2.5 million years to travel to the closest spiral galaxy, Andromeda. Another important distance scale in astronomy is the distance from the Sun to the Earth: 1 Astronomical Unit or 1 AU. As we look out into the universe, determining the distances to other stars is accomplished with high school level trigonometry: measuring the shift of foreground stars relative to more distant stars.

Review Questions

Summary Questions

1. How does the “cosmic calendar” help in understanding when key events happened for the development of life on Earth?
2. How do you compare the size (or any property) of two objects using powers of ten?
3. How does the smallest size we can observe compare with the largest size? Give an example of two objects that vary by 8 orders of magnitude. How about 17 orders of magnitude?
4. What are light units? Why is it convenient for astronomers to use light units when expressing distances?

5. How many light years are there in 1 parsec? Describe the unit parsec and why it is named the way it is.
6. How can parallax be used to estimate the distance to an object?

Exercises

1. Using the speed in Table 2, how long will your hair grow in one month?
2. The Voyager 1 spacecraft is traveling at a speed of 38,000 miles per hour (17 km/s). Express this speed in light years per second (ly/s). How long would it take Voyager 1 to reach the closest star to the Sun, Proxima Centauri, which is 4.3 light years away?
3. There are some additional exercises using powers of ten and scientific notation [here](#) (with solutions).

1.2: Energy and Light

Life needs a source of energy to thrive. This energy can come in different forms, from sunlight to energy released when chemicals react. Light is a type of pure energy, and coded into the light and other kinds of radiation that reach us from objects in the universe is a wide range of information about what those objects are like and how they work. If we can decipher this code and read the messages it contains, we can learn an enormous amount about the cosmos without ever having to leave Earth or its immediate environment.

Learning Objectives

By the end of this chapter, you will be able to:

- Understand that there are different types of energy and that energy is conserved
- Explain the wavelength, frequency and energy of a wave or particle of light
- Discuss the different parts of the electromagnetic spectrum
- Explain how and why the light emitted by an object depends on its temperature
- Explain the difference between the luminosity and brightness of an object
- Explain how the total energy emitted differs for hot and cold objects

Energy

Different types of energy are all around us — the electrical energy that powers our technology, the chemical energy that is stored in the food we eat, and the heat energy we feel when walking in sunlight. These different types of energy can be changed, or converted, from one type to another. For example, the chemical energy in food is converted to energy that allows us to think and move around. We will see other examples of energy transformations in astrobiology, such as the gravitational energy that is converted into heat when a giant cloud of gas and dust collapses and eventually forms stars.

The amount of the energy in the universe is always the same and it is continuously being changed from one form into another. This is the essence of the law of conservation of energy. Other quantities in nature, such as mass and momentum, are also conserved.

Though energy cannot be created or destroyed, it can be transformed into different forms. The two main forms of energy are kinetic energy and potential energy. Kinetic energy is the energy of motion. When an object is moving more quickly, it has more energy. A fast-moving car has more energy than a slow moving car. Thermal energy is actually a form of kinetic energy since higher temperature is really a measure of the average speed of atoms and molecules — the higher the temperature, the greater the kinetic energy of the atoms and molecules.

Potential energy is the energy that is associated with different positions in space. When you throw a ball high into the air, it has more gravitational potential energy than when it is held in your hand. When an object moves from an area of high potential energy to low potential energy, that energy will be transformed into another form. For example, when an apple falls to the Earth, it will speed up gaining kinetic energy. But potential energy does not need to be related just to gravity. For example, the energy stored in chemicals is in the form of chemical potential energy which can be determined based on the position of different atoms bonded to each other by electric forces of attraction.

A more recently discovered form of energy is the energy associated with mass itself. Einstein's famous equation $E = mc^2$ tells us, that mass-energy equivalence is a fundamental feature of the universe. Any object that has mass therefore also contains energy — this type of energy is called rest-mass energy. Inside the core of stars, nuclear fusion converts some of the rest-mass of two protons into energy. This process occurs an astounding 10^{38} times every second and is the source of the energy from the Sun that we receive on Earth.

Light

A great deal of the energy that life on Earth uses comes in one way or another from the Sun's light, although geothermal energy from inside the Earth can also be harnessed. Light from the Sun, or any star, is an example of electromagnetic radiation. Here, the

words light and radiation are synonymous, and both are a form of energy. Radiation can sometimes be thought of as dangerous and some types are (for example, gamma rays and X rays), but radiation is a neutral term and is synonymous with energy.

One of the more bizarre aspects of light is the way that it acts, specifically the fact that it can behave like a wave and a particle. In the 17th century, the nature of light was debated and both Isaac Newton and Christiaan Huygens tried to explain it. Newton experimented with lenses and prisms and believed that light was a stream of particles, which he called corpuscles. Huygens, on the other hand, believed that light was a wave that traveled outward in all directions from a source, much like the ripples on a pond that expand outward through the water after a stone is dropped into it. Building upon pioneering work by James Clerk Maxwell and Max Planck, Albert Einstein proposed in 1905 that light can indeed behave both ways, thus solidifying the concept of wave-particle duality, one of the tenets of quantum mechanics. The wave and particle nature of light has now been experimentally verified countless times.

Light as a Wave

Light carries energy and information from one place to another, and this energy is carried in the form of electromagnetic (EM) waves. Other familiar types of waves in nature include sound waves and water waves. All three of these types of waves have some differences and some similarities.

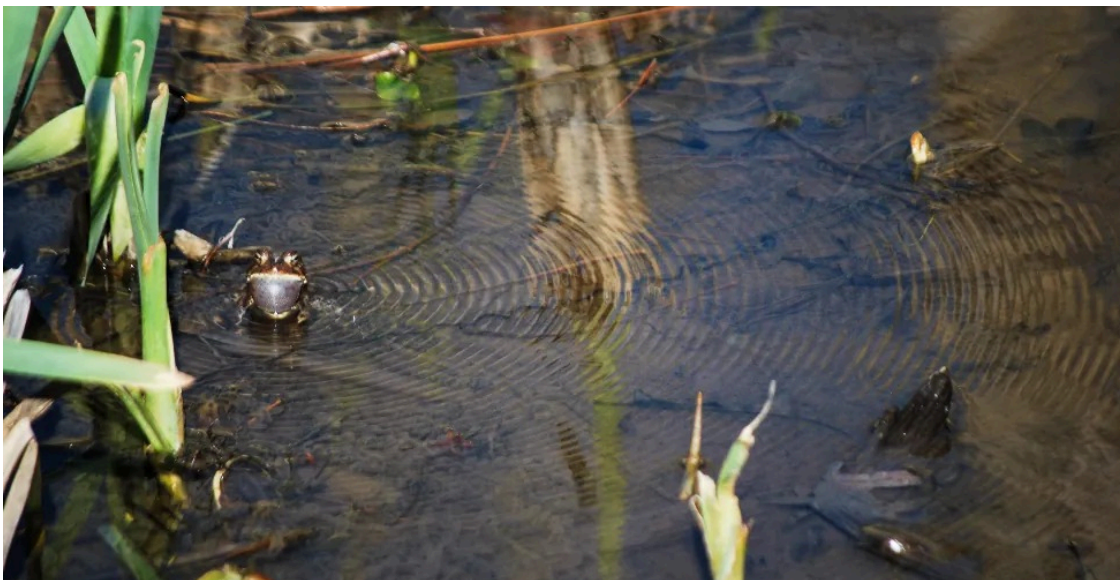


Figure 1 – Making Waves. Water waves created when a frog jumps into a pond. The waves move outward away from the frog, and the distance between each wave peak is called the wavelength. Credit: [OpenStax Astronomy 2e](#), CC BY

Water and sound waves are mechanical waves and require a medium to travel in. For sound waves, the medium is air and the medium is water for water waves. If you screamed on the surface of the Moon, which has no atmosphere, nobody would hear you as the sound waves have no medium to travel through. (The director Stanley Kubrick got it right in the film *2001: A Space Odyssey*.) EM waves do not require water or air or any medium to travel through – they can travel through empty space. This was such a disturbing idea to nineteenth-century scientists that they actually made up a substance to fill all of space—one for which there was no evidence—just so light waves could have something to travel through: they called it the *aether*. Today, we know that there is no aether and that EM waves have no trouble at all moving through empty space (as all the starlight visible on a clear night must surely be doing).

The speed of a sound wave depends on the medium through which it is traveling. A sound wave moves faster through water than it does through air, for example. However, *all* electromagnetic waves move at the same speed in empty space (the speed of light—approximately 300,000 kilometers per second, or 300,000,000 meters per second, which can also be written as 3×10^8 m/s), which turns out to be the fastest possible speed in the universe.

Now for some similarities. All waves are a kind of repeating phenomenon. Whether it is the up-and-down motion of a water wave or the changing electric and magnetic fields in a wave of light, the pattern of disturbance repeats in a cyclical way. Thus, any wave motion can be characterized by a series of crests and troughs. Moving from one crest through a trough to the next crest completes one cycle. The horizontal length covered by one cycle is called the wavelength (λ). Ocean waves provide an analogy: the wavelength is the distance that separates successive wave crests. An example of water waves created when a frog leaps into

a pond is shown in Figure 1.

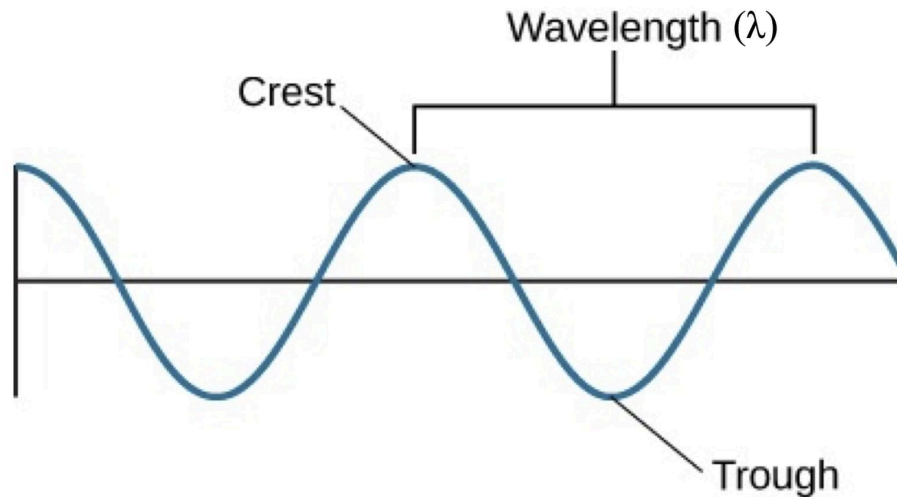


Figure 2 – One cycle of a wave. Credit: [OpenStax Astronomy 2e](#), CC BY

We can also characterize different waves by their frequency, which is the number of wave cycles that pass by per second. If you count 10 crests moving by each second, for example, then the frequency is 10 cycles per second (cps). In honor of Heinrich Hertz, the late nineteenth physicist who, inspired by physicist James Clerk Maxwell's work, discovered radio waves, a cps is also called a *hertz* (Hz). Take a look at your radio, for example, and you will see the channel assigned to each radio station is characterized by its frequency, usually in units of kHz (kilohertz, or thousands of hertz) or MHz (megahertz, or millions of hertz).

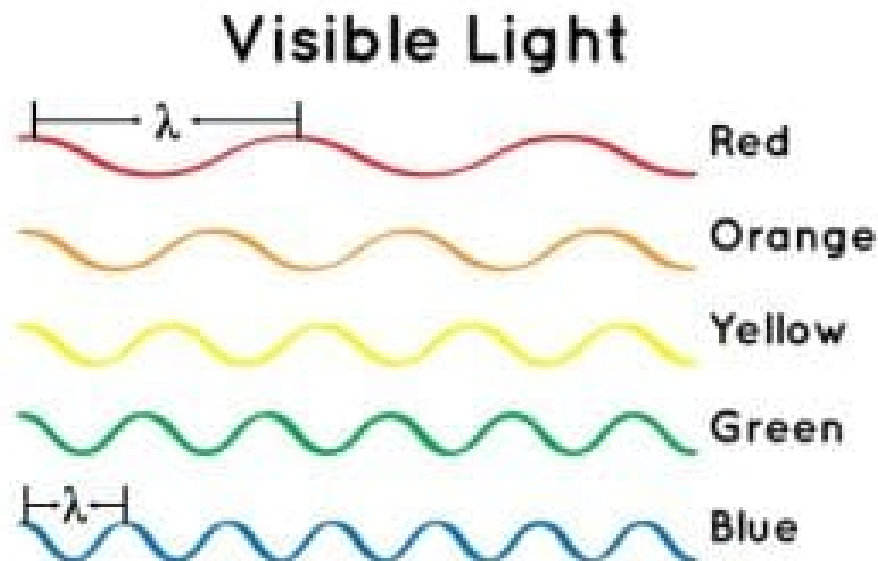


Figure 3 – Wavelength and frequency for different colors of light. Credit, [NASA](#) (modified), public domain

Concept Check: Wavelength and Frequency

Figure 3 is a schematic showing the wavelengths and frequencies for different colors of light. Red light is at the top and its wavelength (λ) is shown as the distance between two peaks, where the peaks are marked by the vertical black lines. Blue light is on the bottom; notice that it has a shorter wavelength than red light (the length between the two peaks is noticeably shorter).

The **frequency** of a wave is a measure of how many waves pass by in one second. Let's imagine that the amount of time that elapsed for each of the waves in Figure 3 is 1 second. Looking at red light, two full waves can pass by in this 1 second period, so we say the frequency is 2 cycles per second, or 2 Hz. (Actually, a bit more than two red light wave cycles can pass by, about 2.5, but we will say 2 full waves to keep this example simple.)

1. Which color of light has the highest frequency? How about the lowest frequency?
2. Which color of light has the shortest wavelength?

Show Answer

1. Looking at Figure 3, the most crests pass by in one second for blue light (6 Hz) while the fewest pass by for red light. Thus, blue light has the highest frequency and red light has the lowest.
2. Since blue light has the highest frequency, it also has the shortest wavelength. It is easy to see in Figure 3 that λ is shortest for blue light.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Wavelength (λ) and frequency (f) are related because all electromagnetic waves travel at the same speed. The formula for this relationship can be expressed as follows: for any wave motion, the speed at which a wave moves equals the frequency times the wavelength. Waves with longer wavelengths have lower frequencies. Mathematically, we can express this as

$$c = \lambda f \quad (1.2.1)$$

where the Greek letter for "l"—lambda, λ —is used to denote wavelength and c is the scientific symbol for the speed of light.

The electromagnetic wave model of light (as formulated by James Clerk Maxwell) was one of the great triumphs of nineteenth-century science. In 1887, when Heinrich Hertz actually made invisible electromagnetic waves (what today are called radio waves) on one side of a room and detected them on the other side, it ushered in a new era that led to the modern age of telecommunications. His experiment ultimately led to the technologies of television, cell phones, and today's wireless networks around the globe.

Light as a Photon

By the beginning of the twentieth century, physicists had to reluctantly accept that sometimes light behaves more like a "particle"—or at least a self-contained packet of energy—than a wave. We call such a packet of electromagnetic energy a photon.

A photon carries a specific amount of energy. How much energy a photon has depends on its frequency. We can use the idea of energy to connect the photon and wave models. How much energy a photon has depends on its frequency. A low-energy radio wave has a low frequency, while a high-energy X-ray at your dentist's office is a high-frequency wave. Among the colors of visible light, violet-light photons have the highest energy and red-light photons have the lowest.

Test whether the connection between photons and waves is clear to you. In the above example, which photon would have the longer wavelength as a wave: the radio wave or the X-ray? If you answered the radio wave, you are correct. Radio waves have a lower frequency, so the wave cycles are longer

The Electromagnetic Spectrum

Objects in the universe send out an enormous range of radiation, or light. Scientists call this range the electromagnetic (EM) spectrum, which are divided into a number of regions. The spectrum is shown in Figure 4, with some information about the waves in each part or band.

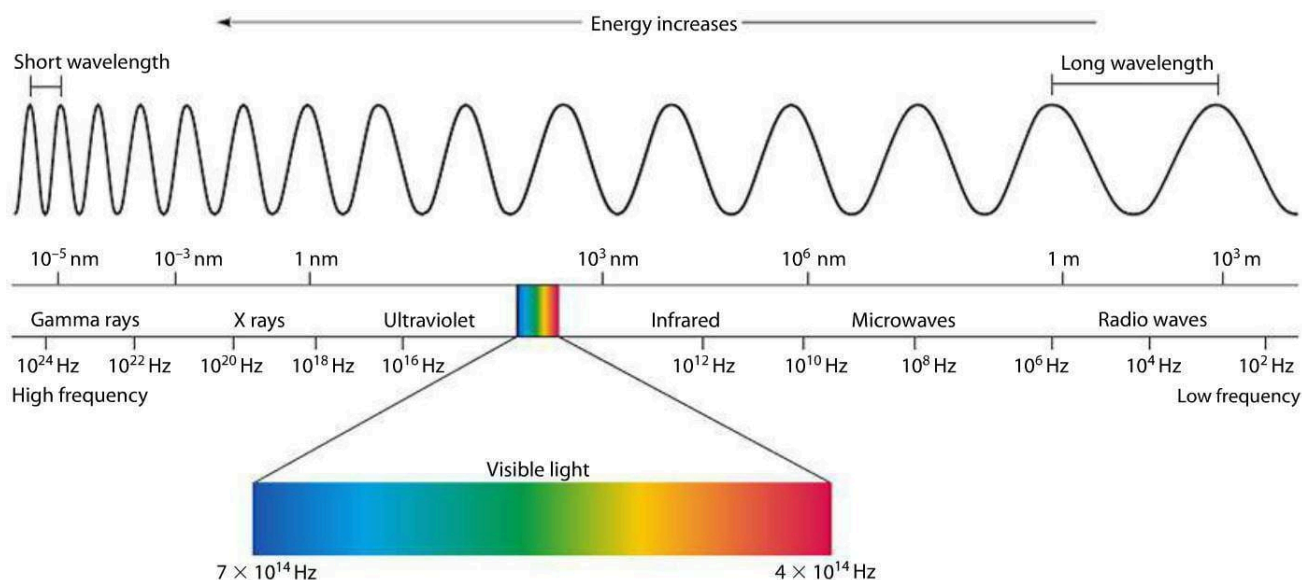


Figure 4 – The electromagnetic spectrum. This diagram of the EM spectrum shows the wavelength ranges for different types of radiation. Credit: https://jila.colorado.edu/~ajsh/courses/astr2030_19/electromagneticspectrum.html

Looking at the EM spectrum in Figure 4, we see that:

1. Gamma rays have the highest energy and radio waves have the lowest energy
2. Radio waves have the longest wavelength and gamma rays have the shortest wavelength
3. Radio waves have the highest frequency and gamma rays have the lowest frequency

These properties of light are summarized by the beautifully simple equation that relates the energy of a photon to its frequency (or wavelength), where h is Planck's constant:

$$E = hf \quad (1.2.2)$$

This relationship shows the wave-particle duality of light, as the energy of photon (a particle of light) is directly related to its frequency (a wave property). Since h has a constant value, you can immediately get the energy of a particular color of light simply by knowing its frequency (or wavelength, which can be expressed as $\lambda = c/f$).

Another property of light that has already been mentioned but is worth re-emphasizing is that **all types of light travel at the speed of light**. Since the speed of light is constant, all light travels at the same speed.

Worked Example: Frequency and Wavelength

Red light has a wavelength of about 650 nm and blue light has a wavelength of about 450 nm. Recall that nm is the abbreviation for nanometers.

What is the frequency of red light?

The frequency of any type of light is related to its wavelength as $f = c/\lambda$, where c is the speed of light.

To use this equation, we must be sure that the units are all consistent. If we use 3×10^8 m/s for the speed of light, then the wavelength must be in units of meters and the frequency will be in units of Hz (where 1 Hz = 1 cycle per second). So, we must first convert the wavelength from nanometers into meters. One nanometer is one-billionth of a meter: 1 nm = 10^{-9} m so it follows that 650 nm = 650×10^{-9} m. Expressed in scientific notation, the wavelength of red light is 6.50×10^{-7} m.

$$f = c/\lambda \quad (1.2.3)$$

That's a huge number of cycles in one second! Expressing this in scientific notation, $f = 4.62 \times 10^{14}$ Hz.

The wavelength of visible light (like red and blue) is often reported in THz, where T is the prefix one trillion, or 10^{12} .

Extra 1: Show that the frequency can also be expressed as 462 THz.

Show Answer

The frequency 4.62×10^{14} Hz can also be written as $4.62 \times 10^2 \times 10^{12}$ Hz

$4.62 \times 10^2 = 462$ and 10^{12} Hz = 1 THz, so $f = 462$ THz

Extra 2: Show that the frequency of blue light (with a wavelength of 450 nm) is 666 THz.

Show Answer

$$f = c/\lambda \quad (1.2.4)$$

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Types of Electromagnetic Radiation

Throughout our study of astrobiology, we will encounter all types of light. When we collect light using a camera or sensor on the Earth's surface, that light must first travel through the Earth's atmosphere before it reaches the ground. Depending on the type of light, some of it may be partially or completely absorbed by molecules in different parts of the Earth's atmosphere. This means that some types of telescopes need to be above the Earth's atmosphere in order to detect any light. Also, we will see that this absorbed light will show up as a kind of set of Earth's chemical fingerprints in any observations that contain the Earth's atmosphere; this is the basic idea behind biosignatures for life, and we will probe it more deeply in the SETI chapters.

The way that light is absorbed in Earth's atmosphere for each part of the EM spectrum is shown in Figure 5.

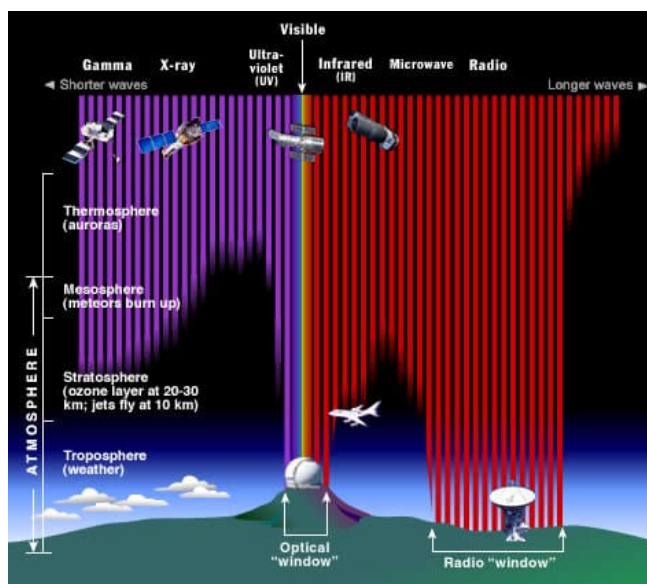


Figure 5 – Absorption of light by Earth's atmosphere. The Earth's atmosphere stops most types of electromagnetic radiation from space from reaching Earth's surface. This illustration shows how far into the atmosphere different parts of the EM spectrum can go before being absorbed. Only portions of radio and visible light reach the surface. Credit: [OpenStax Astronomy 2e](#), CC BY

We highlight a few properties of each type of light that are of relevance to astrobiology, as well as what types of telescopes are used to study the universe in each type of light.

Radio Waves

Radio waves have the longest wavelengths of any type of radiation — they can range from a few millimeters all the way up to

hundreds of kilometers (for reference, Switzerland is about 300 km across). Radio waves have very low energies and are all around us in our everyday lives. On Earth, cell phones, radio and TV transmissions, satellites and radar all emit radio waves and these pass right through our bodies. Radio waves are also produced in nature by a number of astrophysical objects, including stars and pulsars. The most abundant element in the cosmos — hydrogen — also naturally emits radio waves at a very specific wavelength of 21 cm (which corresponds to a frequency of 1420 MHz). As we will see later in the chapter on SETI, astronomers search for signals from advanced extraterrestrial civilizations that are close to the 21 cm radiation that hydrogen produces.

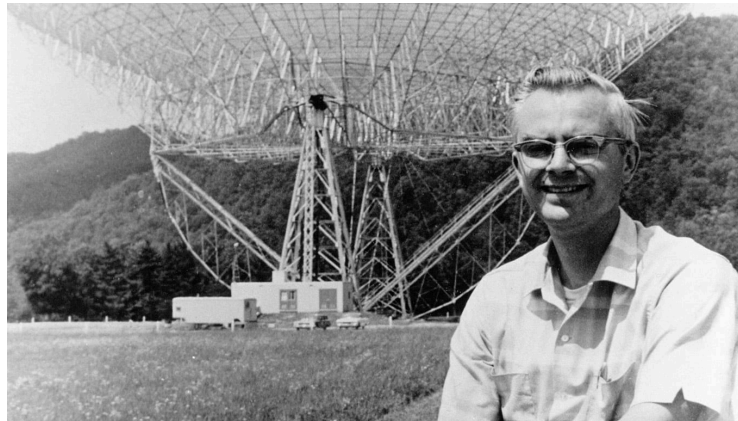


Figure 6 – Astronomer Frank Drake standing in front of the 300-foot radio telescope in Green Bank, West Virginia in 1964. Credit: NRAO (SETI)

Most radio waves reach the ground, as seen by the “radio window” in Figure 5. The lowest frequency radio waves do not make it to the Earth’s surface and this region of the EM spectrum remains unexplored. Because radio waves are long, in general the telescopes (in the case of radio astronomy, these are called dishes or antennae) need to be very large. Some of the first radio observations made to look for signs of intelligent life were carried out by Frank Drake in 1960 using the 85-foot Green Bank Telescope. There is another way to work around the large size needed for dishes through a technique called interferometry. Using this technique, a number of small radio dishes can be made to work in concert and achieve the same precision as a larger dish. The Hat Creek Radio Observatory in northern California, for example, has 28 dishes that are each 6 meters and can all point to the same target on the sky and collectively function as a 40 m dish.

Microwaves

Microwaves are also low energy radiation and have sizes that range from about 1 mm up to 300 mm. The most familiar source of microwaves on Earth is probably microwave ovens. Microwaves are used extensively by satellites to monitor weather on the Earth (Doppler radar maps use microwaves) and for remote sensing of the Earth’s surface. The ancient radiation left over from the Big Bang also happens to fall into the microwave part of the EM spectrum. This Cosmic Microwave Background (CMB) radiation was discovered serendipitously in 1965 and has a wavelength of 1.9 mm. The CMB is literally everywhere on the sky and will appear as “background noise” in any observations that include this wavelength.

As seen in Figure 5, some microwave radiation makes it to the ground (the telescope used to detect the CMB in 1965 was on the ground in New Jersey, USA) but some parts are also absorbed by the Earth’s atmosphere. The COBE, WMAP, and Planck missions all studied the CMB in more detail and were space-based missions.

Infrared

Infrared light is given off by anything with a temperature. This discovery was made in 1800 by the astronomer William Herschel. Herschel studied the rainbow created when white light was sent through a prism and he found that the invisible region beyond the red edge of the rainbow was hotter than the temperature of the rainbow! Herschel called this type of radiation “calorific rays”, or heat rays, and today we refer to these heat rays as infrared (IR) light. Our eyes are only sensitive to visible light (some animals, such as vampire bats and goldfish, can see into the infrared), although we can still “see” this type of light by using an IR camera or night vision goggles. These types of sensors pick up differences in infrared intensity. Since most IR light gets absorbed by the Earth’s atmosphere, many IR telescopes are in space.

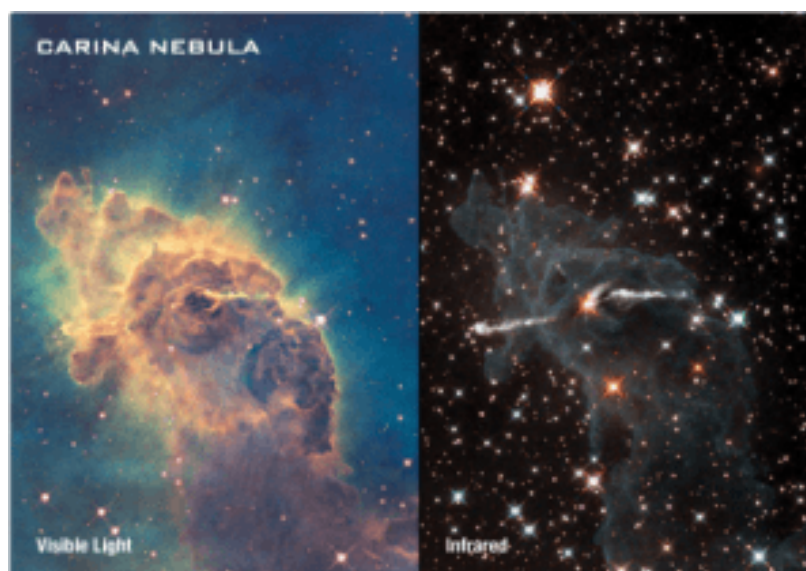


Figure 7 – Visible and infrared light comparison. Images of the Carina Nebula taken in visible light (left) and infrared light (left) by the Hubble Space Telescope. Credit: NASA, ESA and the Hubble SM4 ERO Team

IR light is of great importance in astronomy because it can penetrate, or travel through, dust that is opaque to visible light. This idea is best shown with a comparison of the same picture taken in visible light and infrared light. Figure 7 shows pictures of the Carina Nebula taken by the Hubble Space Telescope (HST) in visible light (left) and infrared light (right). You can see an enormous amount of stars and other objects in the infrared that are not seen in the visible, as infrared light travels right through the dust while visible light is absorbed by the dust.

The James Webb Space Telescope (JWST), which launched in 2021 and has a mirror 2.7 times larger than the HST (meaning it can collect over 7 times as much light), detects IR light. IR light covers the range of about 1-1000 micrometers, where micrometers are abbreviated as μm and are often referred to as microns; for reference, E. Coli bacteria are 2 microns and a human hair is 100 microns. JWST can sense wavelengths in the range of 0.6-28 microns. One of the main goals of JWST is to study planetary systems and the origin of life. One way JWST is doing this is by studying the atmospheres of exoplanets to see if any molecules needed for life, as we know it, are present. This will be discussed in more detail in the chapters on exoplanets.

Visible Light

Human eyes evolved to detect wavelengths spanning the visible part of the EM spectrum, which covers 400 nm to 750 nm. Stars emit visible light and life on Earth would not exist without the light from the Sun. Microscopes and optical telescopes both take visible light and focus it into an image — the study of visible light is essential to astrobiology. For example, detailed pictures of the surface of Mars that show evidence for past liquid water are taken with visible cameras like Mastcam-Z on the Perseverance rover.

Visible light is not absorbed by the Earth's atmosphere but it does get distorted as it passes through, thus causing images to become blurry. Stars as viewed from the Earth's surface appear to twinkle for this reason. However, optical (visible) telescopes on the Earth's surface can use techniques such as adaptive optics to correct for the atmospheric distortion of visible light. The Hubble Space Telescope does not need to worry about this when it collects visible light.

Some searches for intelligent extraterrestrial life in the universe are carried out using optical telescopes; these searches typically scan the sky and look for anomalous pulses, perhaps from lasers being used by a distant civilization.

Ultraviolet

Ultraviolet (UV) light has a shorter wavelength than visible light and ranges from about 10-400 nm. Stars hotter than the Sun emit primarily UV light but the Sun emits some UV light, too. This is probably most apparent when you are outside on a sunny day and forgot to bring some sunblock — your skin will absorb some of the UV (specifically, UVA) light. The fact that stars emit UV radiation is important for many aspects of life, in addition to protecting your skin. UV light can actually break the bonds between the hydrogen and oxygen atoms in a water molecule in a process called photolysis and this can create ozone in the atmosphere,

which in turn protects life from dangerous radiation that can damage cells.

UV light is emitted from a range of astrophysical sources, including sites of star formation and planetary aurorae. As seen in Figure 5, nearly all UV radiation is absorbed by the Earth's atmosphere so UV telescopes are almost always in space, although some telescopes on high mountain tops, like the Keck telescopes atop Mauna Kea in Hawaii, have UV instruments. NASA's SWIFT satellite is actively studying the universe in UV light, and the Hubble Space Telescope has been probing a wide range of UV sources since 1990.

X Rays

X rays are a high energy type of radiation, with short wavelengths of just 0.01-10 nm, and they can be absorbed by tissue and bones in our bodies. They can destroy DNA so great care is taken when X ray machines are used to take medical images. Fortunately for life, X rays are absorbed by the Earth's atmosphere so they are not a concern on Earth's surface (but are a great concern for astronauts traveling above the Earth's atmosphere).

X rays are generated by high energy astrophysical phenomena such as supernova as well as the Sun's corona. During a total solar eclipse, the very hot gas in the Sun's corona can be seen. The composition of soil on Mars is studied using X ray instruments on Martian rovers, such as Spirit and Opportunity and the active Perseverance mission.

Gamma Rays

Gamma rays are the most energetic type of radiation in the cosmos and have wavelengths on the order of the size of the nucleus inside an atom. On Earth, gamma rays are produced naturally by lightning and radioactivity. Gamma rays are also created when two neutron stars merge, in addition to the gravitational waves that these events create. Gamma rays produced in neutron star mergers produce many of the heavy elements on the periodic table, such as gold and platinum.

Gamma rays do not make it to the Earth's surface so gamma ray observatories need to be in space. The Fermi Space Telescope has been studying gamma rays on the sky since 2008 and has created a complete map of the sky in gamma rays. The surface composition of a planet can be studied using gamma ray instruments — the MESSENGER mission to Mercury carried a gamma ray spectrometer.

Multiple Wavebands

To fully understand an object, we ideally want to look at in every type of light that it emits. For many object, this can include all types of radiation, from radio waves to gamma ray. An example of this is shown in Figure 8 for the Crab Nebula:

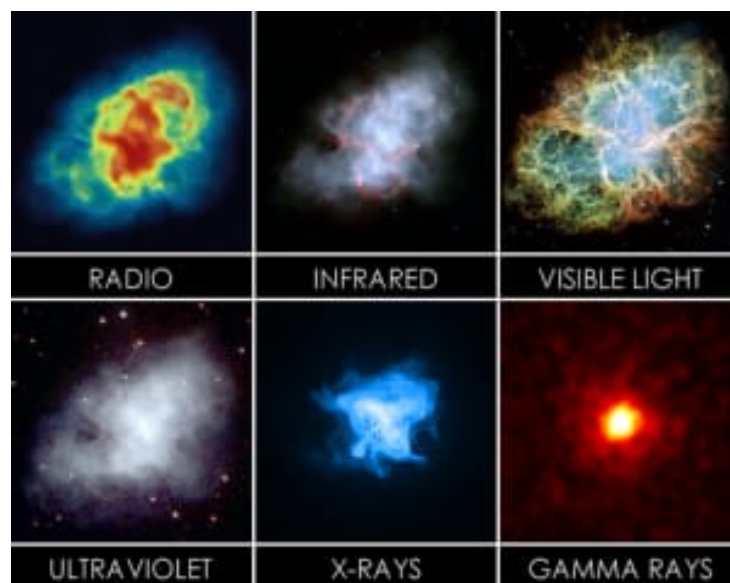


Figure 8 – The Crab Nebula in different parts of the EM spectrum. Telescopes used to create the images: Very Large Array (radio), Spitzer (IR), Hubble Space Telescope (visible), Swift (UV), Chandra (X), Fermi (gamma). Credit: [wikimedia](#), CC BY-SA

Understanding Radiation

Some astronomical objects emit mostly infrared radiation, others mostly visible light, and still others mostly ultraviolet radiation. What determines the type of electromagnetic radiation emitted by the Sun, stars, and other dense astronomical objects? The answer often turns out to be their *temperature*.

At the microscopic level, everything in nature is in motion. A solid is composed of molecules and atoms in continuous vibration: they move back and forth in place, but their motion is much too small for our eyes to make out. A gas consists of atoms or molecules that are flying about freely at high speed, continually bumping into one another and bombarding the surrounding matter. The hotter the solid or gas, the faster the motion of its atoms or molecules. The temperature of something is thus a measure of the average motion energy of the particles that make it up.

This motion at the microscopic level is responsible for much of the EM radiation on Earth and in the universe. As atoms and molecules move about and collide, or vibrate in place, their electrons give off EM radiation. The characteristics of this radiation are determined by the temperature of those atoms and molecules. In a hot material, for example, the individual particles vibrate in place or move rapidly from collisions, so the emitted waves are, on average, more energetic. And recall that higher energy waves have a higher frequency. In cooler materials, the particles have lower energy atomic and molecular motions and thus generate lower energy waves.

Blackbody Radiation and Wien's Law

To further understand the relationship between temperature and light (EM radiation), we consider a type of ideal object called a blackbody. A blackbody absorbs all light that hits it; none of the incoming light is reflected away or passes through the object. The object heats up and gains energy; the object then loses this energy by emitting light at all wavelengths (recall that light is a type of energy). In other words, the object cools off until it reaches the same temperature as its surroundings — this is called thermodynamic equilibrium — and it does this by converting the energy it absorbed into energy of motion that is radiated in a specific way (this radiated energy is called thermal or blackbody radiation). A glowing ember of coal cools by radiating energy until it reaches the same temperature as its environment.

Any dense, solid object that has a temperature emits blackbody radiation; this includes people, animals, stove tops, light bulbs, stars and planets. For example, the Earth absorbs EM radiation from the Sun, heats up, and then emits radiation as mostly infrared light. Note that any objects that reflect light, such as a book with a red cover, are not blackbodies. Similarly, if light passes through the object, as with a glass table, it is also not a blackbody. In that respect, the Earth is not a perfect blackbody, since clouds and other particulate matter do reflect some sunlight.

No object in nature is a perfect blackbody but many opaque objects behave like a blackbody so it is a very good approximation and we can use some fairly simple math to learn more about the object. As mentioned above, blackbody radiation is not emitted at just one wavelength or one part of the EM spectrum, but covers *all* wavelengths. At first this may sound strange — since humans act like blackbodies, does that mean that our bodies emit dangerous X rays and gamma rays? No, that is not the case (I don't have X ray eyes) and any blackbody emits the most intense radiation at one particular wavelength. Humans have a typical temperature of 98.6°F (or 310 K) and emit the most intense radiation at 9.3 microns (an infrared wavelength, meaning we radiate heat!). The peak wavelength of the radiation that a blackbody gives off depends only on its temperature. This means two very different objects with the same temperature emit identically with the same wavelength of peak intensity. The Earth is just a little bit cooler than a human at 288 K, and its most intense radiation is at 10.1 microns. This relationship between the temperature of a blackbody and its peak wavelength is called **Wien's Law** and is:

$$\lambda_{\text{max}} = \frac{0.0029}{T} \text{ m} \quad (1.2.5)$$

where the wavelength is in meters and the temperature is in K (the constant 0.0029 has units of m × K). You can easily verify λ_{max} for a human at 310 K: $(0.0029 \text{ m K})/(310 \text{ K}) = 9.3 \times 10^{-6} \text{ m} = 9.3 \text{ microns}$.

The thermal radiation emitted by any blackbody shows a similar shape in the intensity of light at different wavelengths. These shapes are called blackbody curves or thermal energy curves. Figure 9 shows this general shape for blackbody curves for objects of different temperatures. The shape is similar for all temperatures but notice that the intensity for each blackbody curve peaks at a different maximum wavelength.

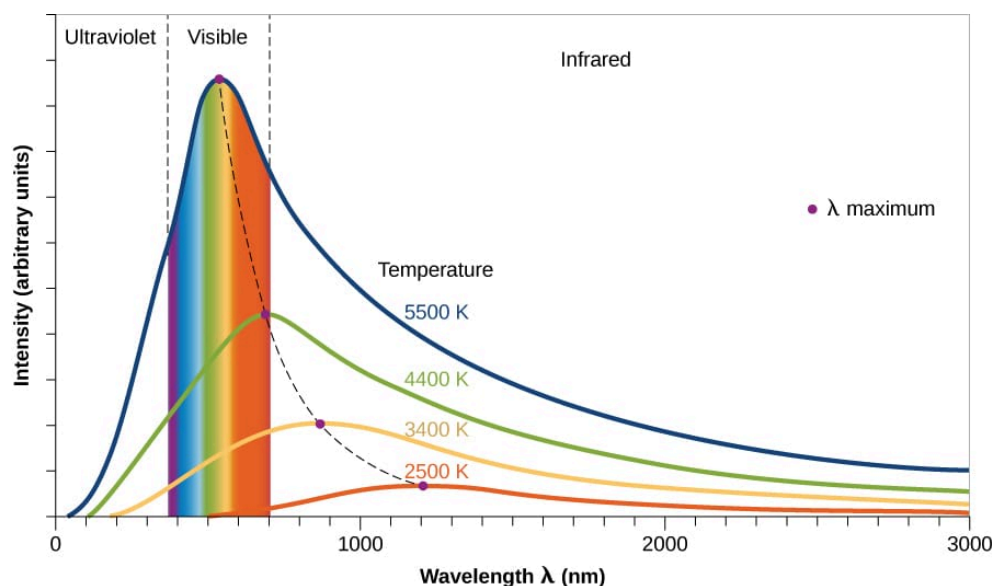


Figure 9 – Wien's Law Illustrated. This graph shows how much light is given off at each wavelength for objects at four different temperatures. The wavelengths corresponding to visible light are shown by the colored bands, with UV and IR to the left and right. Note that at hotter temperatures, more energy (in the form of photons) is emitted at all wavelengths. The higher the temperature, the shorter the wavelength at which the peak amount of energy is radiated (this is known as Wien's law). Credit: [OpenStax Astronomy 2e](#), CC BY

Notice in Figure 9 that the curves show that, at each temperature, the blackbody objects emit radiation at all wavelengths (all colors) but that the most energy is emitted at a peak wavelength. Physically, this peak corresponds to the average speed of atoms or molecules inside the object, which is a manifestation of its temperature.

The hotter the object, the shorter the peak wavelength — the object at 5500 K has a shorter λ_{max} (530 nm) than the object at 2500 K (which has λ_{max} of 1160 nm and is in the infrared portion of the EM spectrum). It makes sense, then, that hot objects give off a larger fraction of their energy at shorter wavelengths (higher energies) than do cool objects. You may have observed examples of this rule in everyday life. When a burner on an electric stove is turned on low, it emits only heat, which is infrared radiation, but does not glow with visible light. If the burner is set to a higher temperature, it starts to glow a dull red. At a still-higher setting, it glows a brighter orange-red (shorter wavelength). At even higher temperatures, which cannot be reached with ordinary stoves, metal can appear brilliant yellow or even blue-white.

We can use these ideas to come up with a thermometer for measuring the temperatures of stars. Because many stars give off most of their energy in visible light, the color of light that dominates a star's appearance is an indicator of its temperature. If one star looks red and another looks blue, which one has the higher temperature? Because blue is the shorter-wavelength color, it is the sign of a hotter star. (Note that the temperatures we associate with different colors in science are not the same as the ones artists use. In art, red is often called a "hot" color and blue a "cool" color. Likewise, we commonly see red on faucet or air conditioning controls to indicate hot temperatures and blue to indicate cold temperatures. Although these are common uses to us in daily life, in nature, it's the other way around.)

What about the Sun? The Sun has a surface temperature of 5800 K so Wien's law immediately tells us that λ_{max} is 500 nm. Which corresponds to green light. But the Sun doesn't look green, it looks yellow or white! What's going on?! This makes sense when you consider that the Sun is also emitting red light and blue light; these all mix together as a nearly white color. Pure purple stars are also never seen for similar reasons — the blue and violet mix into a more deep blue color.

Color Temperature

Different wavelengths (or frequencies) of light correspond to different colors. But what exactly is color? Is it a property of an object, like its size, or is it something that our brains perceive? Our eyes receive light, which is sent to the brain for processing. When we look at an object, we are seeing the light that is reflected off that object. If there is no light in the room, then we do not see anything at all. When light strikes an object, some of the light is reflected and some of it is absorbed. Human eyes have evolved to see visible light; some other animals, such as birds, can see ultraviolet light! Photoreceptors called cones in the retina of our eyes are sensitive to red, blue, and green light. In that sense, every color can be broken down into a combination of red, blue, and green. This is why television only needs those three colors to accurately simulate reality. In low light conditions, rods in

the retina take over and are only sensitive to one color; hence, we see shades of gray.

The Sun emits white light, which contains all colors. As an example, let's consider what happens when sunlight strikes a house painted red. The white light hits the red house, and only red light is reflected to our eyes; all of the other colors are absorbed by the paint molecules. Pigments are substances that selectively absorb very specific wavelengths of light, leaving only the desired wavelength of light reflected.

How does the color emitted by an object relate to its temperature? Wien's law gives a mathematical relationship for this and shows that the temperature of a blackbody is inversely proportional to the wavelength of the light emitted. As a more visual example, consider a piece of metal being heated up. First it will turn a dull red color, then orange to white, and eventually it becomes blue when it is hottest. The scale is shown in Figure 10, where the temperatures are in Kelvin.

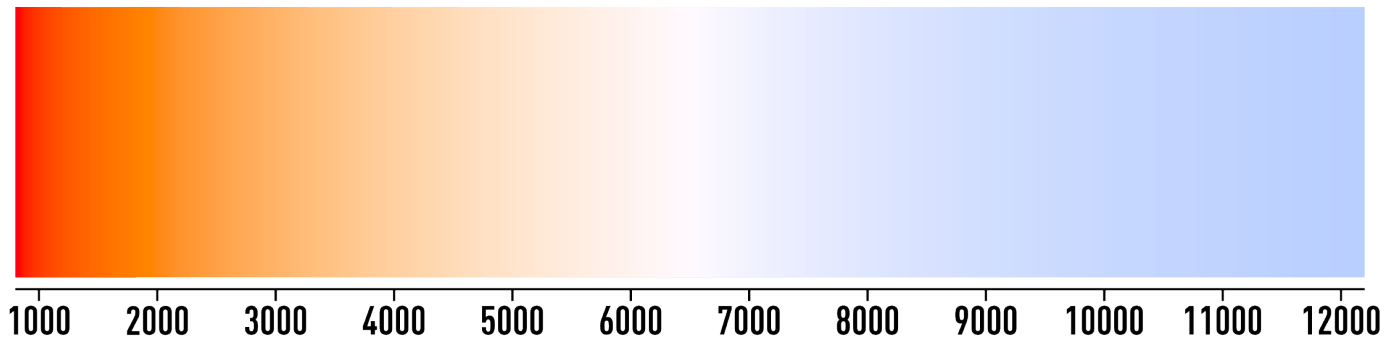


Figure 10 – Color-temperature scale (values in Kelvin). Credit: Wikipedia, CC SA-BY

Light bulbs are designed with color temperature in mind. This is likely a familiar concept to photographers when selecting back lighting for a shoot. For a “candlelight” type background, a bulb of around 2000 K will be best, while a “natural daylight” background is best achieved with a bulb at 4500 K. Most fluorescent light bulbs in classrooms are around 5000 K while those in a library will be closer to 3000 K.

Luminosity and Brightness

Wien's Law shows that blackbodies, like stars, emit the most intense radiation at a specific wavelength, λ_{max} . What exactly is this intensity a measure of?

Luminosity is the total amount of energy that an object (like a star) puts out each second. It has dimensional units of energy per second. In the same way that a 100 W bulb will always put out 100 Watts whether we are standing close or farther away, the luminosity of a star does not depend on our distance from it.

However, astronomers do not measure luminosity directly with a telescope; they measure brightness: the luminosity that is intercepted by a detector such as a photographic plate or a digital camera. If you imagine a spherical surface – a bubble – around a star, then the luminosity is the integrated (total) light from the surface of that bubble. The brightness, which is the luminosity per unit area, decreases as the surface area of the spherical volume increases. This is the same phenomenon that happens with expanding balloons. The balloon has a certain amount of material, usually latex or rubber. Analogous to luminosity, that amount of material is constant, no matter how much air is in the balloon. However, as the balloon expands, that constant amount of material is stretched over a larger surface area. The walls of the balloon get thinner and the amount of material per unit area decreases. The luminosity of a star is constant. The brightness that we measure depends on whether we are “up close” or far away from the star.

This is shown in Figure 11, where the light from a star with luminosity L spreads out into spheres of increasing surface area as it travels away from the star. At a distance d from the star, we can measure the brightness (B) by dividing the total power emitted by the star (L) by the surface area of the sphere that the light has now spread into:

$$B = \frac{L}{4\pi d^2} \quad (1.2.6)$$

This idea—that the apparent brightness of a source (how bright it looks to us) gets weaker with distance in the way we have described—is shown in Figure 11 below. At point 1, the light is concentrated into one box. By the time the light reaches point 2,

which is twice as far as point 1, it is now spread out into *four* squares.

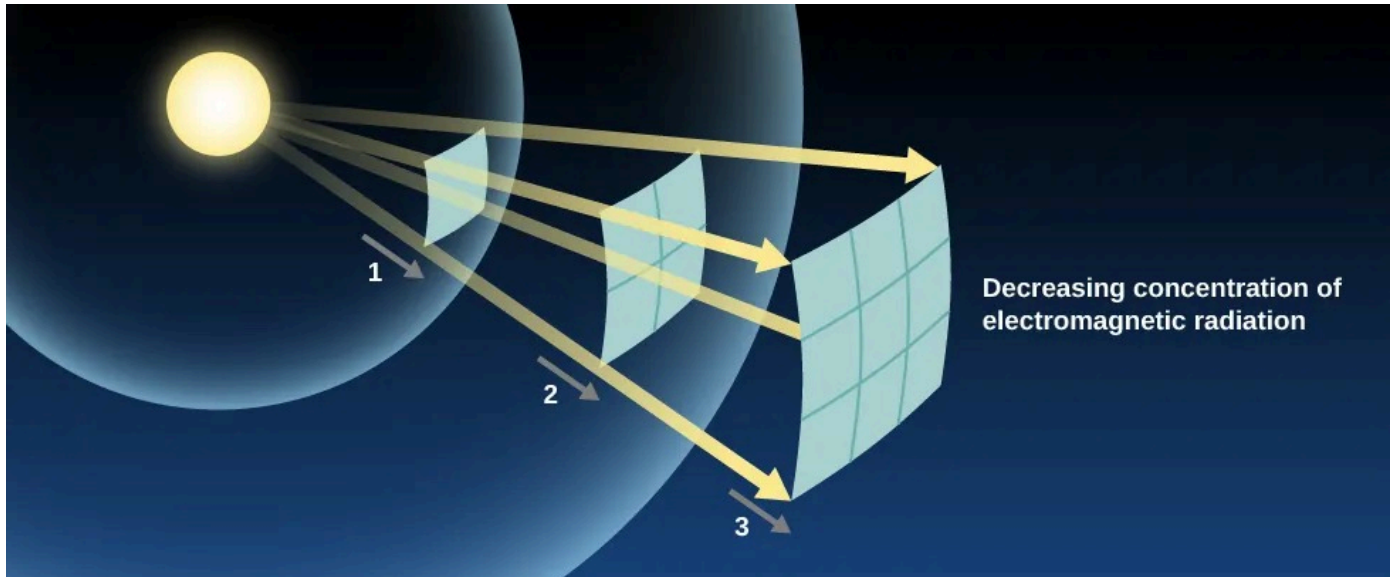


Figure 11 – Inverse Square Law for Light. As light radiates away from its source, it spreads out in such a way that the energy per unit area (the amount of energy passing through one of the small squares) decreases as the square of the distance from its source. Credit: [OpenStax Astronomy 2e](#), CC BY

Concept Check: Brightness and distances of stars

You observe two stars, named Sol-2 and Sol-3, that have the exact same luminosity as the Sun (this means they have the same temperature, radius, and luminosity as the Sun). You measure the brightness of both stars with the same instrument and find that the light from Sol-2 is twenty-five times brighter than the light from Sol-3. Which star is closer to the Earth and by how much? Explain your reasoning.

Show Answer

Ans: Sol-2 is five times closer than Sol-3. Brightness drops as the distance squared, so the ratio of the distances to Sol-2 and Sol-3 is given by $\sqrt{B_{Sol2}/B_{Sol3}}$ or $\sqrt{25} = 5$.

Let's consider the luminosity of the Sun and the amount of energy from the Sun that reaches the Earth. This will be important to know when we study exoplanets and want to know how much light from the host star reaches the exoplanet. The luminosity of the Sun is $3.9 \times 10^{26} \text{ W}$ — in other words, it emits the same amount of power as 3.9×10^{24} 100 W light bulbs (or 39 trillion trillion 100 W bulbs!). As the light travels away from the Sun, it is spread out more and more, as all of that light needs to cover spheres (bubbles) that have greater surface area. The distance from the Earth to the Sun is 151 million km, or $1.51 \times 10^{11} \text{ m}$. Putting this value into the brightness equation, we find that we receive 1360 W/m^2 of sunlight on the Earth. That means every 1 m^2 patch on the Earth, which is about the size of a standard chess board, receives 1360 Joules of energy every second. Venus is closer to the Sun than the Earth and receives more energy for every square meter, 2600 W/m^2 , while Mars is further and receives only 600 W/m^2 .

Worked Example: Calculating the energy received from the Sun

How much radiation from the Sun is received at Saturn's moon Titan? You can assume that Titan and Saturn are at the same distance from Earth.

We can find the value by using the relationship for brightness, luminosity and distance:

$$B = \frac{L}{4\pi d^2} \quad (1.2.7)$$

Here, L is the luminosity of the Sun and d is the distance of Titan from the Sun. The distance from the Sun to Titan (Saturn)

is 1.45 billion km (or $1.45 \times 10^{12} \text{ m}$).

$$B = (3.9 \times 10^{26} \text{ W}) / (4 \pi (1.45 \times 10^{12} \text{ m})^2) = 14.8 \text{ W/m}^2$$

This makes sense — Saturn is almost 10 times further from the Sun than the Earth (Earth is 1 AU and Saturn is 9.57 AU), so the value should be significantly lower than Earth's. [In fact, you could also find this value by taking the ratio $(1/9.57)^2$ and multiplying by 1360 W/m^2]

Show mathematically that the Earth receives 1360 W/m^2 from the Sun.

Show Answer

The only difference here is the distance: now it is the distance from the Earth to the Sun, which is 151 million km.

$$B = (3.9 \times 10^{26} \text{ W}) / (4 \pi (1.51 \times 10^{11} \text{ m})^2) = 1361 \text{ W/m}^2$$

Stefan-Boltzmann Law

It is also apparent from Figure 9 that an object at a higher temperature emits more power at all wavelengths than does a cooler one. In a hot object (the taller blackbody curves in Figure 9), for example, the atoms have more collisions and give off more energy. In the world of stars, this means that hotter stars give off more energy at every wavelength than do cooler stars.

We can also describe our observation that hotter objects radiate more power at all wavelengths in a mathematical form. If we sum up the contributions from all parts of the electromagnetic spectrum, we obtain the total energy emitted by a blackbody. What we usually measure from a large object like a star is the energy flux, the power emitted per square meter. The word *flux* means “flow” here: we are interested in the flow of power into an area (like the area of a telescope mirror). It turns out that the energy flux from a blackbody at temperature T is proportional to the fourth power of its absolute temperature. This relationship is known as the Stefan-Boltzmann law and can be written in the form of an equation as

$$F = \sigma T^4 \quad (1.2.8)$$

where F stands for the energy flux (in units of watts per square meter), T is given in Kelvins, and σ (Greek letter sigma) is a constant number ($5.67 \times 10^{-8} \text{ W/m}^2 \text{ K}^{-4}$).

Notice how impressive this result is. Increasing the temperature of a star would have a tremendous effect on the power it radiates. If the Sun, for example, were twice as hot—that is, if it had a temperature of $11,600 \text{ K}$ —it would radiate 2^4 , or 16 times more power than it does now. Hot stars really shine away a tremendous amount of energy.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Key Concepts and Summary

The energy of light is our messenger about the universe. Different colors of light are simply different wavelengths and different wavelengths of light contain different amounts of energy. The longest wavelengths in the electromagnetic spectrum are radio waves and these have the lowest energy. The shortest wavelengths are high energy gamma rays. Unsurprisingly, the visible band of light (where our brain interprets different wavelengths as different colors) matches the key energy output of the Sun – we evolved to optimally navigate this range of energies. By observing the universe with detectors that are sensitive to different energies of light, we can learn about the energy output and the peak temperatures of the objects we detect. Because the brightness of a star or other celestial object decreases with distance, a measurement of distance is needed to back out luminosity (energy per second).

Review Questions

Summary Questions

1. What is energy? Describe two different kinds of energy and give examples of each.

2. What does it mean to say the energy is “conserved”?
3. How are the frequency and wavelength of light related to each other?
4. What is a photon?
5. Which property of light determine how much energy a photon has?
6. What are the seven main parts of the electromagnetic spectrum? For each band (part), give an example of a source of that type of radiation.
7. How does the type of light emitted by an object depend on its temperature?
8. What is the difference between the brightness of an object (such as a star) and its luminosity?
9. How does the total amount of energy emitted differ for hot and cold objects?

Activities

1. **Understanding blackbody radiation.** Open the simulation at https://phet.colorado.edu/sims/html/blackbody-spectrum/latest/blackbody-spectrum_en.html.
 - The default is set to the temperature of the Sun. Click on the “labels” box to show different parts of the EM spectrum (UV, visible, and infrared). What part of the EM spectrum has the most intense light from the Sun.
 - Now set the thermometer to the temperature of the star Sirius. The y-axis now needs to be rescaled to see the value for the peak intensity. Adjust this by using the zoom out button. Once you can see the peak, click on the “Graph Values” box. What is the wavelength of the peak intensity light for Sirius?

Energy flux. Calculate the solar flux (in W/m^2) at the planet Mercury.

1.3: Spectroscopy and Stars

Light from objects is filled with information. By taking the light from an object and splitting it apart into a spectrum, we can decode this light. Spectroscopy is a fundamental tool for astrobiology and allows us to detect exoplanets and study the chemical makeup of their atmospheres. Stars can be classified by their spectra and we can use these classes to understand how much energy different types of stars emit, helping us to determine if a planet around another star could have the right conditions for life.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe how spectroscopy can provide information about an object, such as a star.
- Discuss the difference between continuous, absorption and emission spectra.
- Describe the sequence of spectral types for stars and how they are ordered
- Describe the main stages of life for a low-mass or a high-mass star.
- Explain why the H-R diagram is useful for understanding the properties of different stars.
- Describe the processes that created all of the elements on the periodic table.

Spectroscopy

The most familiar example of spectroscopy is the rainbow of colors that is produced when white light is sent through a prism, as seen in Figure 1. A water droplet can also split up (or disperse) light and create a rainbow when a light beam passes through it. Both prisms and water droplets can serve as spectrometers or spectrographs — tools used to disperse light into a spectrum.

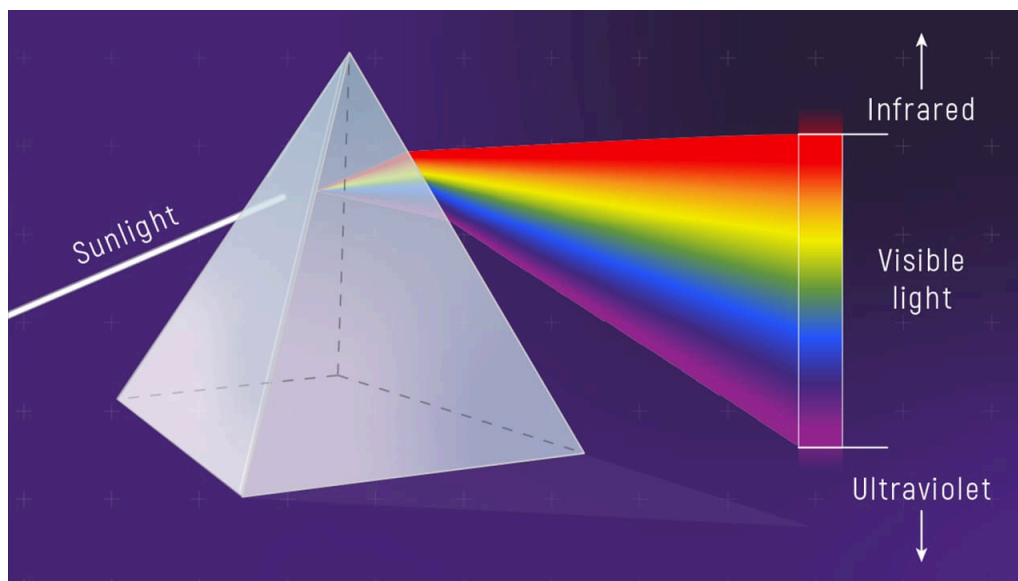


Figure 1 – Different wavelengths of light disperse through a glass prism. Credit: NASA, ESA, Leah Hustak (STScI), public domain

Stars emit all types of electromagnetic radiation but emit the most intense light at a specific wavelength (see [Wien's Law](#)). Note that infrared and ultraviolet light are also emitted by the Sun and are split out as seen in Figure 1 — our eyes just can't perceive these types of radiation. The distribution of colors, which is naturally organized by wavelength (or frequency), is called a **spectrum**. There are three main types of spectra: continuous, absorption, and emission as shown in Figure 2. If the shape of the spectrum on the bottom left of Figure 2 looks familiar, it is because you have seen it before when learning about [blackbody \(thermal\) radiation curves](#).

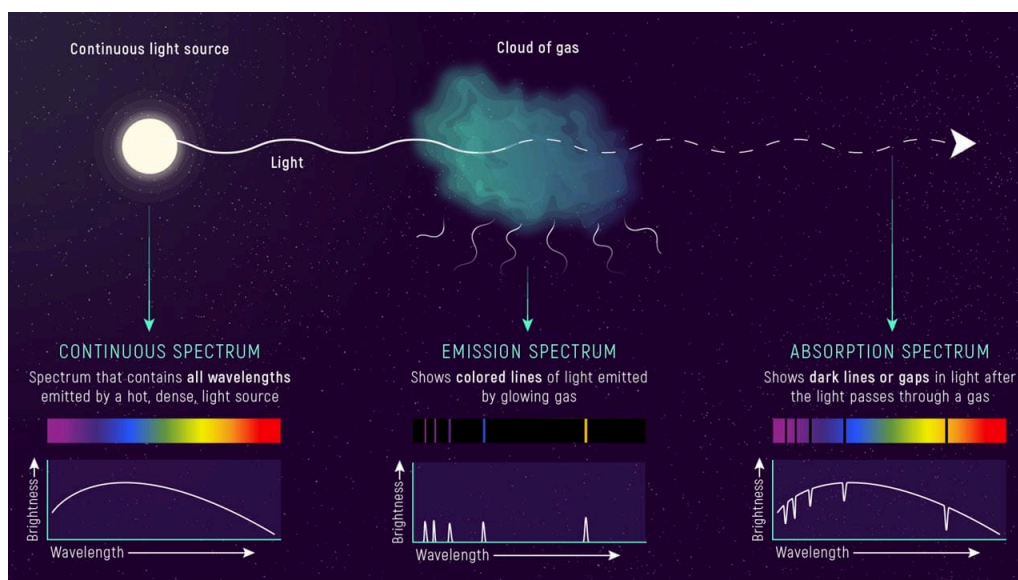


Figure 2 – Formation of different types of spectra: continuous, emission, and absorption. Credit: NASA / STScI, public domain

A continuous spectrum looks like the rainbow shown in Figure 1. However, sunlight that reaches the Earth actually is an absorption spectrum, not a continuous spectrum. Why? The Sun does emit a continuous spectrum from its core but some of that light is absorbed by different atoms as it passes through the Sun's atmosphere. This is shown in Figure 2. On the left, the Sun emits a continuous spectrum at its surface. But that light passes through the atmosphere (the cloud of gas in the middle) and some of that light is absorbed by atoms in the gas. The spectrum on the right is the type of spectrum seen for the Sun, where the dark lines correspond to the wavelengths associated with the atoms or molecules that absorbed the light. Figure 3 shows an absorption spectrum for the Sun, with some of the atoms and molecules present in the Sun's atmosphere labeled. This spectrum was taken from the Earth's surface, so the O_2 (molecular oxygen) lines are from Earth's atmosphere.

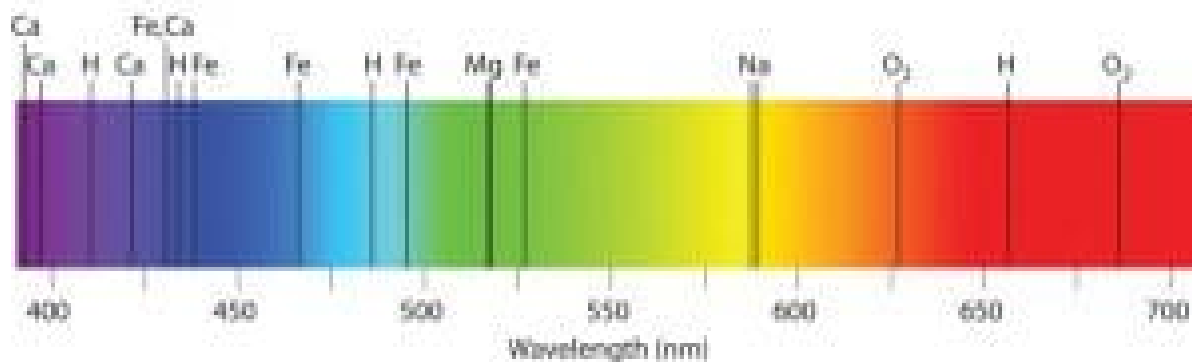


Figure 3 – The absorption spectrum for the Sun. Each black line indicates the presence of a specific atom in the Sun's atmosphere. For example, we can see two sodium (Na) lines at 590 nm, which tells us that there are sodium atoms in the Sun's atmosphere. Credit: continuous spectrum

What does it mean to say “the wavelengths associated with the atoms”? Every element on the periodic table has a different number of protons and thus a different structure. Hydrogen is the simplest atom, with just one proton and one electron orbiting the proton in a neutral atom. To get the electron out of the hydrogen atom entirely requires a certain amount of energy. If there is no energy supplied to the atom, it is stable and the electron is in what we call the ground state of energy. However, the electron can move to different energy levels above the ground state — imagine rungs on a ladder, with the ground state at the bottom and the top rung being the amount of energy needed to kick the electron out of the atom. Hydrogen has a unique set of energy levels and each level has an exact amount of energy. And recall that this energy has an exact frequency or wavelength given by $E = hf = hc/\lambda$. When light particles (photons) with exactly one of those wavelengths encounters the atom, it is absorbed. And every element has a different set of energy levels, meaning a distinct set of its own characteristic wavelengths. We thus have a very powerful way of learning which elements are inside a star's atmosphere.

Using the unique pattern of an element to identify whether it is inside of an atmosphere is similar to DNA barcoding used in

forensics to identify a match to evidence at a crime scene. Figure 4 shows an example of some of the colors (wavelengths) for a few different elements. No two are the same and we can use these patterns to identify an element in a star or planet's atmosphere. This same technique is used extensively in biochemistry and many other scientific fields. We can think of the unique emission lines associated with an atom or molecule as its **chemical fingerprint**.

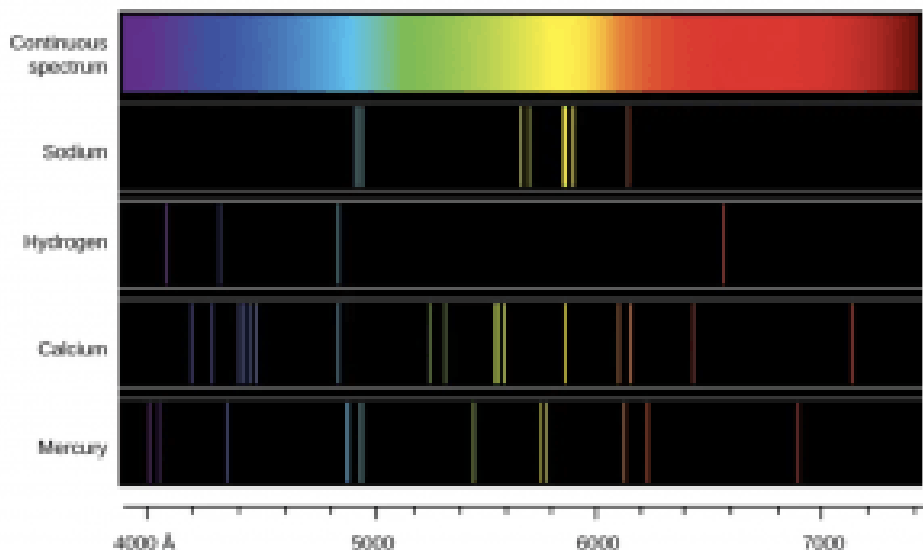


Figure 4 – The “chemical fingerprints” for a few different elements. Every element has its own unique set of spectral lines. Astronomers can use these fingerprints to determine what atoms or molecules are in a star or exoplanet’s atmosphere. **The units for the wavelength axis are called angstroms (Å), and 1 Å is equal to 10^{-10} m. How many angstroms are in 1 nanometer (nm)?** Credit: OpenStax Astronomy 2e, CC BY

The middle part of Figure 2 shows the physical idea of how an emission spectrum is created. Let’s say you have a gas made from atoms of just one type of element. If you heat up that gas, the electrons can move up to higher energy levels and then drop back down, emitting light in the process. This is almost the same situation as the absorption spectrum on the left except we are not looking at the continuous source and the gas but instead just at the gas...so there is no continuous spectrum in the background, just the bright lines that correspond to whatever element is in the gas...so there is no continuous spectrum in the background, just the bright lines that correspond to whatever element is in the gas. The atom is emitting energy at the unique set of wavelengths (colors) for that particular element. Of course, the gas can contain many different atoms and molecules at the same time.

You can take a look at the fingerprints for some other elements using the simulation below. Click on any element to see its unique set of spectral lines. Notice that you can choose to see both absorption (top) and emission (bottom) spectra.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Spectral Types for Stars

In 1814 Joseph Fraunhofer studied the spectrum of the Sun and cataloged the dark lines superposed on a continuous spectrum. A decade later, Gustav Kirchhoff and Robert Bunsen (famous for the Bunsen burner) realized that the dark lines in the spectrum were caused by absorption of light in the atmosphere of the Sun. When other stars were observed, they sometimes had different intensities or completely different absorption lines. The absorption lines in spectra provided a quantitative tool for classifying stars.

Want to Know More: The Harvard “Computers”

In 1890, the Director of the Harvard College Observatory, Edward Pickering, needed a workforce to analyze and catalog the extensive photographic data that had been collected. He employed dozens of women “computers” as his assistants. He believed that these women (unappealingly referred to as “Pickering’s Harem”) were ideal for the job because they did excellent work and they could be paid half the salary of men. Henrietta Swan Leavitt, who classified Cepheid variable stars, and Annie Jump Cannon were two of Pickering’s women computers.

Annie Jump Cannon was nearly deaf throughout her career. She was a suffragist and a member of the National Women's Party. In 1935, she created the Annie J. Cannon Prize for "the woman of any country, whose contributions to the science of astronomy are the most distinguished.



Figure 5 – Annie Jump Cannon (1863 – 1941) was one of the “women computers” at Harvard Observatory. She developed a careful classification of stars with seven main groups: O, B, A, F, G, K, M. Credit: [Wikipedia](#), public domain

The Director of the Harvard Observatory, Edward Pickering, obtained spectra for 10,000 stars. He and his staff grouped the spectra according to the strength of hydrogen lines with letters from A through Q. Director Pickering enlisted the help of Annie Jump Cannon, one of his “women computers.” Cannon and Antonia Maury found a more natural ordering that merged the stellar types into the smooth continuous sequence shown in the spectral sequence below in Figure 6. But it was not clear what this classification meant – what physical processes were driving the differences in the spectra?

Annie Jump Cannon expanded the catalog to include more than a million stars, and she developed a new system for spectral classification with seven main groups: **O, B, A, F, G, K, M** and three classes of peculiar stars. The original classification of stars was in alphabetical form, but it was later realized that the letters needed to be rearranged to sort them by temperature to have a measurable parameter. Each spectral type is further subdivided from 0 to 9 (increasing numbers moving toward the next cooler spectral type). Stars of type O are the hottest and type M stars are the coolest. Our Sun is a G2 star. The Sun is cooler than a G1 star and hotter than a G3 star.

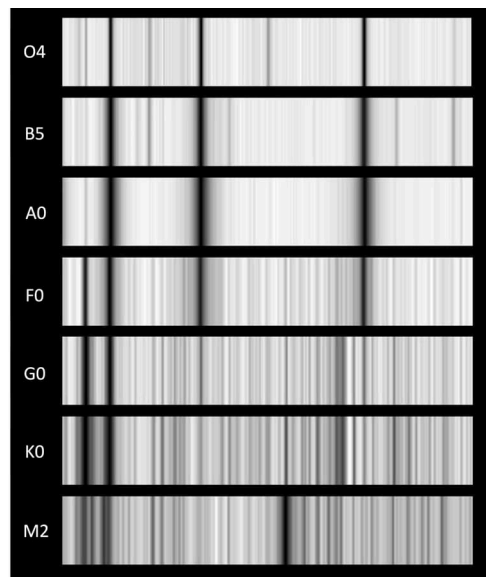


Figure 6 – Spectra were originally recorded with photographic plates. This series of seven different types of spectra show characteristic patterns of dark lines, the signature of missing light at particular wavelengths where atoms like hydrogen in the atmosphere of the star absorb light. It is always true that the spectrum of a K0 or K1 star looks more like the spectrum here for a K0 star than the spectrum of the G0 or M2 star. Credit: [CK-12 Flexbooks](#), CC BY-NC

The spectral type sequence **OBAFGM** is a temperature sequence. The mnemonic “Oh Be A Fine Girl Kiss Me” is an outdated but still useful way to remember the order of the sequence. The Stefan-Boltzmann law shows that the luminosity of a star is related

to its temperature and radius. What determines the luminosity of a star after it is born? The answer is its mass. We'll go into this more deeply in the Star Formation section, but we note here that everything about a star's life is determined by the mass of the cloud of gas and dust that it forms from. A massive O or B star races through its fuel and has a short life, while a low mass M star can take trillions of years to use up all of its fuel and die.

For any type of star, the main source of energy comes from the fusion of hydrogen into helium. All stars are made mostly of hydrogen so there is a lot of it to start with — how much depends on the mass of the star. All stars spend the vast majority of their life in this hydrogen-burning phase. Stars in the hydrogen-burning phase are called **main sequence** stars.

The lifetime of a main sequence star depends on two factors: how much hydrogen exists and the rate of hydrogen fusion. For massive O-type stars, there is 100 times more hydrogen fuel than for stars like the Sun. However, the rate of fusion is so high that the O-type stars burn through this fuel in a million years. In comparison, stars like the Sun will burn hydrogen for about 10 billion years. And the lowest mass M stars will burn hydrogen for more than 100 billion years (since the galaxy is only 13.8 billion years old, there are no M stars that have depleted all of their hydrogen yet). Table 1 shows the different properties of a star in its hydrogen-burning phase.

Class	Temperature	Conventional color	Apparent color	Mass (solar masses)	Radius (solar radii)	Luminosity (solar luminosity)	Hydrogen lines	% of all Main Sequence Stars
O	30,000–60,000 K	blue	blue	60	15	1,400,000	Weak	0.00005%
B	10,000–30,000 K	blue white	blue white to white	18	7	20,000	Medium	0.13%
A	7,500–10,000 K	white	white	3.1	2.1	80	Strong	0.6%
F	6,000–7,500 K	yellowish white	white	1.7	1.3	6	Medium	3%
G	5,000–6,000 K	yellow	yellow	1.1	1.1	1.2	Weak	8%
K	3,500–5,000 K	orange	yellow orange	0.8	0.9	0.4	Very weak	13%
M	2,000–3,500 K	red	orange-red	0.3	0.4	0.04	Very weak	>78%

Table 1 – Properties of stars in their hydrogen-burning phase (main sequence) sorted by spectral type. Hot O stars are at the top and cool M stars are at the bottom. Credit: [CK-12 Flexbooks](#), CC BY-NC

Looking at Table 1 we can note some immediate trends: High mass stars are hotter, more luminous and have a larger radius than low mass stars. Hot stars have a blue color while cooler stars are orange or red.

Concept Check: stellar luminosity

The Sun is a G star.

- About how many times more luminous is the Sun as compared to an M star?
- Does an O star have a smaller or larger radius than the Sun? By how many times?

Show Answer

(a) 30 (b) the O star's radius is 15 times larger

The location of the spectral lines in a star's spectrum can also be used to determine the line-of-sight motion of the star — its radial velocity — through the **Doppler effect**. We explore this concept more deeply when learning about [detecting exoplanets](#).

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Stellar Evolution

Our world has risen from the ashes of massive stars. Those massive stars are rare: they make up just 0.1% of all stars. The synthesis of heavy elements occurs during the final 10% of the lifetime of these stars. Except for hydrogen, the atoms in our bodies were all formed by nucleosynthesis in stars. Without those stars, or specifically without the death of those stars, we would not have the material needed to build planets and needed for prebiotic and biological chemistry.

It bears repeating that the the mass of the star is what drives the rate of fusion reactions and therefore the evolution of a star. The definition of “high mass” is generally taken to be around eight times the mass of the Sun. Low-mass stars are much more common, as seen in Table 1. These lower mass stars take the upper path in Figure 7 below and end their lives as compact white dwarfs. Higher mass stars follow the lower path in Figure 7 and end their lives as exotic neutron stars or black holes.

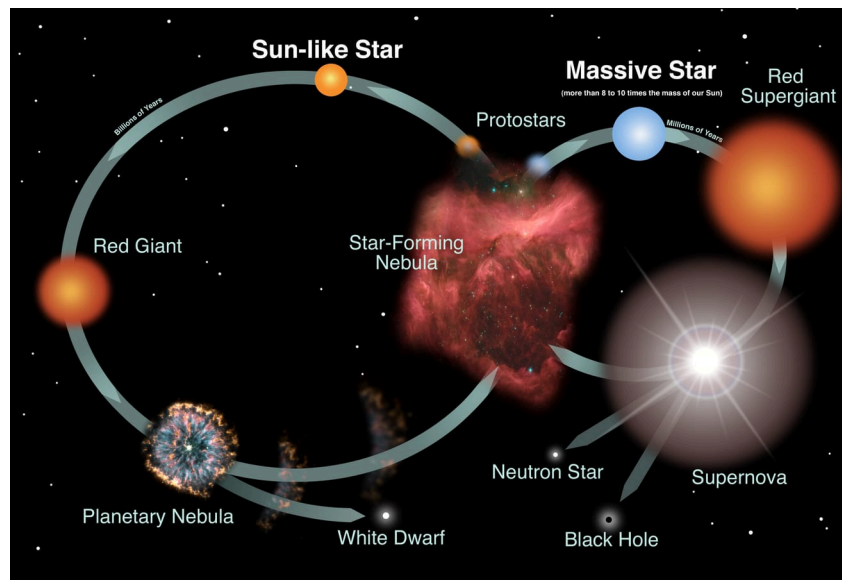


Figure 7 – The evolution and final state of a star are largely determined by the initial mass of the star. **What will happen to our Sun after it is no longer able to carry out nucleosynthesis?** Credit: NASA/Night Sky Network, public domain

Pressure vs. Gravity

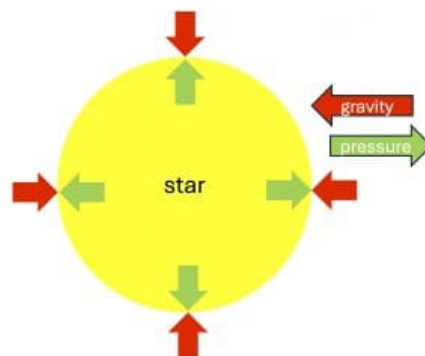


Figure 8 – The competition between gravity and gas pressure for a star. Credit: Allyson Sheffield

As we discuss the different stages of a star’s life, it helps to keep in mind that from the moment a star is born, there is a competition between gravity and pressure. Gravity tends to make the star collapse in on itself, while thermal pressure tends to push the star apart. When the two are balanced, we say a star is in hydrostatic equilibrium.

A main sequence star is in hydrostatic equilibrium. Throughout its life, stars find a way to keep pressure and gravity balanced. However, as we will see, for the most massive stars gravity will eventually win and create a black hole.

Stellar evolution of low mass stars

Red giants

Roughly 99.9% of the stars are AFGKM spectral type stars. These stars fuse hydrogen in their cores on the main sequence and then evolve into red giants when the hydrogen is depleted. When hydrogen runs out, the core collapses and heats up. The red giant stars puff off about half of their mass (hydrogen enriched with helium, and trace metals like carbon, nitrogen, and oxygen), polluting the interstellar medium like an industrial chimney stack. Pressure takes over as a star becomes a red giant and this causes its radius to swell up to 100 times its size as a main sequence star. The star also becomes more luminous at this stage, as additional heating causes fusion to continue more quickly in a shell around the core.

As the outer envelope of the red giant (confusingly called a planetary nebula because astronomers were originally uncertain about these objects) mixes with the interstellar medium, the core of the remnant red giant collapses. If the core mass is less than 1.4 solar masses, it becomes a white dwarf. This is the ultimate fate of our Sun.

White dwarfs

A white dwarf is about the size of the Earth, and it is an incredibly dense object – one teaspoon of white dwarf material weighs several tons. Nuclear fusion is no longer taking place in a white dwarf (except for a short period of hydrogen fusion on the surface). Here, electron degeneracy pressure supports the white dwarf against gravity. This supporting pressure arises from the quantum mechanical nature of electrons. As is the case with atoms, identical electrons are forbidden from occupying the same energy levels and so are forced into higher energy levels that give atoms and white dwarfs their increasing sizes as the number of electrons increase. As long as the mass of the remnant white dwarf is less than about 1.4 times the mass of the Sun, electron degeneracy can support it against further collapse.

Stellar evolution of high mass stars

Red supergiants

As fragments of a cold molecular cloud begin to contract and form stars, hundreds to thousands of stars are born; however, only about 0.1% of the newborn stars will have enough mass to become O or B type stars. These massive stars contract quickly and carry out hydrogen fusion at a furious pace for 1 or 2 million years. At that point they have burned through their endowment of hydrogen. The core collapses until helium fusion begins and the outer shell expands, forming a red supergiant star. There are several cycles of fuel depletion, contraction, and re-ignition as the core of the red supergiant develops an onion layer structure, with stratified shell burning of different elements. Once the core contains iron, the star collapses again, but iron fusion does not produce energy and cannot support the star against gravitational collapse. Therefore, iron is the end of the line for elements formed via fusion; heavier elements are created during the supernova explosion. After the star explodes as a supernova, the core of the star now hits a fork in the road:

- if the core mass is between 1.4 and 3 times the mass of the Sun, then the star becomes a neutron star
- if the core is greater than 3 times the mass of the Sun, then the star becomes a black hole.

Betelgeuse: ready to explode?

Betelgeuse is a red supergiant star that is 12 – 20 times the mass of the Sun with a radius that is almost 900 times the radius of the Sun. Betelgeuse would sweep out almost to Jupiter if it were the center of our solar system. The star can be seen in the Orion constellation – at the shoulder of the famed hunter. Betelgeuse has been in the news lately, because astronomers noticed that this red supergiant started dimming in October 2019 and think that it may explode as a supernova soon. We have a ring-side seat to watch the evolution of this star, but it is impossible to know if we will see this happen next week, or over the next several thousand years. Because Betelgeuse is 640 light years away, it is possible that the star has already gone supernova and we just haven't gotten the memo yet.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=61>

Credit: Space.com / produced & edited by Steve Spaleta

Neutron stars

When the mass of the remnant core is greater than 1.4 times the mass of the Sun, electron degeneracy can no longer support the core against gravitational collapse. As the core collapses, electrons and protons squeeze together to form neutrons with a density similar to an atomic nucleus. The neutrons exert a resisting pressure to gravitational collapse that is similar to that of electrons. Objects supported by neutron degeneracy are called neutron stars. They are far more compact and far more dense than white dwarfs. The diameter of a neutron star is about the size of San Francisco, measuring roughly 10 kilometers (or about 6 miles); a teaspoon of neutron star material would weigh a billion tons. Neutron degeneracy can support a star against further collapse as long as the total mass of the remnant is less than 2 or 3 times the mass of the Sun.

Pulsars

As the remnant core collapses into a neutron star, it spins up, conserving angular momentum. In other words, as its radius gets smaller, it spins faster (just like an ice skater spins faster when they pull their arms in). Neutron stars, when they first form, can rotate hundreds of times per second with a narrow beam of electromagnetic radiation that spins with the star like a lighthouse. If the Earth happens to reside in the path of this beam of light, we see the neutron star as a rapidly blinking source – a pulsar. Pulsars were theoretically predicted going back to the 1930s. In 1968, Jocelyn Bell observed radio emission pulses that confirmed the existence of neutron stars.



Figure 9 – Jocelyn Bell. In 1967, the Irish astronomer, Jocelyn Bell, used radio telescope observations to measure periodic pulses from neutron stars. **Why are the dishes of radio telescopes made from meshes rather than mirrors?** Credit: CK-12 Flexbooks, CC BY-NC

As an interesting historical side note to Bell's discovery of pulsars, when the first pulsar was detected, the source of its regular pulses was unclear. The original name for the first pulsar was LGM-1, where LGM stands for little green men. This was considered to be a possible sign of contact from an extraterrestrial civilization.

Pulsars are of additional interest to the search for life in the universe as the first exoplanet detected in 1992 was found orbiting a pulsar. If this seems surprising, given that a pulsar is what's left after a star explodes as a supernova, it is speculated that the exoplanets may have formed after the explosion.

Black Holes

Electron degeneracy can support a stellar core (or white dwarf) against collapse if the total mass is less than $1.4 M_{\text{sun}}$. Neutron degeneracy can support a core against collapse if the remnant stellar core is between 1.4 and $2-3 M_{\text{sun}}$. If the mass of the remnant stellar core is greater than about 3 times the mass of the Sun, there is nothing that can stop the collapse, and a black hole is formed. Nothing escapes the black hole — not even light — so it is very difficult to find these stellar ghosts. But there are about a dozen candidate black holes in binary star systems where the second star is still visible. We can measure the orbit of the visible star and deduce the presence of a massive, but invisible star, and in some cases we can see gas being funneled off the visible star and heated up to tens of millions of degrees as it spirals onto an accretion disk around something that cannot be seen. If it walks like a duck and it quacks like a duck....

Neutron Star Mergers and Gravitational Waves

There is another possible fate for the most massive stars that are part of binary systems. Once they become either neutron stars or black holes, these objects will spiral in towards each other and eventually merge, releasing an enormous amount of energy in the form of gravitational waves. The first gravitational waves from such a cataclysmic event were detected in 2015 with the LIGO instruments. This process is believed to be the source of production for many elements.

H-R Diagram

The different stages in the life of a star are nicely summarized in Figure 7 above. But we have learned an immense amount about the properties of a star — its temperature, luminosity, and radius — and how they vary throughout a star's life. These properties can be seen dynamically in a figure that shows the luminosity of a star plotted against its temperature. This type of diagram is called the **Hertzsprung-Russell (H-R) diagram**, in honor of the two astronomers who discovered it around 1910.

Figure 10 shows an H-R diagram. Since the temperature of a star is directly related to its spectral type, the x-axis can show either of these variables. The y-axis shows the luminosity compared to the Sun (notice that the Sun has luminosity of 1 on this scale). The majority of the stars shown fall along the main sequence, which runs from the upper left to the lower right. The main sequence is

the phase of a star's life when it is burning hydrogen into helium.

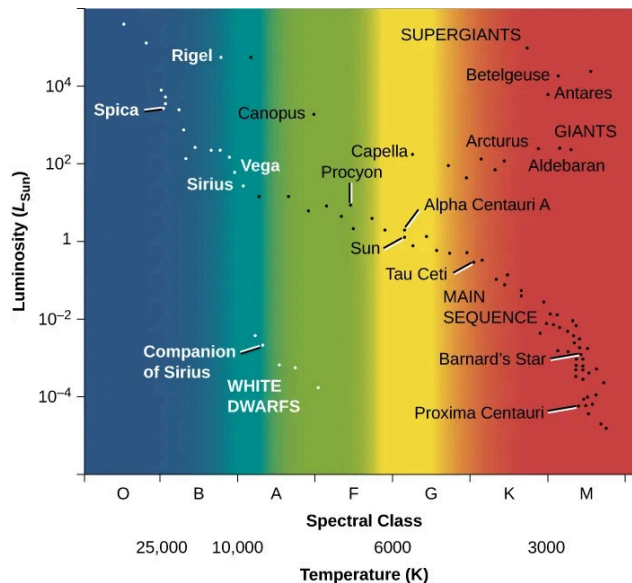


Figure 10 – The H-R diagram for a sample of nearby stars. **Is the supergiant Betelgeuse hotter or colder than the Sun?** Credit: OpenStax Astronomy 2e, CC BY

You can further explore the H-R diagram using the simulation below.

To start, click on “show luminosity class” and plot the nearest stars and the brightest stars. Click on different parts on the H-R diagram (the default, shown by a red X, is the Sun). How does the size of the star change as you move to different luminosity classes?

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Want to know more: Luminosity Class

Stars have both a spectral type (OBAFGKM) and a luminosity class (I, II, III, IV, V). Spectral type is directly tied to the temperature of a star. The energy output of a star (luminosity) is directly tied to *both* temperature and radius. Therefore, two stars can have the same temperature, but they can have different luminosities if their radii are different. You can verify this by locating the star Capella on the H-R diagram in Figure HR. Capella has nearly the same spectral type (and temperature) as the Sun, at G3, but is more than 100 times more luminous. Capella has finished burning through its hydrogen on the main sequence and is now a red giant star.

The radius of a star is constant while it is burning hydrogen – this is the “main sequence” or luminosity class V for a star. Stars of luminosity class V are also called “dwarfs.” When hydrogen is depleted from the core in a star that is similar in mass to the Sun, the core of the star collapses, but the outer envelope expands so that overall, the radius of the star increases. Now, the star evolves across the subgiant branch (luminosity class IV) and the surface temperature cools as the star expands. Even as the outer envelope of the star is increasing, the core of the star is collapsing; when the pressure is high enough in the core, helium fusion begins and the star enters the giant phase (luminosity class III). Stars that are more massive than the Sun will evolve from class V to giants (class III) or supergiants (class II or class I) depending on their mass. There are also subdwarfs (luminosity class VI) and white dwarfs (luminosity class VII) but by convention, those luminosity classes are not often used. Instead, those stars are usually just called “subdwarfs” or “white dwarfs.”

Here is a summary of the luminosity classes:

- **I** – supergiants
- **II** – supergiants
- **III** – giants
- **IV** – subgiants
- **V** – main sequence (dwarfs)

Assembling the Periodic Table

In discussing spectroscopy and stars, a couple of facts can be highlighted: (1) all stars are made mostly of hydrogen; (2) every element has a unique “chemical fingerprint,” or distinct set of spectral lines. We can ask a question related to both of these points: Where did all of the elements in the universe come from?

Following the Big Bang, the only atomic elements that existed in the early universe were hydrogen, helium, and a small amount of lithium. All other elements in the periodic table were formed later through nucleosynthesis and the mergers of neutron stars. Since we need those other elements to form planets and life, we review and summarize the processes that form these elements.

Nucleosynthesis

Nucleosynthesis is the process of building up heavy elements from lighter elements through fusion. In the Sun and other stars, the most common atom is hydrogen. It is possible for hydrogen nuclei (protons) to overcome the strong electrostatic repulsive forces because of the enormous temperature and pressure in the core of a star and fuse into heavier nuclei. The energy output of the Sun is $3.8 \times 10^{26} \text{ W}$ or $3.8 \times 10^{26} \text{ J/s}$ (one watt is equal to 1 joule of energy every second). This energy comes from the fusion of hydrogen into helium in the core of the Sun.

Theories of nucleosynthesis inside stars are able to predict the *relative abundances* with which the elements occur in nature. The way stars build up elements during various nuclear reactions really can explain why some elements (oxygen, carbon, and iron) are common and others are quite rare (gold, silver, and uranium).

Want to know more: Hydrogen Fusion

The exact pathway for nucleosynthesis is very temperature-dependent. In a star like the Sun (or less massive than the Sun), the process is proton-proton (or p-p) fusion and follows the steps outlined in the animation below.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

In addition to making helium, the p-p chain produces the energy that supports the Sun against collapse. One reaction cycle in the p-p chain produces 25 MeV of energy (eV are a very tiny unit of energy... 25 MeV is just 4×10^{-12} joules of energy). In the second step of the p-p chain, a proton is transformed into a neutron. How does that happen?

In the first step of p-p nucleosynthesis, a small fraction of hydrogen transforms into deuterium (an isotope of hydrogen with a neutron added to the nucleus, written as ^2H). This is the rate-limiting step in the p-p chain because the transformation of a proton into a neutron is statistically improbable. Once deuterium is created, the rest of the p-p pathway proceeds relatively quickly.

As hydrogen fusion occurs in the Sun, most of the mass of the hydrogen is converted into an “ash” of helium. However, there is a slight difference of 0.71% between the mass of 4 protons and the mass of a helium nucleus; this fraction of the original mass is converted into energy and the mass loss rate from the Sun, Δm , can be calculated with Einstein’s famous equation:

$$E = \Delta m c^2$$

where c is the speed of light. We can solve this equation for Δm and find how mass is converted into energy every second in the Sun:

$$\Delta m = \frac{E}{c^2} \quad (1.3.1)$$

$$\Delta m = (3.8 \times 10^{26} \text{ J}) / (3 \times 10^8 \text{ m/s})^2 = 4 \times 10^9 \text{ kg} = 4 \text{ billion kg}$$

Think about that... every second, 4 billion kg (or 4.4 million tons) of the mass of our Sun vanishes — it is transformed into energy that is radiated away. However, this represents only 0.71% of the mass of hydrogen that is converted into helium. How much hydrogen is being burned in the sun every second?

$$m_{\text{H}} = (4 \times 10^9 \text{ kg}) / (0.0071) = 5.63 \times 10^{11} \text{ kg} = 563 \text{ billion kg} \sim 600 \text{ million tons}$$

To produce the Sun's luminosity of 4×10^{26} watts, some 600 million tons of hydrogen must be converted into helium *each second*, of which about 4 million tons are converted from matter into energy. As large as these numbers are, the store of hydrogen (and thus of nuclear energy) in the Sun is still *more* enormous, and can last a long time—billions of years, in fact.

Stars that are more massive than the Sun have more gravitational pressure in their cores and hotter temperatures. Massive stars that have trace amounts of carbon, nitrogen, and oxygen (CNO) can take a more efficient path for synthesizing helium, using CNO as catalysts to form helium nuclei. The energy yield is higher and therefore more efficient: 26.72 MeV per helium nucleus instead of 25 MeV per helium nucleus with the p-p pathway.

Manufacturing heavier elements

So, if hydrogen fusion is forming helium, how does that get us to carbon, oxygen, nitrogen, phosphorus, sulfur.... all the good stuff that we need for life? Stars on the main sequence (dwarfs) burn only hydrogen in their cores. Elements with atomic masses greater than helium can be formed only after the star leaves the main sequence. After most of the hydrogen in the core has been fused to helium, hydrogen fusion slows down. This is a problem for the star because H-fusion was supporting the star against gravitational collapse. As the rate of H-fusion declines, gravity begins to win, and the core of the star collapses. This causes the pressure and temperature to increase. When the pressure and temperature are high enough, helium fusion in the core is initiated. During this gravitational collapse, the pressure and temperature of hydrogen gas in a shell surrounding the core “reignites” and begins forming helium.

Next, the helium nucleus (a.k.a. an alpha particle) fuses into carbon and oxygen. For low mass stars (masses less than about 3 solar masses) this is the end of nuclear fusion reactions. Low mass stars do not form most of the elements that we need for life. For this, we need high mass stars.

Massive stars are rare with relatively short lifetimes (millions rather than billions of years). But, massive stars do all of the heavy lifting when it comes to manufacturing high atomic mass elements. Like their lower mass counterparts, massive stars burn hydrogen on the main sequence. But when hydrogen fusion ends, there is a rapid succession of fusion and collapse events. When the core depletes the element it is burning in the core, gravity squeezes the star, and the next stage of core nuclear fusion along with shell burning of lighter elements begins. Close to the end of its life, the massive star has developed a stratified, onion skin structure with the heaviest elements in the core and surrounding shells of successively lighter elements, as shown in Figure 11 below. However, that fusion stops when iron is formed in the core. For elements with lower atomic mass than iron, fusion reactions are exothermic – they give off energy and support the star against gravity. Iron fusion is endothermic – energy is used in this reaction. In other words, iron is not capable of fusing to a heavier element.

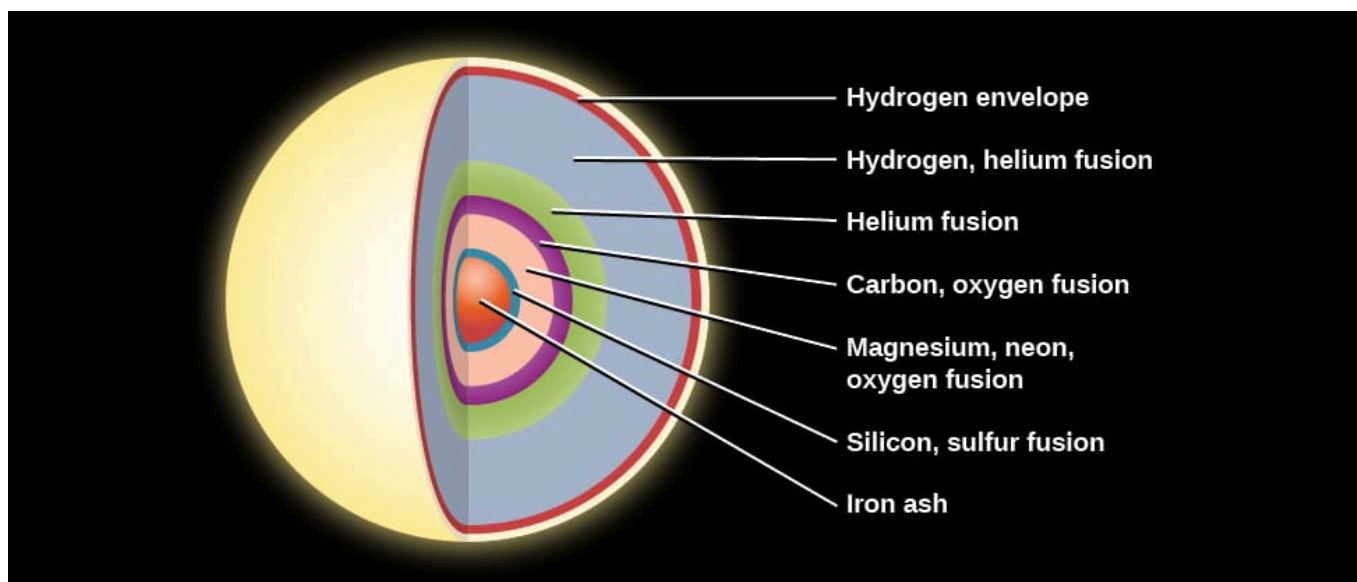


Figure 11 – Nuclear fusion in massive stars. The core of a massive evolved star has shells of material, differentiated by mass that develop through a succession of core-burning and collapse. An iron ash core is the end of the line for a massive star. Why doesn't iron fuse in the core as the star collapses? How are heavier atomic elements manufactured? Credit: [OpenStax Astronomy 2e](#), CC BY

The next move belongs to gravity. Once the core collapse begins, the onion shells in the core mix, and the density increases.

The outer layers of the star fall in, hit the high density core, and bounce off in a fierce supernova event that blows out about half of the material in the star. During the supernova, the shock waves generate high enough pressures to produce the heaviest elements in the periodic table. The final distribution of elements has some interesting structure. The addition of alpha particles favors production of even-numbered atomic mass elements and leaves behind an even-odd pattern in the logarithmic elemental abundances. There is an obvious iron peak in the elemental abundances with higher-mass elements being increasingly rare.

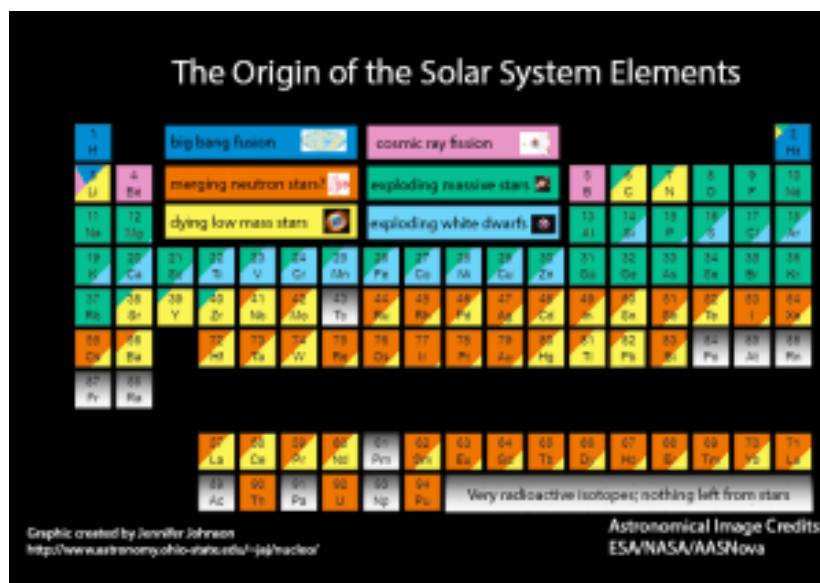


Figure 12 – Periodic table of the elements color-coded by the site of formation.

Key Concepts and Summary

The wavelengths of light encode the energy and chemical composition of stars and other celestial bodies – astronomers use spectrographs to disperse light collected at a telescope into a spectrum. Objects emit a continuous (sometimes called “blackbody”) spectrum with an intensity distribution that depends on temperature. When atomic transitions take place in cooler outer layers of stars, the spectrum can also have absorption or emission lines. Absorption lines occur when atoms absorb photons of light pushing electrons to a higher energy state. Emission lines occur when atoms release photons and the electrons cascade to a lower energy state. The process of absorption (and emission) is sensitive to temperature, so the peak intensity, coupled with the spectrum of a star reveals its temperature and the chemical composition. Stellar spectra have been organized into a temperature sequence: OBAFGKM, from the hottest to the coolest stars. Stars like the Sun burn hydrogen in their cores for about 8-9 billion years – our middle aged Sun is halfway through this process. Once the hydrogen in the core has been fused, the star begins to evolve through a giant phase, fusing successively higher atomic mass elements to produce the elements in the Periodic Table. Stars end their life as a white dwarf, neutron star, or black hole, depending on the mass of the star.

Review Questions

Summary Questions

1. How does spectroscopy give information about an object? What happens when starlight passes through a prism?
2. What are the differences between continuous, absorption and emission spectra? What are the similarities between all three?
3. Why are the spectral lines for an element sometimes called “chemical fingerprints”?
4. What does the spectral type of a star depend on? Why isn’t the sequence of spectral types alphabetical?
5. How does the lifetime of a star depend on its mass?
6. Explain how thermonuclear fusion in the core of a star creates energy. Which element is being fused when a star is on the main sequence?
7. Describe the five main luminosity classes for stars. What luminosity class is the Sun? Betelgeuse?
8. What is displayed on a H-R diagram? How does an H-R diagram help us understand the properties and evolution of stars?
9. How do the spectral type and luminosity class of a star differ from each other?
10. How were the elements on the periodic table created? Explain a few of the different processes and how elements were

created (nucleosynthesis).

Activities

1. **H-R Diagram simulator.** Open the H-R Diagram simulator above (or you can open it in a separate window at <https://astro.unl.edu/classaction/animations/stellarprops/hrexplorer.html>).
 1. Click on “show luminosity classes”; next, plot the nearest stars and then plot the brightest stars. Why do the nearest stars and brightest stars fall into different regions on the H-R diagram?
 2. Look up the temperature and luminosity for the star Vega (be sure that the temperature is in K and that the luminosity is reported relative to the Sun) and enter them in the boxes in the “Cursor Properties” box on the left. How do the size (radius) and color of Vega compare with the Sun?

1.4: Laws of Motion and Gravity

The cosmos is vast and searching for other life within it requires an understanding of how and why objects move. These questions are not new — Aristotle and other Greek philosophers thought about them — and it was through rigorous mathematical testing of observations against theory that allowed Isaac Newton to uncover gravity as the underlying force for the motion of the planets around the Sun. In this chapter, we examine the fundamental laws governing objects in motion and show examples of how to apply these laws to studying planets in other solar systems.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain how Galileo disproved the idea that heavy objects fall faster than light ones.
- Understand how Newton's Laws of Gravity explain Kepler's Laws of Planetary Motion.
- Describe how free fall occurs for satellites in stable orbits about the Earth.
- Explain the difference between mass and weight and the concept of surface gravity.

Understanding Motion

Ancient Greece, from about 600 B.C. to 200 A.D, is notable in the history of science as a time when observations in nature began to be explained using rational thought and reasoning, rather than invoking supernatural explanations. As discussed in [Appendix on the History of Astrobiology](#), the geocentric (meaning Earth is at the center) model of Ptolemy emerged as the accepted explanation of how the cosmos is organized and this model remained unchallenged until Copernicus proposed a heliocentric model in the early 1500s.

Contemplating different models of the cosmos necessitated thinking about why the bodies, celestial or otherwise, move (or don't move) in the way that they do. Physics, particularly kinematics and dynamics, was fundamental to this inquiry and the word physics itself is thought to come from the Greek word *phusis*, meaning nature. Aristotle posited that all objects on Earth seek their natural state of motion, which is to be at rest, unless some force is being applied to the object to keep it in motion. This idea has some aspects of the correct explanation of the motion of objects but misses the fact that an object in constant motion is also in its natural state. Aristotle also stated that heavier objects fall more quickly to the Earth than lighter objects. This is not correct but is easily tested. While Aristotle (or anyone else at this time) did not test this, Galileo conducted many physics experiments and did test this idea and ultimately falsified it.

Galileo also began to show that motions of objects in the heavens and objects on the ground were connected. Greek astronomy held that the motions in the celestial realm were different than the motions seen near the ground and, while the motion of planets and stars could be predicted very accurately, there was no unifying explanation for why things were moving the way they were seen to move. The general consensus was that things in the heaven moved "perfectly" in their own fashion based on ideal laws and patterns and many believed that the objects in the sky were made of substances that were fantastical, brilliant, and glorious unlike the dark, dirty, and "base" material found on Earth.

As for Aristotle's assertion that heavier objects fall faster to the Earth, Galileo found results that yielded a different rule. Galileo's motion experiments were consistent with all objects, regardless of their mass, falling at the same rate. Galileo found that this rate (known as an acceleration) is approximately a change in speed of 10 m/s for every second that an object is falling provided there is no resistance (such as friction) or extra force applied in the direction of falling. This constant acceleration of free fall would feature prominently in future developments of our understanding of gravity. It was used by Isaac Newton to show that mass is what causes gravity and was used by Albert Einstein to argue that gravity was indistinguishable from a reference frame that is under a similar sort of acceleration (this is his so-called equivalence principle).

The video clip below shows confirmation by Apollo 15 astronauts on the Moon that all objects fall at the same speed (the Moon is a great place to test this since there is no air) in the absence of air resistance.

Gravity experiment on Apollo 15



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=64>

Video source: @NASAGovVideo

The original version of this chapter contained H5P content. You may want to remove or replace this element.

In regards to the natural state of an object, Galileo showed that the motion of objects on Earth under the right conditions can continue to move forever without any forces being applied to them. This is the basis of **inertia**, which is the tendency of objects to keep doing what they are already doing. In other words, a stationary object stays put, and a moving object keeps moving unless some force intervenes. Objects speeding along at a constant velocity were physically indistinguishable from objects at rest and Galileo proved this by comparing the movement of objects falling to the ground, rolling down ramps, and flying through the air as projectiles. No matter what the situation, the motion of objects parallel to the ground was completely independent of the falling towards the Earth, and it seemed that the ideal scenario was one where such motion would continue forever unless or until something stopped it. Isaac **Newton's first law of motion** would further clarify this state of affairs by elegantly stating: in the absence of external forces, objects in motion stay in motion and objects at rest stay at rest.

To show that the motion of objects in the heavens obeyed these same rules was a bit harder to explain and some more further insight was needed to connect the two. Kepler's law of motion provided a mathematical framework for the motions of planets but Kepler didn't know that gravity was the unifying concept that would explain this motion. This was left to Newton who figured out that the same force that caused objects to fall to Earth could also explain all the celestial motions.

Kepler's Laws of Motion

Kepler inherited a very accurate catalog of planetary positions that Tycho Brahe meticulously collected over several decades. Through his analysis of the motions of the planets, Kepler empirically developed a series of principles, now known as Kepler's three laws, which described the behavior of planets based on their paths through space. The first two laws of planetary motion were published in 1609 in *The New Astronomy*. Their discovery was a profound step in the development of modern science.

Kepler's First Law

Kepler's First Law states planets orbit the Sun in the shape of an ellipse, with the Sun at one focus. The discovery that planetary orbits are ellipses brought down the long-held belief, going back to Plato and Aristotle, that the orbits were perfect circles.

The path of an object through space is called its orbit. Kepler initially assumed that the orbits of planets were circles, but doing so did not allow him to find orbits that were consistent with Brahe's observations. Working with the data for Mars, he eventually

discovered that the orbit of that planet had the shape of a somewhat flattened circle, or **ellipse**. Next to the circle, the ellipse is the simplest kind of closed curve, belonging to a family of curves known as conic sections.

A circle is a special case of an ellipse. We can think of an ellipse in general as a flattened circle. The amount of flattening of an ellipse is called its **eccentricity** or **e**. The eccentricity of an ellipse varies from 0 to 1. A circle has an eccentricity of 0. Any ellipse also has two interior points called **foci** (singular: **focus**); these are labeled f_1 and f_2 in Figure 1 below. As the eccentricity of an ellipse increases, the foci get further away from the center. For a circle, there is only one focus and it is at the center. As we'll see, the eccentricity for the orbits of most of the planets in our solar system are very close to zero, with Mercury being the exception with an eccentricity of 0.21. If the eccentricity of a planet's orbit is close to zero, this means that it is very close to being circular motion but, importantly, the distance between the planet and the Sun is not constant.

In the mathematical description of an ellipse, the distance from the center of the ellipse to the widest part of the ellipse is the **semi-major axis**, a . This means the diameter of an ellipse is $2a$. The semi-major axis of a planetary orbit is the average distance of the planet from the Sun. Note that the second focus of the ellipse is empty, meaning there is physically nothing there. We can define two important points on a planet's orbit: **perihelion** and **aphelion**. Perihelion is when a planet is closest to the Sun and aphelion is when the planet is at its furthest distance from the Sun.

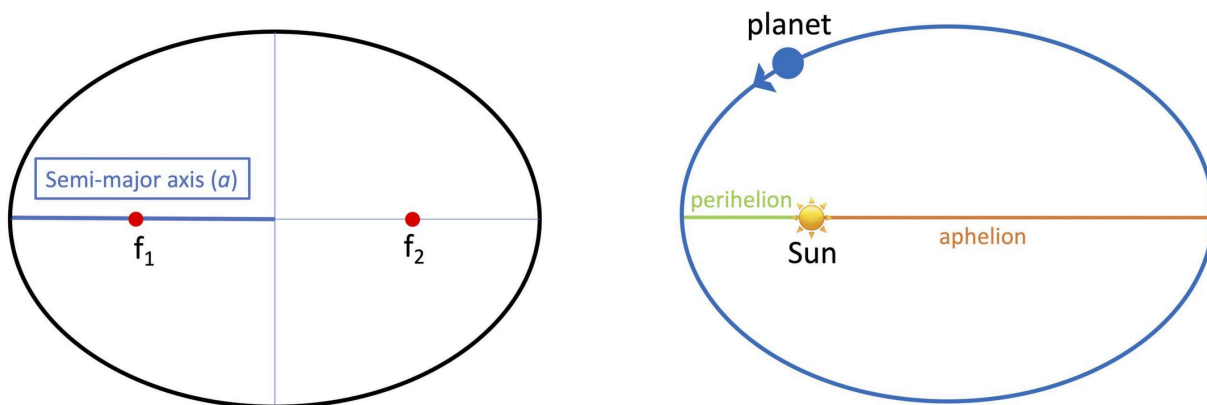


Figure 1 – Properties of an ellipse. **Left:** The two foci are shown as the red points f_1 and f_2 and are the same distance from the center of the ellipse. The semi-major axis, a , is the distance from the center to the edge along the longer axis. **Right:** Planets orbit the Sun in an elliptical shape, with the Sun at one focus. The closest distance the planet gets to the Sun is called perihelion (green line), and the furthest distance is called aphelion (orange line). Credit: Allyson Sheffield

You can explore the properties of an ellipse using the simulation at the end of the chapter below.

Kepler's Second Law

Kepler's second law deals with the speed with which each planet moves along its ellipse, also known as its **orbital speed**. Working with Brahe's observations of Mars, Kepler discovered that the planet speeds up as it comes closer to the Sun and slows down as it pulls away from the Sun. Kepler's second law is called the law of equal areas and states that a planet will sweep out equal areas in an orbit in equal time intervals. Note that the planet does not sweep out equal angles in equal time intervals unless the orbit is a circle. In Figure 2 below, the triangular area B, defined by the Sun and points 1 and 2, has an area equal to the triangular area A, defined by the Sun and points 3 and 4.

Kepler's second law has important implications. Clearly the distance between points 3 and 4 is shorter than the distance between points 1 and 2 in Figure 2 below. Since the time interval is the same for these different distances, the planet is traveling at different velocities. A planet moves fastest when it is closest to the Sun (at perihelion) and slowest when it is furthest from the Sun (at aphelion).

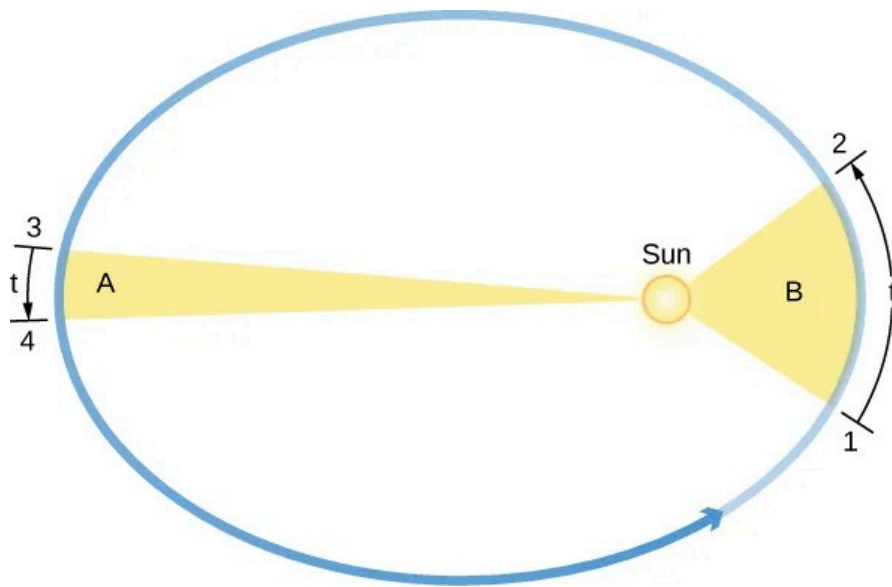


Figure 2 – Kepler's Second Law: a planets sweeps out "equal areas in equal times" as it orbits the Sun. The time for a planet to travel between points 1 & 2 and points 3 & 4 are the same, as are the areas of shapes A and B. Since the distance from point 1 to 2 is greater than the distance from point 3 to 4, the planet is moving fastest when it is closest to the Sun. Credit: [OpenStax Astronomy 2e](#), CC BY

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Kepler's Third Law

About a decade after Kepler had formulated his first two rules for the motion of planets in the solar system, he published a third law based on his empirical data, which says that the square of the orbital period is proportional to the cube of the semi-major axis:

$$P^2 \propto a^3 \quad (1.4.1)$$

where a is the semi-major axis of the elliptical orbit and P is the orbital period and the $\propto$ symbol is read as "proportional to". When P is measured in years, and a is expressed in astronomical units (AU), the two sides of the formula are not only proportional but equal:

$$P^2 = a^3 \quad (1.4.2)$$

This means that a planet that is closer to the Sun has a shorter orbital period than a planet that orbits farther away. Mercury completes one orbit in 88 days, Venus orbits in 225 days, and the Earth takes 365 days to travel around the Sun. Kepler's third law applies to all objects orbiting the Sun, including Earth, and provides a means for calculating their relative distances from the Sun from the time they take to orbit.

Worked Example: Using Kepler's Third Law

Example 1. Let's start by using Kepler's Third Law to find the orbital period of Jupiter. The semi-major axis of Jupiter's orbit around the Sun is 5.20 AU. Recall that the semi-major axis, a , is equivalent to the average distance of a planet from the Sun.

$$P^2 = a^3 \quad (1.4.3)$$

We can solve for P by taking the square root (which is the same as raising a number to the $1/2$ power) of each side of the equation: $(P^2)^{1/2} = (a^3)^{1/2}$

$$P = a^{3/2} \quad (1.4.4)$$

Putting in the value of a for Jupiter, and noting that if we have the semi-major axis in AU, then the orbital period will be in years:

$$P_J = a_J^{\frac{3}{2}} \quad (1.4.5)$$

Example 2. Now let's use Kepler's Third Law to find the average distance that Halley's Comet is from the Sun. It takes Halley's Comet 76 years to make one orbit around the Sun (and it last visited the solar neighborhood in 1986). The average distance of Halley's comet is found by solving Kepler's Third Law equation for a , this time by taking the cube root (or 1/3 power) of each side: $(a^3)^{\frac{1}{3}} = (P^2)^{\frac{1}{3}}$

$$a = P^{2/3} \quad (1.4.6)$$

This means the average distance of Halley's comet from the Sun is 17.9 AU. In reality, Halley's Comet has a very elliptical orbit, with an eccentricity of $e = 0.97$. You can easily carry out one more step to find the minimum and maximum distance (in this case, the perihelion and aphelion distances) of the comet from the Sun:

$$d_{\text{perihelion}} = a(1 - e) \text{ and } d_{\text{aphelion}} = a(1 + e)$$

$$d_{\text{perihelion}} = 17.9 \text{ AU } (1 - 0.97) = 0.537 \text{ AU}$$

$$d_{\text{aphelion}} = 17.9 \text{ AU } (1 + 0.97) = 35.3 \text{ AU}$$

Newton's *Principia*

Isaac Newton was born the same year that Galileo died: 1642. While he was a college student at Cambridge in the UK, the Bubonic plague swept through London. Newton worked from home during the quarantine (sound familiar?) and had an immensely productive time during his isolation in 1665-1666. Newton continued to work on his ideas in private, even inventing new mathematical tools, such as Calculus, to help him deal with the complexities involved. Eventually, his friend Edmund Halley prevailed on him to collect and publish the results of his remarkable investigations on motion and gravity. The result was a volume that set out the underlying system of the physical world, *Philosophiae Naturalis Principia Mathematica*. The *Principia*, as the book is generally known, was published at Halley's expense in 1687.

Newton's Laws of Motion are:

- **Newton's First Law (Law of inertia):** Every object will continue to be in a state of rest or move at a constant speed in a straight line unless it is compelled to change by an outside force.
- **Newton's Second Law:** The change of motion of a body is proportional to and in the direction of the force acting on it. This is often stated as the equation $F = ma$, or force equals mass times acceleration.
- **Newton's Third Law:** For every action, there is an equal and opposite reaction.

Here are some examples of each of Newton's laws of motion to help further elucidate their meaning in the context of astrobiology. Newton's first law means that an object that is coasting in space, far from the forces from any other objects, will continue coasting at the same speed. This is seen for the Voyager spacecraft. Voyager 1 and 2 were launched in 1977 and have travelled out past the edge of the Kuiper Belt. Unless their orbits take them close to another object, the spacecraft will both continue coasting at a constant speed (Voyager 1 is moving at about 38,000 miles per hour and Voyager 2 at 34,400 miles per hour).

In the [Space Flight chapter](#), we discuss sending a light sail out to visit the nearby (4 light years away) exoplanet Proxima b. After a force is initially applied to speed up the light sail, it will coast at a constant speed of about 100 million miles per hour (45,000 km/s) until it reaches the Proxima star system.

Newton's second law tells us how to speed up (or slow down or change the direction of motion of) an object: by applying a force. The rate of change in an object's velocity is called *acceleration*. Newton showed that the acceleration of a body was proportional to the force being applied to it. A larger force will lead to a larger acceleration, meaning a greater change in motion. For example, the larger the force applied to a light sail sent to another star, the faster the sail will be able to travel.

Newton's third law means that if you push on something, it pushes back. You pushing on something is the action, and the something pushing back with the exact same force is the reaction. An example of action-reaction forces is familiar to all who have batted a baseball (or shot a rifle). The recoil you feel as you swing your bat shows that the ball exerts a force on it during the impact, just as the bat does on the ball. A more pertinent example for astrobiology is rocket motion. The force that discharges the exhaust gases from the rear of the rocket is accompanied by the force that pushes the rocket forward; the exhaust gas pushed through the back is the action and the rocket moving forward is the reaction. The exhaust gases need not push against air or

Earth; a rocket actually operates best in a vacuum.

Newton's Universal Law of Gravitation

Kepler used Tycho's measurements of the planet positions over time to empirically derive the mathematical relationship captured in Kepler's third law. However, Kepler did not have a physical explanation for *why* the planets orbited the Sun. This is a good example of how science works. We make observations, try to fit the observations with a mathematical model, and then look for an elegant generalization that is both descriptive and predictive. If the mathematical generalization holds up under intense scrutiny, it is elevated to have the status of a theory.

About sixty years after Kepler published his third law for how the motion of planets in our solar system behaved, Isaac Newton helped us to understand the "why." Newton developed a mathematical description of gravity as a force between two objects that is proportional to the product of the object masses (M_1 and M_2) and inversely proportional to the square of the distance, d , between them measured from center to center; G is Newton's Universal Gravitational Constant which sets the strength of gravity throughout the universe and is equal to :

$$F_{grav} = G \frac{M_1 M_2}{d^2} \quad (1.4.7)$$

This is Newton's Universal Law of Gravitation and it tells us that the the closer together the objects are (that is, the smaller the value of d), the stronger the force between the two will be. It also tells us that the more mass an object has, the stronger the pull of its gravitational force.

We make an important note here that Newtonian Gravity, as we are discussing it here, is not the full story. Albert Einstein showed that mass itself creates gravity by warping the spacetime around it.

Revisiting Kepler's Third Law

Newton's law of gravitation shows that mass and gravity are directly related to each other. When Newton reanalyzed Kepler's Third Law through the lens of gravity, he found that the mass of the star and the planet also appear in the relationship as:

$$(M_1 + M_2)P^2 = a^3 \quad (1.4.8)$$

Here, M_1 and M_2 are the masses of the two objects that are in orbit, for example a planet orbiting the Sun or a moon orbiting a planet. Technically, both objects are orbiting what is know as the center-of-mass between them, but we will visit this concept in more detail when discussing methods for detecting exoplanets.

The mass of a star is vastly greater than the mass of a planet (or even more so for a comet) so we can ignore the mass of the smaller object and use only the mass of the star. This same logic applies if a moon is orbiting a planet and the planet's mass is much greater than the moon's mass, which is the case for Earth. We can thus eliminate the mass of the much lighter object and write the relationship as:

$$MP^2 = a^3 \quad (1.4.9)$$

where M is the mass of the much heavier object. To use Newton's version of Kepler's Third Law as shown above, the semi-major axis a needs to be in AU, the orbital period P in years, and the mass must be reported in solar masses (where the Sun has a value of 1 and a star twice as massive as the Sun has the value 2). We will look at some examples in the box below to show how fundamental Newton's version of Kepler's Third Law is for understanding other solar systems.

Worked Example: Using Newton's Version of Kepler's Third Law

Example 1. Consider a planet orbiting a star that is one-half the mass of the Sun ($M_{\text{star}} = 0.5 M_{\text{Sun}}$). The planet takes 730.5 days to make one orbit around the star. What is the semi-major axis of the planet's orbit?

First, we need to convert the units for the orbital period into years:

$$(730.5 \text{ days}) \times (1 \text{ year}/365.25 \text{ days}) = 2 \text{ years}$$

Now, we can use this value, along with the mass of the star, to find a ; first, we solve for a by taking the cube root of both sides:

$$(a^3)^{\frac{1}{3}} = (MP^2)^{\frac{1}{3}} \quad (1.4.10)$$

$$a = (M \times P^2)^{\frac{1}{3}} = (0.5 \times 2^2)^{\frac{1}{3}} = (0.5 \times 4)^{\frac{1}{3}} = 2^{\frac{1}{3}} = 1.26 \quad (1.4.11)$$

Example 2. Now let's turn the problem around and find the orbital period for a planet if we are given the semi-major axis of the planet's orbit and the mass of the star. If the planet is orbiting with a semi-major axis a of 4.0 AU around a star that is three-times more massive than the Sun, how long does it take this planet to make one orbit around the star?

This time, let's solve Newton's version of Kepler's Third Law for P :

$$(P^2)^{\frac{1}{2}} = (a^3/M)^{\frac{1}{2}} \quad (1.4.12)$$

$$P = (a^3/M)^{\frac{1}{2}} = (4^3/3)^{\frac{1}{2}} = (64/3)^{\frac{1}{2}} = 4.62 \quad (1.4.13)$$

Learn More About: Orbital and Escape Speed

Newton's Law of gravity can be combined with his laws of motion to derive the speed of a planet as it orbits a star or, more generally, the speed of any object revolving around another object.

For an object in a circular orbit, the **orbital speed** is given by:

$$v_{orbit} = \sqrt{\frac{GM}{R}} \quad (1.4.14)$$

where M is the mass of the body about which the object orbits, R is the radius of the orbit (how far away it is from the center) and G is the universal gravitation constant. This is the speed needed to launch an object such as a satellite into orbit around the Earth (v_c in Figure 3a) or the average speed of a planet orbiting a star.

What if we want the object to leave (escape) Earth's orbit? The equation for the **escape speed** is very similar, but now the speed is higher by a factor of $\sqrt{2}$:

$$v_{escape} = \sqrt{\frac{2GM}{R}} \quad (1.4.15)$$

It makes sense that a higher speed, meaning more force, is needed to get an object out of orbit. What about speeds that are in between those for a circular orbit and the escape speed? These will result in bound elliptical orbits for the object; at the escape speed, the object becomes unbound from the Earth's gravity.

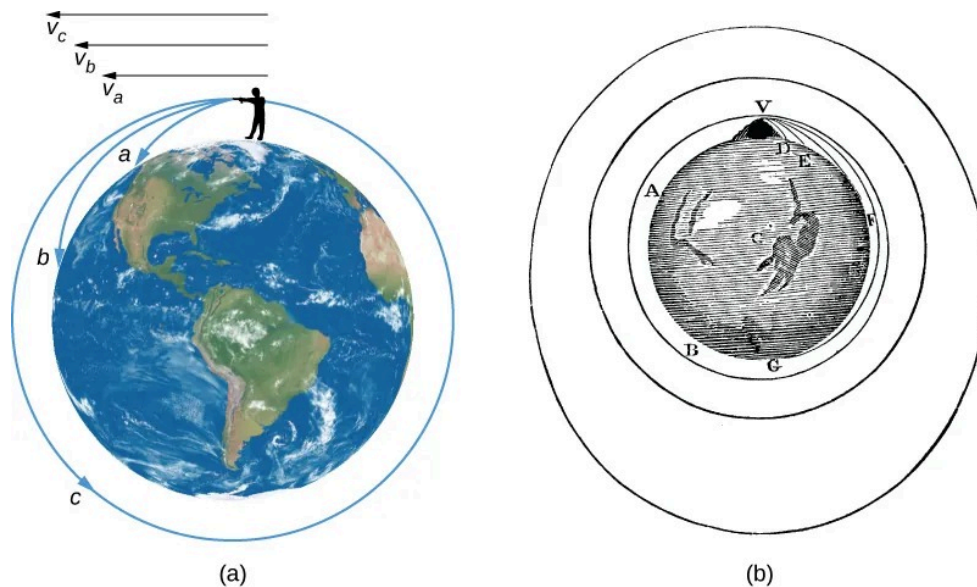


Figure 3 – Firing a bullet into orbit. (a) For paths a and b, the velocity is not enough to prevent gravity from pulling the bullet back to Earth; in case c, the velocity allows the bullet to fall completely around Earth. (b) This diagram by Newton in his *De Mundi Systemate*, 1731 edition, illustrates the same concept shown in (a). Credit: [OpenStax Astronomy 2e](#), CC BY

The equation above for the orbital speed of an object is specifically for a circular orbit. Planetary orbits around the Sun are not circular but are elliptical, so we must keep in mind that the actual speed of the planet varies as it revolves around the Sun. For a nearly circular orbit like the Earth, with an eccentricity e of 0.017, the deviation from the circular orbit is only a few km/s, but you can imagine that this breaks down for a highly eccentric object like that for Halley's comet. An exact speed can be found for elliptical orbits by taking the eccentricity into account.

One final note is that any satellite orbiting another object — be it a weather satellite orbiting the Earth, a moon orbiting a planet, or a planet orbiting a star — obey these laws. Kepler's and Newton's laws are universal.

Free Fall, Weight and Surface Gravity

Free Fall and Weightlessness

An object that only has the force of gravity acting upon it is in free fall. This can be a confusing concept, as it certainly looks like the astronauts on the International Space Station (ISS) are in an environment with no gravity, as they are floating around — but there is in fact gravity at the height that the ISS orbits. What is happening? The astronauts are experiencing the sensation of weightlessness; you may have felt this sensation if you have been on a ride at an amusement park that lifts you up out of your seat momentarily. But the astronauts still have weight, just as you still do on a roller coaster ride. With the astronauts on the ISS, they are in free fall *around* the Earth, not *to* Earth. As the ISS is pulled to the Earth's surface due to gravity, the motion (speed) of the ISS is fast enough to prevent it from actually falling to the Earth...and it stays in orbit. This is shown in Figure 3(a) above.

Mass, Weight and Surface Gravity

You may have heard that you weigh less on the Moon than you do on Earth. This is true! How can that be? The key here is that mass and weight are not the same. Your mass (measured in kilograms, usually) is the same no matter where you are; your weight depends on the surface gravity of where the scale you are standing on is located. And the Moon has a lower surface gravity than the Earth.

Using Newton's laws of motion and gravity, we can show that the surface gravity g on a world is:

$$g = \frac{GM}{R^2} \quad (1.4.16)$$

where M is the mass of the world and R is its radius. We can look up the mass and radius of the Earth ($M_{\text{Earth}} = 5.97 \times 10^{24}$ kg;

$R_{\text{Earth}} = 6.38 \times 10^6 \text{ m}$) and the value of the constant G ($6.67 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$) and find that:

$$g_{\text{Earth}} = 9.8 \text{ m/s}^2$$

Note that the units for the radius and mass need to be consistent with those for the constant G for the equation to give the correct result.

We are interested in knowing the surface gravity of planets orbiting other stars to assess the habitability of the planet. We can find this quickly if we know the mass and radius relative to Earth, as shown in the example below.

Worked Example: Surface Gravity of an Exoplanet

The exoplanet Proxima b is of interest to astrobiologists as it is orbits in the “habitable zone” around its star. Proxima b has a mass that is 1.07 times that of the Earth ($M_{\text{Prox}} = 1.07 M_{\text{Earth}}$) and a radius that is 1.03 times the radius of the Earth ($R_{\text{Prox}} = 1.03 R_{\text{Earth}}$). Note that this means that Proxima is almost the same size as the Earth but has a slightly larger mass (7% larger than the Earth’s mass) and radius (3% greater than the Earth’s radius). If we take the ratio of the surface gravity of the Earth to that of Proxima b, the gravitational constant G cancels out and we are left with the following ratio:

$$\frac{g_P}{g_E} = \frac{GM_P/R_P^2}{GM_E/R_E^2} = \frac{M_P}{M_E} \sqrt{\frac{R_P}{R_E}} \quad (1.4.17)$$

$$g_P = \frac{1.07 M_E}{M_E} \sqrt{\frac{1.03 R_E}{R_E}} g_E = 1.07 \sqrt{1.03} g_E = 1.08 g_E \quad (1.4.18)$$

The surface gravity on Proxima b is very close to that on the Earth’s surface and is 8% higher. In other words: $g_P = (1.08) g_{\text{Earth}} = 1.08 (9.8 \text{ m/s}^2) = 10.6 \text{ m/s}^2$

Data source: https://exoplanets.nasa.gov/eyes-on-exoplanets/#/planet/Proxima_Cen_b/

Once you know the surface gravity of a world, you can easily calculate how much you would weigh if you were standing on a scale on its surface. From Newton’s second law, we know that force is equal to mass times acceleration. Standing on the surface, your acceleration is that due to gravity:

$$\text{Weight} = (\text{mass}) \times (\text{surface gravity}) = mg$$

Worked Example: Weight

The surface gravity on the Moon is 1.62 m/s^2 , which is about one-sixth the value on Earth’s surface. How much would you weigh if you stepped on a scale on the Moon?

We can use the relationship between weight, mass and surface gravity to solve this. Let’s assume a weight of 100 kg. We can quickly show the weight on Earth’s surface for a person with this mass:

$$W_{\text{Earth}} = mg_{\text{Earth}} = (100 \text{ kg}) \times (9.8 \text{ m/s}^2) = 980 \text{ newtons} = 980 \text{ N}$$

Notice that the units of weight are in newtons. This is a unit of force, named in honor of Sir Isaac Newton, that is abbreviated as N. In the U.S., many scales show weight in pounds, and 1 pound is equal to 4.45 N. Therefore,

$$(980 \text{ N}) \times (1 \text{ pound}/4.45 \text{ N}) = 220 \text{ pounds} = 220 \text{ lbs}$$

Now, let’s do this calculation for the Moon:

$$W_{\text{Moon}} = mg_{\text{Moon}} = (100 \text{ kg}) \times (1.62 \text{ m/s}^2) = 162 \text{ N}$$

$$(162 \text{ N}) \times (1 \text{ pound}/4.45 \text{ N}) = 36.4 \text{ pounds} = 36.4 \text{ lbs}$$

Another way to solve this problem is to take the ratio of your weight on the Moon to your weight on Earth; doing this leaves the simple formula:

$$W_{\text{Moon}} = (g_{\text{Moon}}/g_{\text{Earth}}) \times W_{\text{Earth}}$$

$$W_{\text{Moon}} = (1.62 \text{ m/s}^2 / 9.8 \text{ m/s}^2) \times 220 \text{ lbs}$$

$$W_{\text{Moon}} = 0.1653 \times 220 \text{ lbs} = 36.4 \text{ lbs}$$

This way of solving the problem avoids the need to convert the units for the weight.

Calculate the weight for this same person (with a mass of 100 kg) on the surface of Proxima b.

Show Answer

We found in the previous example that the surface gravity on Proxima b is 10.6 m/s^2 . Using this value:

$$W_{\text{Proxb}} = mg_{\text{Proxb}} = (100 \text{ kg}) \times (10.6 \text{ m/s}^2) = 1060 \text{ N}$$

$$(1060 \text{ N}) \times (1 \text{ pound} / 4.45 \text{ N}) = 238 \text{ pounds} = 238 \text{ lbs}$$

General Relativity

In 1915, Einstein published his General Theory of Relativity, treating gravity as a warping of the space-time continuum and an attribute of mass rather than a fundamental force. Einstein's description of gravity employs mathematically sophisticated metric tensors, while Newtonian gravity uses more commonly accessible algebra or calculus. On the scales that humans are accustomed to working with (and certainly for the physics included in this course), general relativity converges to the Newtonian description, so we are justified in defaulting to the Newtonian description of gravity.

Key Concepts and Summary

Galileo demonstrated that heavier, more massive objects fall at the same rate as light objects. This can seem counter-intuitive because light objects often experience more wind resistance than heavier ones, but was beautifully demonstrated by astronauts on the airless moon. Kepler used data from his mentor, Tycho Brahe, to demonstrate three laws of planetary motion: planetary orbits are ellipses (not circles), the speed of planets was fastest during the closest approach to the Sun, the orbital distance (a) and orbital period (P) are mathematically related by $P^2 = a^3$. Isaac Newton developed the branch of mathematics known as calculus to describe gravitational forces and showed that this matched Kepler's Third Law. Although gravity is now understood to be a property of mass that is more completely described by Einstein's General Theory of Relativity, Kepler's Law and Newton's Law of Gravity are accurate descriptions for most observations of orbital motion.

Review Questions

Summary Questions

1. How did Galileo show that objects all fall at the same rate, regardless of their mass?
2. Summarize Newton's Laws of Motion and give a simple example that illustrates each law.
3. Explain each of Kepler's Laws of Planetary Motion.
4. What does Kepler's Second Law say about the speed of a planet as it orbits the Sun?
5. Explain in words Kepler's Third Law. Give an example of a problem we can solve using Kepler's Third Law. Now solve the problem!
6. Explain Newtonian gravity and how gravity depends on mass.
7. What does it mean to say an object is "in free fall"?
8. What is the difference between mass and weight? Do your mass and weight change on the Moon? Explain why or why not.

9. Explain what the surface gravity of a world means. Give an example of two planets or Moon's with different surface gravities and explain how they differ.
10. Briefly explain why we do not need to use Einstein's laws of relativity to study the orbits of planets.

Exercises

1. **Planetary Orbit Simulator.** Open the simulator [here](#).
 - The default is set to the planet Mercury. Click on "show center". How close is the Sun to the center? Now switch to planet Earth. How close is the Sun to the center? Note the difference in eccentricities for Mercury and Earth.
 - Click on the "Kepler's Second Law" tab.

Gravity and Orbits. Open the Gravity and Orbits simulation [here](#).

- To start, select the **To Scale** option in the simulation as we would like to have an experiment that is as close as possible to the way things are in reality. This option adjusts the physical sizes of objects so that they take up the correct amount of space relative to their orbital motions. Because the scale of orbits is so much larger than the physical size of the stars, planets, moons, and satellites, in the **To Scale** option, you won't be able to see many of the objects beyond tiny points.
- Enable the "Mass", "Path", and "Grid" options. You will notice that the system we are simulating is the Earth-Sun system. Note the position of the Earth at the start of the simulation. It is at the intersection of a particular set of grid points.
- Now start to run the simulation. You will see the planet (Earth) begin to orbit the star (the Sun). Try to time it so that you can determine the amount of time it takes for the planet to complete one revolution around the Sun. You should note that it takes about 365 days for the planet to return to that initial grid point. This means the simulation is working since that is indeed the amount of time it should take a planet that is the distance from the Sun of one astronomical unit (AU) to make a complete orbit!
- If the Earth suddenly gained mass, what do you think would happen to the length of a year? Run the simulation for a more massive Earth to see if your intuition was correct.
- What if the Earth suddenly lost mass?
- If the Sun suddenly gained mass, what do you think would happen to the length of a year?

2: THE CHEMISTRY OF LIFE

2.1: Introduction to Atoms and Elements

2.2: Basic Chemistry of Life

2.3: Key Biochemicals- Amino Acids, Nucleic Acids, and the Genetic Code

2.4: Alternative Biochemistries- Speculations about Strange Possibilities

2.1: Introduction to Atoms and Elements

All matter, life included, is made up of atoms. Here we discuss the inner workings of atoms. This will be relevant as we work to understand the chemistry and biology of life.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe atoms and how they relate to chemical elements
- Understand the properties of the subatomic particles electrons, protons, and neutrons
- Explain what ions and isotopes of an atom are
- Describe the cosmic abundance of different elements

Atomic Theory

Atomic theory provides a microscopic explanation of the many macroscopic properties of matter that you can directly test yourself in a chemistry laboratory. Every piece of matter we interact with, whether living or non-living, is made out of atoms.

The word *atom* comes from the Greek word for “indivisible.” This naming is based on an idea held by some ancient philosophers—the **Atomists**, as discussed in the History of Astrobiology chapter—that all matter is composed of small, finite particles that differ in shape and size and that join together in different ways.

A modern understanding of atomic theory began in the early 1800s with John Dalton, an English schoolteacher. Through observations and experiments, Dalton demonstrated that atoms exist and correspond to different elements. Pure elements consist of only one type of atom, which has certain properties characteristic of that element. As an example, Figure 1 shows a pure copper penny, which means the penny is entirely made up of copper atoms that are linked or bonded together.

Because an element consists of only one kind of atom, an element cannot be broken down into a different substance. All atoms of a given element have identical chemical properties. Inversely, atoms of a given element differ in properties from atoms of all other elements.



Figure 1 – Copper atoms in a penny. A pre-1982 copper penny (left) contains approximately 3×10^{22} copper atoms (several dozen are represented as brown spheres at the right), each of which has the same chemical properties.

An atom is the smallest unit of an element that can participate in a chemical reaction. Atoms are neither created nor destroyed during a chemical reaction; they are instead rearranged to yield substances that are different from the substances present before the reaction (Figure 2). This property is captured in the law of conservation of matter. Because the number and nature of atoms

remains constant before and after a chemical reaction (it is only the configuration of atoms that changes), the total mass of matter before and after a chemical reaction also remains constant.

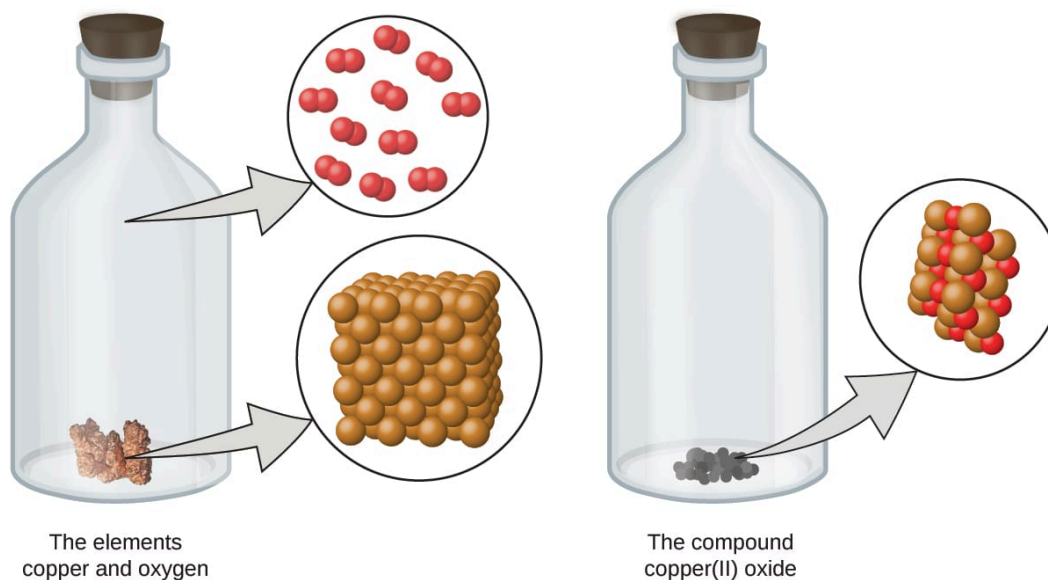


Figure 2 – Atoms are rearranged in a chemical reaction. In this example, we start on the left with copper atoms (shown here as brown spheres) and oxygen (shown here as red spheres). These atoms react and rearrange (right) to form a compound containing copper and oxygen (a powdery, black solid).

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Going Subatomic

After establishing the existence of atoms, the inquiring mind is tempted to explore deeper: what are atoms themselves made of?

Discovery of Subatomic Particles

The first evidence of subatomic particles came in the late 1800s when different rays, meaning beams of light or radiation, were used to probe the structure of atoms. Rays differ in energy level and source. Some rays were found to emanate from certain elements, which we now call radioactive.

In 1897, the English physicist J. J. Thomson was the first to discover a subatomic particle from investigations of a type of ray called cathode rays. Thomson found that cathode rays are made up of actual particles that are much smaller than atoms. Experiments showed that these particles could be extracted from any atom, and that they carried a discrete amount of negative electric charge. He named these particles electrons.

Soon thereafter, Marie Curie discovered radium, a radioactive element, in 1898. Curie's pioneering work in developing the theory of radiation was recognized by a Nobel Prize in Physics, and her work discovering the radioactive elements polonium and radium was awarded a Nobel Prize in Chemistry. Marie Curie was the first woman to be awarded a Nobel Prize, and she is one of only four people (as well as the only woman) to win the Nobel Prize twice.

Ernest Rutherford, a physicist from New Zealand who largely spent his scientific career in Canada and England, showed that radium can be used to produce alpha (α) rays, beams of tiny, high-speed, positively charged particles called alpha particles. Between 1908 and 1913, Rutherford directed a series of experiments using alpha rays in collaboration with, Hans Geiger, later famous for the Geiger counter, and Ernest Marsden, an undergraduate student at the time.

Geiger and Marsden aimed alpha rays at a very thin foil of gold and other various metals. They found that alpha particles did one of three things: (1) most of the alpha particles passed right through the foil, (2) some particles were diverted slightly, and (3) a very small number of particles were deflected almost straight back towards the source of the alpha rays. Rutherford described the behavior of the small number of particles being deflected back as follows:

It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you¹

Let us examine what can be learned by the three different behaviors exhibited by alpha particles in this experiment. Undeflected alpha particles were not affected by the foil at all, so they must have travelled through essentially empty space inside the atom. Because most particles passed right through, this suggests that an atom largely consists of empty space.

We know that alpha particles are positively charged, and that like charges repel each other (like how the ends of magnets that have the same charge will repel each other rather than attract). The deflected alpha particles therefore must have encountered another positive charge within an atom that repelled the alpha particles, causing their paths to be diverted or deflecting the particles straight back towards the source. The smaller number of diverted and deflected particles suggests the existence of a nucleus, a small, positively charged body at the center of each atom.

These experiments led to an understanding of the atom as a small, positively charged nucleus, in which most of the mass of the atom is concentrated, surrounded by negatively charged electrons. Because atoms are electrically neutral, the positive charge of the nucleus must be balanced by the negative charge of electrons.

After many more experiments, Rutherford established that the nuclei (the plural of nucleus) of all other elements contain the nucleus of a hydrogen atom. He named this more fundamental “building block” particle a proton to refer to the positively charged, subatomic particle found in the nucleus of atoms. We now know that the nucleus of hydrogen atoms consist of just a single proton.

The nucleus was known to contain almost all the mass of an atom, but it was found that the number of protons accounted for only half, or less, of the mass of the nucleus. It was proposed that there existed neutral particles in the nucleus as well as the positively charged protons. However, detecting uncharged particles is very challenging as they do not interact with charged particles. It was not until 1932 that James Chadwick, a British physicist, found evidence of these neutral particles now called neutrons. Neutrons are uncharged, subatomic particles found in the nuclei of atoms with approximately the same mass as a proton.

Atomic Structure

We now understand that each atom is composed of a very small nucleus, made up of positively charged protons and uncharged neutrons, and that this nucleus is surrounded by a much larger volume of space in which negatively charged electrons exist.

The nucleus contains the majority of an atom’s mass because protons and neutrons are much heavier than electrons, whereas electrons occupy almost all of an atom’s volume. The diameter of an atom is on the order of 10^{-10} m, whereas the diameter of the nucleus is roughly 10^{-15} m—about 100,000 times smaller. To contextualize this, if an atom’s nucleus was the size of a blueberry, the diameter of the atom would be roughly the size of a football stadium (Figure 3).

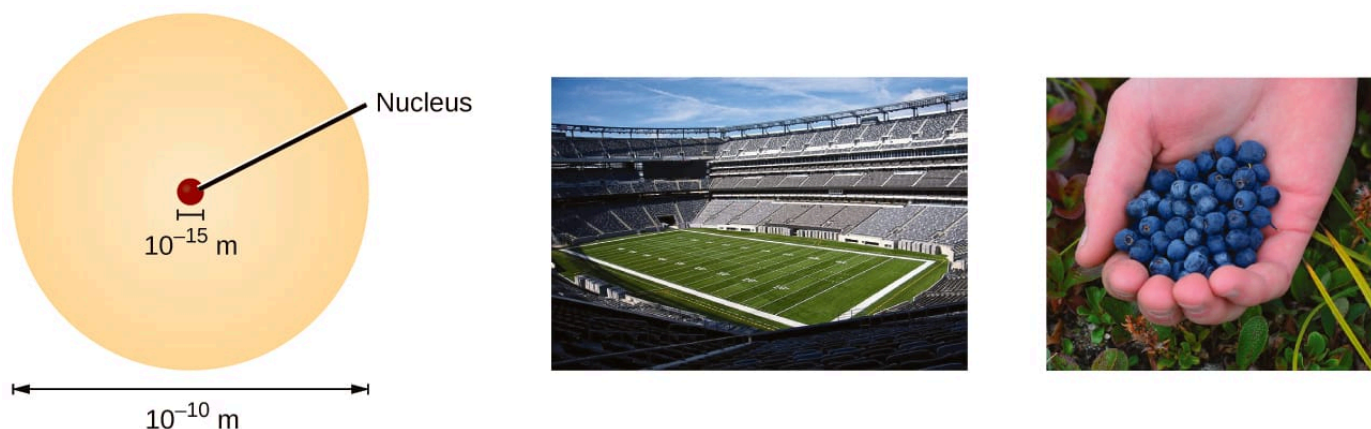


Figure 3 – A scaling of the size of an atomic nucleus. If the size of an atom corresponded to the size of a football stadium, then the nucleus would be the size of a single blueberry.

Properties of Electrons, Protons, and Neutrons

Atoms—and the protons, neutrons, and electrons that compose them—are extremely small. For example, mass may typically be measured in terms of grams (g); a classic M&M is about 1 g. The mass of a single carbon atom is less than 2×10^{-23} g. An electron is over a thousand times less massive, with a mass on the order of 1×10^{-26} g. Protons and electrons also carry electric charge in very small amounts that are unwieldy to express using more typical units.

When describing the properties of atoms and their subatomic particles, we therefore make use of appropriately small units of measure. Rather than grams, we use the atomic mass unit (amu), which is approximately the mass of one proton. For electric charge, we use the fundamental unit of charge (e), which is equivalent to the electric charge on a single proton. In other words, a proton has a mass of ~1 amu and a charge of +1 e. Table 1 summarizes the properties of these subatomic particles.

Table 1 – Properties of Subatomic Particles

Name	Location	Charge (e)	Mass (amu)	Mass (g)
electron	outside nucleus	-1	0.00055	0.00091×10^{-24}
proton	nucleus	+1	1.00727	1.67262×10^{-24}
neutron	nucleus	0	1.00866	1.67493×10^{-24}

Characterizing Atoms

Recall that all atoms of a given element have identical and unique chemical properties. The chemical properties of an atom are a function of the subatomic particles contained with that atom.

A defining trait of an element is its atomic number (Z), the amount of protons in the nucleus of an atom of that element. The atomic number determines the identity of the atom. For example, helium has an atomic number of 2. All helium atoms therefore have two protons, and all atoms with two protons are helium atoms. This is true regardless of how many neutrons and electrons an atom has.

For neutral atoms, atoms that carry no electric charge, the positive charge from protons must be balanced out by the negative charge of electrons. For example, we know that a helium atom contains two protons. A neutral helium atom must therefore also contain two electrons. For neutral atoms, this means that the atomic number also corresponds to the number of electrons in the atom.

An atom's mass number (A) gives the total number of protons and neutrons in the atom. Recall that an element is defined by its atomic number (Z), the number of protons in its atoms nuclei, regardless of the number of neutrons. We can use the mass number to calculate the number of neutrons in an atom as the difference between the mass number and the atomic number, i.e. $A - Z = \text{the number of neutrons}$. For example, if a helium atom, which has two protons, has a mass number of 4, we can calculate that this helium atom has 2 neutrons ($4-2$).

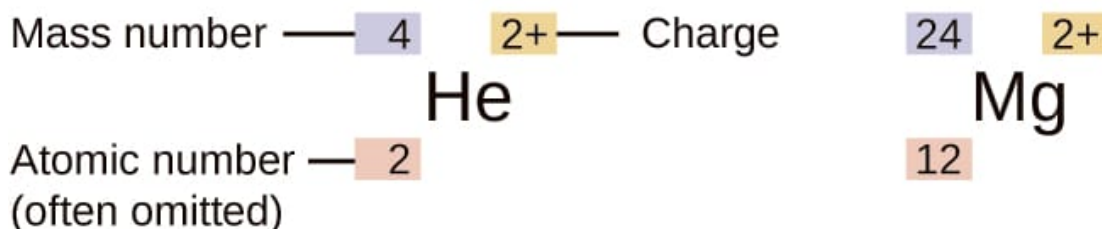


Figure 4 — Examples of a helium (He) and magnesium (Mg) atom characterized by its element's two-letter symbol, the mass number as a left superscript, the atomic number as a left subscript (sometimes omitted), and the charge as a right superscript.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Ions

Recall that neutral atoms contain the same number of positively charged protons and negatively charged electrons; the charges from the protons and electrons cancel each other out so that the atom has a net zero electric charge. An ion is an atom that does carry an electric charge because there are either more electrons or less electrons than there are protons. The electric charge of an atom is defined as follows:

$$\text{Atomic charge} = \text{number of protons} - \text{number of electrons}$$

Atoms typically acquire charge by gaining or losing electrons. An atom that gains one or more electrons will carry a negative charge. These negatively charged ions are called anions. Positively charged atoms, called cations, have lost one or more electrons.

As an example, consider the helium (He) atom characterized in Figure 4. Recall that helium has an atomic number of 2, which indicates that all helium atoms have two protons. A neutral helium atom therefore has two electrons. If a helium atom gains one electron, it will have three electrons total (2+1) and become an anion with a charge of -1 e. If a helium atom loses one electron, it will have one electron left (2-1) and become a cation with a charge of +1 e. The helium atom in Figure 4 has a charge of +2 e, meaning it has lost both of its electrons.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Isotopes

The number of protons in an atom defines the type of element it is; if the number of protons in an atom changes, the type of atom it is also changes. The number of electrons can change and defines the charge of the atom. What about the number of neutrons?

During the early 1900s, it was discovered that an element could have atoms with different masses. Despite the different masses of these atoms, they were otherwise indistinguishable from one another. The difference in masses corresponded more closely to the mass of neutrons and protons, which have similar masses, than the mass of electrons, which are less massive by about a factor of 2,000 (see Table 1). It was therefore deduced that these atoms have different masses because they possess different numbers of neutrons.

Such atoms are called isotopes—atoms of the same element that have differing amounts of neutrons.

An isotope of an element is specified by the mass number (A) written as a superscript to the left of the element symbol (see Figure 4). For example, magnesium naturally occurs in the form of three different isotopes that have mass numbers of 24, 25, and 26. These isotopes would be written as ^{24}Mg , ^{25}Mg , and ^{26}Mg respectively. These isotope symbols are read “element, mass number.” For instance, ^{24}Mg is read as “magnesium 24,” and can be written as “magnesium-24” or “Mg-24.”

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Unstable Isotopes

Not all isotopes of an element are stable. Unstable elements undergo a process known as *radioactive decay*, in which the subatomic particles in an atom change. These unstable elements are described as radioactive. As unstable nuclei decay, they often change from one isotope into another and sometimes into different elements depending on whether the number of neutrons or protons changes in the nucleus. The rate of radioactive decay is commonly characterized by *half-life*, the amount of time in which approximately half the number of atoms will have decayed.

Let's consider as an example uranium-238, which can also be written $^{238}_{92}\text{U}$, meaning it has 92 protons and $238-92=146$ neutrons. Uranium-238 has a half-life of 4.5 billion years. If we had a sample of 12 uranium-238 atoms, then after 4.5 billion years we expect that 6 of those atoms will have undergone radioactive decay. In other words, over 4.5 billion years, each uranium-238 atom has a half/half or 50% chance of decaying.

Concept Check

We encountered alpha particles earlier in this chapter. How did Ernest Rutherford, Hans Geiger, and Ernest Marsden use alpha particles to understand the components of the atom?

When uranium-238 decays, it emits an alpha particle; an alpha particle consists of two protons and two neutrons. Because it gives rise to an alpha particle, this process is known as alpha decay and can produce alpha rays.

Recall that uranium-238 has 92 protons and $238 - 92 = 146$ neutrons. With the emission of an alpha particle, a uranium-238 atom loses two protons and two neutrons. This decay results in an atom with 90 protons and 144 neutrons. Because the number of protons has changed, the nature of the atom has also changed. An atom with 90 protons, and therefore an atomic number of 90, is a thorium (Th) atom. The mass number of this atom is the combined number of protons and neutrons it has, $90 + 144 = 234$. Therefore, we see that when undergoing radioactive decay, a uranium-238 atom has decayed into a thorium-234 atom.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The process of radioactive decay is happening around us all the time. If a radioactive nucleus decays into another radioactive nucleus, then another decay can happen. This chain of reactions is called a “decay chain”. Eventually, the resulting nucleus may be stable and no longer decay. Most of the nuclei in the natural world appear to be stable, but there is a possibility that all nuclei will eventually decay if given enough time. We do not yet know if “stable” nuclei are just extremely long lived.

Cosmic Abundance of Elements

After the Big Bang, there were only a few elements in existence: hydrogen, helium, and a sprinkling of lithium. As described in the section “[Assembling the Periodic Table](#),” we know that the majority of the remaining elements, including most of those which make up the Earth, were formed at the cores of stars. These elements were then distributed throughout the cosmos through their violent and explosive deaths.

We have used spectroscopy to identify the chemical composition of other stars, giant molecular clouds, and even the gas around and between galaxies. These measurements allow us to calculate the abundance of various elements and isotopes. We see that the universal abundance of elements match what is expected from the processes that create them.

These observations of abundances throughout the universe suggest that the laws governing chemistry and physics that we observe on Earth also operate the same way on any other planet at any other location in our galaxy and beyond. Therefore, if we can understand the chemical origins of life on Earth and in our solar system, we would gain insights into when, where, and how life might arise on other worlds.

Figure 5 shows the relative abundances of different elements in the Solar System. What are the most common elements?

Figure 5 – Relative abundance of elements in the Solar System. Note that this is a log scale so an element that is one major tick mark higher than another is ten times more abundant. Thus, hydrogen and helium make up more than 99% of all atoms in the cosmos as they are more than two major tickmarks above the next most abundant elements: oxygen and carbon.

Concept Check

All life on Earth is based on four elements: hydrogen (H), carbon (C), nitrogen (N) and oxygen (O). How common are each of these elements in the solar system, according to Figure 5?

Why might life not use the element helium (He) for life? Explain your answer!

Hydrogen, carbon, nitrogen, and oxygen form the basis of life, but how do we move from single atoms to complicated living things? To take the first step, we consider the chemistry of how atoms bond together to form molecules.

Key Concepts and Summary

Atoms are fundamental units in chemistry that explain the Periodic Table as a sequence with increasing numbers of protons. The nucleus of an atom is comprised of positively charged protons and uncharged neutrons. It is the number of protons in the nucleus

that defines the element – isotopes of elements have different numbers of neutrons. In an electrically neutral atom, the number of electrons equals the number of protons. After the Big Bang, the only elements that emerged were hydrogen, helium, and traces of lithium. All other elements were forged in the cores of stars or during highly energetic explosions. The elements in the universe today are generally decreasing in abundance with increasing atomic mass (i.e., highest atomic mass elements are rarest), but this trend is modulated by even-odd atomic numbers, a signature of the fusion of alpha particles.

Review Questions

Summary Questions

1. What is an atom?
2. How are atoms relate to chemical elements?
3. What is the difference between an atom and a molecule?
4. What three “subatomic” particles are found inside an atom? Describe the similarities and differences between them.
5. How are different isotopes of a particular element distinguished? What are some reasons that certain isotopes are more common than others?
6. Which elements are the most common and which elements are the least common in our universe? How does astrophysics help explain this?

Activities

1. **Building atoms.** Open the “Build an atom” simulator (<https://phet.colorado.edu/en/simulations/build-an-atom>) and select the Atom option. Build the following:
 1. hydrogen, deuterium, tritium
 2. lithium (is this atom stable?)
 3. carbon, carbon-12, carbon-13

2.2: Basic Chemistry of Life

The way that atoms behave and interact is defined by their subatomic particles, and in particular the number of electrons in an atom. Here we discuss how and why atoms interact to form molecules as well as how to represent these molecules. We conclude with a deep dive on the elements most commonly found in life.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain how electrons are organized within an atom, and how this relates to chemical bonding and behavior
- Describe the different types of chemical bonds that hold molecules and compounds together
- Understand the different ways molecules are represented
- Understand the most common elements found in biological reactions

Chemical Bonds

Recall that the subatomic particles within an atom, the protons, neutrons, and electrons, change the nature and behavior of an atom. Every element has a fixed number of protons (for example, hydrogen has 1 proton and carbon has 6 protons), but the number of electrons and neutrons can vary. For a quick refresher, take a look at this [simulation on building atoms](#). The number of neutrons changes the isotope of an atom and can affect the stability of the atom. The number of negatively-charged electrons in an atom can change the charge of the atom. Here, we discuss how the number of electrons in an atom also controls the number and types of chemical bonds an atom can form.

Electron Shells

We can approximate the position of electrons inside an atom as being organized into different orbital shells surrounding the nucleus, where each orbital shell represents a different energy level. In reality, it is impossible to know the exact location of the electrons in an atom as they orbit the nucleus. As such, we make use of the concept of orbital shells that represent different energy levels as a useful approximation to understand how the number and state of electrons affect the behavior of atoms.

Figure 1 shows a graphical representation of electrons organized into orbital shells around neutral hydrogen (H), carbon (C), nitrogen (N), and oxygen (O) atoms, as represented by the chemical symbol for each element. Recall that for a neutral atom, the number of electrons matches the number of protons. Black, co-centric circles represent the orbital shells for each atom; one orbital shell is shown for hydrogen while two are shown for the other three elements. The smaller, innermost circle represents the lowest energy level. The outer circle represents a higher-energy level.

Only a certain number of electrons can exist at each energy level, which places a limit on the number of electrons that can occupy each shell. The innermost and lowest energy shell can contain only two electrons. The next energy level up can contain eight electrons.

The outermost orbital shell is known as the valence shell. Electrons in this shell are called valence electrons. It is in this valence shell where bonding between atoms most typically occurs via interactions between the atoms' valence electrons. The chemistry of molecular bonds is driven by the stability of an atom with a filled valence electron shell. Atoms form chemical bonds so that they can completely fill their valence electron shell.

For example, in Figure 1 the pictured single carbon atom has four electrons in its outer valence shell. We know that this second orbital shell can hold up to eight electrons. A carbon atom with a valence shell completely filled with eight electrons is more stable, and therefore more chemically favored, than a valence shell with only four electrons.

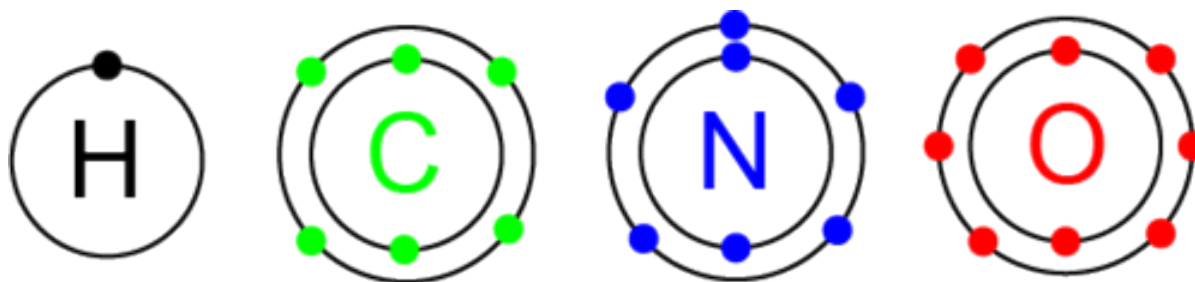


Figure 1 – The electron structure of the four most common elements in life: (from left to right) hydrogen (H), carbon (C), nitrogen (N), and oxygen (O). Large, black circles represent orbital shells containing electrons, which are represented by colored dots. How many valence electrons does each element have? Do you think that electrons could move up to higher (empty) energy levels? If so, how?

Atoms will form chemical bonds to gain access to valence electrons that will fill their valence shells. When more than one atom bonds together, they form a molecule. These molecular bonds are typically divided into two categories: *ionic bonds* and *covalent bonds*.

Ionic Bonds

In an ionic bond, electrons are transferred from one atom to another. This results in two oppositely charged ions that are bonded by the attraction between opposite charges. For example, table salt is representative of a category of salts whose molecules are held together by ionic bonds. Table salt consists of bonded sodium (Na) and chlorine (Cl) atoms.

Sodium has one valence electron in its second orbital shell while chlorine has seven valence electrons in a valence shell that can hold eight electrons total. When sodium forms an ionic bond with chlorine, sodium “donates” its one valence electron to chlorine. Chlorine gains an electron, meaning it now has eight valence electrons and a fully filled valence shell. This transfer of electrons is shown in Figure 2.

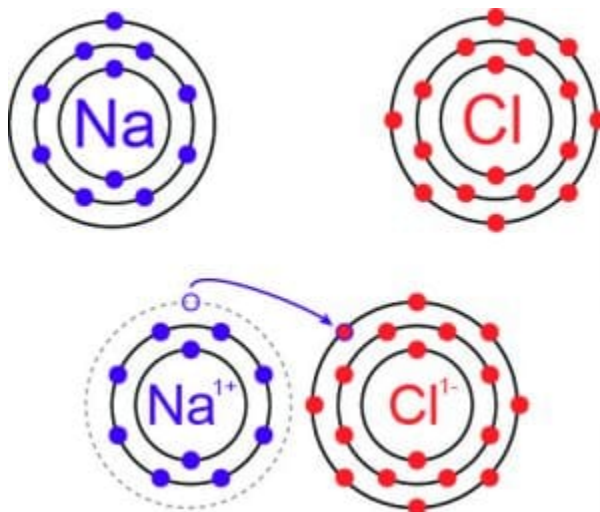


Figure 2 – Top: the electrons of sodium (Na) and chlorine (Cl) organized into the three lowest energy orbital shells. Bottom: sodium donates its one valence electron to chlorine so that both atoms now have filled outer shells.

The sodium atom, whose valence shell is now the filled second orbital shell, has a slightly positive charge. With this extra electron from sodium, the chlorine atom now carries a negative charge. The attraction between the positively and negatively charged sodium and chlorine atoms (respectively) form an ionic bond.

Covalent Bonds

In a covalent bond, the electron orbitals of two or more atoms merge into a new system that allows electron pairs to be shared equally between more than one nucleus. Most chemical bonds in biological organisms are covalent bonds.

There are different types of covalent bonds depending on the number of shared electrons between the two bonded atoms. A single bond between two atoms shares a single pair of electrons, i.e. two electrons total are shared between the two atoms. A double bond shares two pairs (four electrons) and a triple bond shares three pairs (six electrons). Double bonds are shorter and harder to break than single bonds. Triple bonds even more so.

As with all chemical bonds, the details of how and when these bonds form is dependent on an atom's valence electrons. Figure 3 portrays how two hydrogen (H) atoms and one oxygen (O) atom shares electrons to form a water molecule.

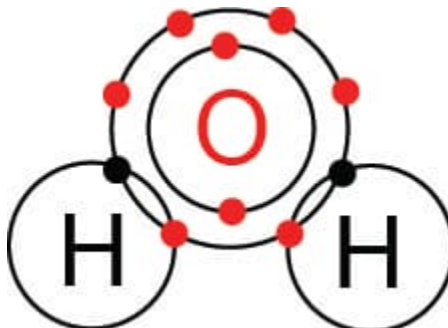


Figure 3 – A molecule of water. Atoms, represented by their chemical signal, are surrounded by black circle(s) representing the atoms' electron shells. Colored dots on these circles represent the atom's electrons. Every hydrogen (H) atom has a full first shell with two electrons while the oxygen (O) atom has a full second electron shell with eight electrons.

A neutral hydrogen atom has one proton and therefore one corresponding electron. With just one electron, hydrogen's valence shell is the lowest energy orbital shell, which can hold up to two electrons. Hydrogen therefore has one valence electron and needs only one additional electron to complete its valence shell.

Oxygen, with eight electrons, has two electrons in the first electron shell and six valence electrons in the second, valence shell. Oxygen needs two additional electrons to fill its outer shell and can form two single or one double bond. In water, we see that oxygen shares electrons with two hydrogen atoms. From this, oxygen acquires access to the two additional electrons it needs to fill its valence shell. Each of the three atoms remains neutral, and each atom has a completely filled outer electron shell.

Chemical Formulas

The molecules, or chemicals, that are formed due to chemical bonding can be represented in a variety of ways. Figure 4 shows examples of different representations for methane, which consists of a carbon (C) atom bond to four hydrogen (H) atoms.

image

Figure A shows C H subscript 4. Figure B shows a carbon atom that is bonded to four hydrogen atoms at right angles: one above, one to the left, one to the right, and one below. Figure C shows a 3-D, ball-and-stick model of the carbon atom bonded to four hydrogen atoms. Figure D shows a space-filling model of a carbon atom with hydrogen atoms partially embedded into the surface of the carbon atom." width="975" height="244" data-media-type="image/jpg"> Figure 4 – A methane molecule can be represented as (a) a molecular formula, (b) a structural formula, (c) a ball-and-stick model, and (d) a space-filling model. Carbon and hydrogen atoms are represented by their chemical symbol in (a) and (b) and as black and white spheres, respectively, in (c) and (d).

A molecular formula uses an element's chemical symbol to indicate the types of atoms in a molecule followed by subscripts to show the number of atoms in the molecule. If there is no subscript, then there is only one atom of that element in the compound. With one carbon (C) atom and four hydrogen (H) atoms, methane's molecular formula is CH₄. Molecular formulas are used in chemistry to specify the molecule and to understand what atoms are needed to make up the molecule.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

It is important to note that a subscript following a chemical symbol and a number in front of a symbol do not represent the same thing. For example, H₂ and 2H represent distinctly different species. With a subscript, H₂ is the molecular formula for two hydrogen molecules bonded to each other. On the other hand, 2H indicates two separate atoms of hydrogen that are not bonded together. The expression 2H₂ indicates two molecules of diatomic hydrogen (see Figure 5).

image

Figure 5 – The symbols H, 2H, H₂, and 2H₂ represent very different entities.

The structural formula for a molecule shows how the atoms are connected in the molecule. Both a molecular and structural formula shows how many of each atom are in a molecule, but a structural formula more clearly indicates how the atoms are bonded together to form the molecule. For methane, the structural formula (see Figure 4 b) reveals that all four hydrogen atoms are bonded to a central carbon atom.

A ball-and-stick model shows the geometric arrangement of the atoms, which are not drawn to scale. As the name suggests, atoms are represented as balls and the bonds between atoms are represented as sticks. A ball-and-stick model expands on the bonds shown in a structural model by representing these bonds in 3D. For methane, the ball-and-stick model (see Figure 4.c) reveals that the four hydrogen atoms bonded to carbon form a pyramidal structure.

A space-filling model shows the 3D arrangement of atoms drawn to scale. It therefore is similar to ball-and-stick models, but incorporates information on atom size. For methane, the space filling model (see Figure 4.d) shows that the carbon atom is much larger than the hydrogen atom. We also see that the chemical bonds between the atoms do not protrude from each atom as suggested in the ball-and-stick model.

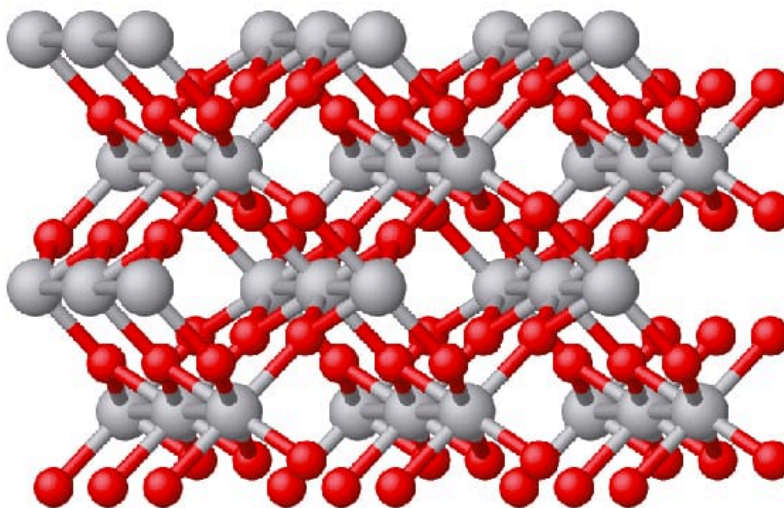
Empirical Formulas

Empirical formulas indicate the types of atoms present in a compound in the smallest, whole-number ratio of the number of atoms in the compound. As an example we consider titanium dioxide, which is commonly used to color white paint and as an ingredient in the thick, white, blocking type of sunscreen (see Figure 6.a).

Titanium dioxide exists as a crystal consisting of numerous titanium (Ti) and oxygen (O) atoms, as depicted as a ball-and-stick model in Figure 6.b. Within a titanium dioxide crystal, there are always twice as many oxygen atoms as there are titanium atoms. The empirical formula for titanium dioxide is therefore TiO_2 , indicating that for every titanium (Ti), there are two oxygen (O) atoms.



(a)



(b)

Figure 6 – The white compound titanium dioxide provides effective protection from the sun. (b) A crystal of titanium dioxide, TiO_2 , contains titanium and oxygen in a ratio of 1 to 2. The titanium atoms are gray and the oxygen atoms are red.

Isomers

With more complex molecular formulas, a variety of isomers become possible. Isomers are compounds with the exact same molecular formula but different structures. Isomers are similar in spirit to homographs—words that have the same spelling but are pronounced differently (like minute, meaning 60 seconds, and minute, meaning small). Isomers cannot be differentiated by their molecular formula alone, it is necessary to use their structural formula to identify an isomer without ambiguity.

There are different types of isomers. Structural isomers describe isomers for which the atoms in the molecules are connected in completely different arrangements. These isomers tend to exhibit completely different behavior. Stereoisomers describe isomers that are mirror-images of each other. It is important to consider the 3D geometry of molecules (for example with a ball-and-stick or space-filling model) to understand the differences in stereoisomers. Lastly, conformers are isomers that differ only in the conformation of the atoms in the molecule. In other words, though the atoms are bonded together in the same way, molecules

can twist along single bonds, which result in molecules with different conformations.

Building Molecules

Most atoms lack some valence electrons in their outermost electron orbital and thus have the potential to form a bond. In general, a molecule is “complete” or no longer prone to reacting when all there are no empty spots for electrons in any of the valence shells of the atoms in the molecule.

You can learn more about building molecules with this [Molecule Building simulation](#)!

CHNOPS: the Most Common Elements in Life

By mass, 95% of all biological systems on Earth are either carbon (C), hydrogen (H), oxygen (O), or nitrogen (N) atoms. This dominance is sometimes referred to as CHON. With phosphorous (P) and sulfur (S), which gives us CHNOPS, these six elements make up 98% of all life on Earth. By considering the valence electrons of these elements, we can understand how they behave and react in biochemistry.

Recall from our discussion of covalent bonds above that hydrogen has one valence electron and needs only an additional electron to complete its valence shell. Because hydrogen is so close to a full valence shell, it is highly reactive—it is easy to form a bond with hydrogen. With its one electron, hydrogen can form bonds wherever valence shells need to be completed; consequently, hydrogen makes up 59% of biological compounds.

Oxygen is short just two electrons of a full valence shell, which also renders oxygen atoms quite reactive. Oxygen is also very electronegative, meaning it can attract shared electrons towards itself. In covalently bonded molecules, this means electrons will spend more time near oxygen atoms than the atom on the other end of the bond. Oxygen in molecules therefore tend to carry a negative charge. In the example of water (see Figure 3), the oxygen atom is slightly negatively charged and the hydrogen atoms are slightly positively charged. This difference in charge makes oxygen a polar molecule.

Nitrogen, with seven electrons, has five valence electrons. This means nitrogen can form up to three bonds: three single bonds, or a single and a double bond, or a triple bond. All these options are commonly found throughout organic chemistry. Oxygen’s ability to attract electrons and nitrogen’s relatively weak bonding of valence electrons drive many biochemical reactions. An example of the possible bonds formed by CHON is given in Figure 7, which diagrams the covalent bonds for a molecule called glycine.

Figure 7 – A molecule of glycine. Left: atoms, represented by their chemical signal, each surrounded by a black circle representing the outermost electron shell, i.e. the valence shell, of the atom. Colored dots on these circles represent the atom’s valence electrons. Every hydrogen (H) atom has a full first shell with two electrons while every carbon (C), nitrogen (N), and oxygen (O) atom has a full second electron shell with eight electrons. Right: the structural formula for glycine. The covalent bonds that are formed are shown as lines between two atoms. Each hydrogen atom has formed one bond while oxygen, nitrogen and carbon have formed 2, 3, and 4 bonds respectively based on their different numbers of valence electrons.

Carbon is so central to organic processes that organic chemistry is sometimes defined as the study of compounds that contain carbon. With six total electrons, carbon has four valence electrons occupying its second electron shell. This allows carbon atoms to form four bonds or a combination of single, double, and triple bonds. Carbon-carbon bonds are particularly easily made and remain very stable. This allows carbon to form large and complicated structures, such as networked lattices (e.g., graphite or diamonds), cyclic structures, and other complex molecules used throughout biology.

Carbon’s ability to form four separate single bonds at once allows it to acquire the important property of chirality. Chirality is defined as the property of an object that can not be superimposed on its mirror image. The classic example, illustrated below in Figure 8, is your left and right hands. When both hands are palm up, they are mirror images of one another. If you lay your hands flat on a table, it is not possible to slide one hand over the other and match the other hand exactly. For this reason, chirality is often also referred to as “handedness”.

Figure 8 – An example of chiral and achiral objects. Hands are chiral objects because their reflections cannot be superimposed, or layed exactly on top of each other. Even though this image draws fingers rather poorly, the point is nonetheless made with the thumbs of either hand that are pointing in different directions. A symmetric vase does not have this property because its reflection is not different. Can you think of something that is not symmetric but still achiral?

Carbon atoms exhibit chirality when they are simultaneously bound to four different things. In such a case, the carbon atom

provides a chiral center. Chirality can have a huge impact on the reactivity of a molecule in biological systems. Further details and an example is give in the next chapter, which discusses chirality in amino acids ([link to anchor/section](#)).

Phosphorus and sulfur, while relatively less abundant, are nonetheless essential to life as we know it. Phosphorus is involved in forming the borders between cells, the back bone of genetic material, and energy storage and transport in cells. Sulfur helps to ensure proteins maintain their shape and carry on important functions. These key biochemicals are discussed more in the following chapter.

Key Concepts and Summary

Biochemistry is driven by electrostatic forces to share electrons of elements as a way of balancing positive and negative charges. The electrons in covalent bonds are shared equally between two or more atoms, while ionic bonds essentially transfer a loosely bound electron to fill the valence shell of another atom. The most common elements used by life on Earth are carbon, hydrogen, nitrogen and oxygen as well as phosphorus and sulfur (CHNOPS). Carbon is central to organic processes with half of its outer valence shell empty and carbon easily forms single, double and triple bonds with other elements. Just two electrons short of filling its outer valence shell, oxygen is highly reactive, driving many reactions in geology and biochemistry.

Review Questions

Exercises

1. Use [Molview](#), a free-to-use software that will allow you to build molecules (especially organic molecules) to learn about the structures of simple and complex molecules. Choose a molecule and investigate how modifying different atoms included in it will affect the structure of the molecule. Find molecules in this text or in other resources that are important for life and see if these molecules are included in its database. Alternatively, build your own molecule using its chemical formula.

Summary Questions

1. How does the electronic structure of atoms affect the way atoms bond?
2. What is the difference between ionic and covalent bonds?
3. What information is contained in a chemical formula? What do the different letters and numbers in a chemical formula indicate?
4. What are CHNOPS? How are they important for life?
5. What is chirality and how does it affect the way different compounds are used by living things?

2.3: Key Biochemicals- Amino Acids, Nucleic Acids, and the Genetic Code

Two key classes of chemicals have been seen to operate in all known living things on Earth: amino acids and nucleic acids. Amino acids are the fundamental building blocks of proteins, which are essential for most biological processes. Nucleic acids make up the genetic code of all living things on Earth. Here, we investigate the chemistry and interplay between these crucial chemicals.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain the structure of amino acids and proteins
- Discuss the importance of proteins for the functioning of life
- Describe the basic features of DNA and RNA and their components
- Discuss the role of an organism's genetic code and the impact of genetic mutations

Amino Acids to Proteins

Amino acids and the proteins they combine to form are found in every living organism on Earth. Proteins are needed to make critical biological reactions happen on timescales relevant to life. Proteins regulate the processes that drive life. Though proteins are found across all living things and carry out a wide variety of roles, every protein on Earth is made up of only twenty different amino acids.

Amino Acids

Amino acids are molecules that are defined by a specific structure; all amino acids consist of a central carbon atom forming four bonds to: (1) a hydrogen atom, (2) an amino group (-NH_2), (3) a carboxyl group (-COOH), and (4) a changeable side chain (see Figure 1 below). In organic chemistry, molecules with a carboxyl group are called carboxylic acids. This along with the amino group gives these compounds the name amino acid.

Figure 1 – The components of an amino acid. Each amino acid contains an amino group (blue) and a carboxyl group (red), which gives them the name amino acid. Amino acids acquire different properties by hosting different side chains (green), often denoted by an R or called the R-group. This allows for the wide variety of functionality seen in proteins.

The amino and carboxyl groups allow amino acids to bond to one another through a process of dehydration, i.e. a process that releases water (H_2O). Figure 2 below diagrams the reaction with structural formulas. Water is formed by the loss of an -OH group and a hydrogen (H) atom.

This process forms a new bond, called a peptide bond, between the carbon and nitrogen atoms. Chains of amino acids are joined by peptide bonds; these chains are therefore also called polypeptide chains. Figure 3 shows an example of a polypeptide chain. Polypeptide chains are then folded into proteins. Note that Figure 3 uses a mixture of a structural and molecular formula; specifically, not all bonds to Hydrogen atoms are shown with a line.

Figure 2 – The formation of a peptide bond. The -OH from the carboxyl group and the -H from the amino group react to form the peptide bond (purple). These atoms are released as a molecule of water, H_2O .

Figure 3 – A polypeptide chain. Pictured above is the chain of the first seven amino acids that make up the OPN1LW gene, which allows humans to distinguish red colors. The different amino acids in the chain feature different side chains, shown and labeled in green. This is simply an example, so there is no need to memorize the amino acid side chains or their names. **Can you identify the repeating pattern of the backbone of the chain?**

The side chains give each amino acid a unique functionality. Examples of amino acids and their side chains are shown in Figure 3 in green. Amino acids can be positively or negatively charged, water-repellent, bulky, bent into different configurations, or have

other properties depending on their side chains. These differences help the polypeptides fold as they form proteins, bind to specific compounds, or chemically react in different ways.

Despite the fact that more than 500 amino acids exist on Earth, living things on Earth incorporate only 20 different amino acids to form the vast array of proteins that are used to regulate chemical reactions in all aspects of life. This is similar in spirit to the concept that even though the English language only uses 26 letters, those same letters make up the words that compose millions of books.

Want to know more: Essential Amino Acids, Body Building, and You

Of the 20 amino acids used by life on Earth, the human body is capable of synthesizing all but nine. These nine amino acids are known as the essential amino acids. It is important to include sources of these amino acids either from meat or plants in a healthy diet since the body has no other source for them.

In fact, many products sell amino acids as supplements targeted towards endurance athletes and body builders. Scientists have been able to trace different amino acids and the role they play in muscle contraction or recovery to identify what the body needs after the coach has said “last set” for the third set in a row. For example, glutamine is drained during intense physical activity. If the body’s glutamine stores become depleted, the body begins to break down muscle cells to compensate.

Proteins

It is commonly said that you are what you eat, but perhaps more correctly, you are what your proteins decide to do. Proteins are the driving force behind the processes of life. Many proteins act as enzymes, which are highly specified molecules that allow complicated organic reactions to progress more easily.

In most organic reactions, the molecules involved must first assume an unfavorable, intermediate configuration (see Figure 5 below) before progressing to the finished product. Enzymes bind to these starting molecules and act to stabilize the intermediary state. This makes it easier for molecules to progress to the desired final products. Enzymes are the primary way to increase reaction rates for biochemical processes.

Figure 5 – An example of how enzymes operate. If the goal is to bend a metal bar until it breaks, the shape of an enzyme will encourage the bent shape of the metal rather than either the starting bar or the ending broken pieces. In this example, the bent bar is the transition state that is stabilized. **If there was an enzyme meant to bend a metal bar into a circular spring, what might that enzyme look like?**

In addition to acting as enzymes, proteins also fulfill several other important roles. Proteins are involved with cell signaling, which helps different cells in the body work together. Antibodies are proteins that work with the body’s immune system to recognize and destroy foreign substances that might cause illness. Structural proteins give shape or rigidity to cells, such as those that make up our nails or hair. Motor proteins allow for the movement of single-celled organisms. In short, proteins are critical for all of the basic functions of life.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Chirality of Amino Acids and Proteins

The central carbon in an amino acid can serve as a chiral center because it is typically bound to four different groups. Recall that chirality is defined as the property of an object that can not be superimposed on its mirror image, like how your palm-up hands cannot lay exactly on one another. The one exception is the amino acid, glycine, whose hydrogen side chain makes it a symmetric molecule.

The chirality of amino acids means there exists both left-handed and right-handed amino acids (Figure 6 below). Oddly, while either configuration is possible, life on Earth only uses left-handed amino acids.

Figure 6 – The two possible configurations of amino acids with differing chiralities. The R group is pointing out of the page in both cases. **Can you think of something that is chiral, such as amino acids or your hands?**

Studies have been done to investigate how the chirality of protein affects how it is made and performs. Synthesis of both left-

handed and right-handed amino acids is not only possible, it is chemically equivalent. From an energy standpoint, protein that is composed entirely of right-handed amino acids should function just as well as proteins made of left-handed amino acids. The only difference is that using only one type of chirality could add an extra layer of regulation in biochemical reactions that may help to reduce synthesis errors.

How did life come to pick left-handed proteins over right-handed ones? There are many competing theories as to the origins of this inequality. One idea is that left-handed amino acids are slightly more water soluble (i.e., easier to dissolve in water), which could have made them easier to incorporate into early life. Amino acids may also have been affected by the polarization of light.

Want to know more: Polarization

The polarization of light defines how the wave component of light oscillates relative to the direction in which the light is moving. Figure 7 shows light traveling from the bottom-right corner to the top-left corner. It shows how the polarization of light can be (1) incoherent, as in the right-most third of Figure 7, (2) oscillate back and forth in only one direction, i.e., linearly polarized as in the center of Figure 7, or (3) circularly polarized as is shown in the first third of Figure 7. Different types of filters (shown in Figure 7 as blue squares) help give rise to these types of polarization. In the early Solar System, it is thought that dust grains could have caused all light to be circularly polarized.

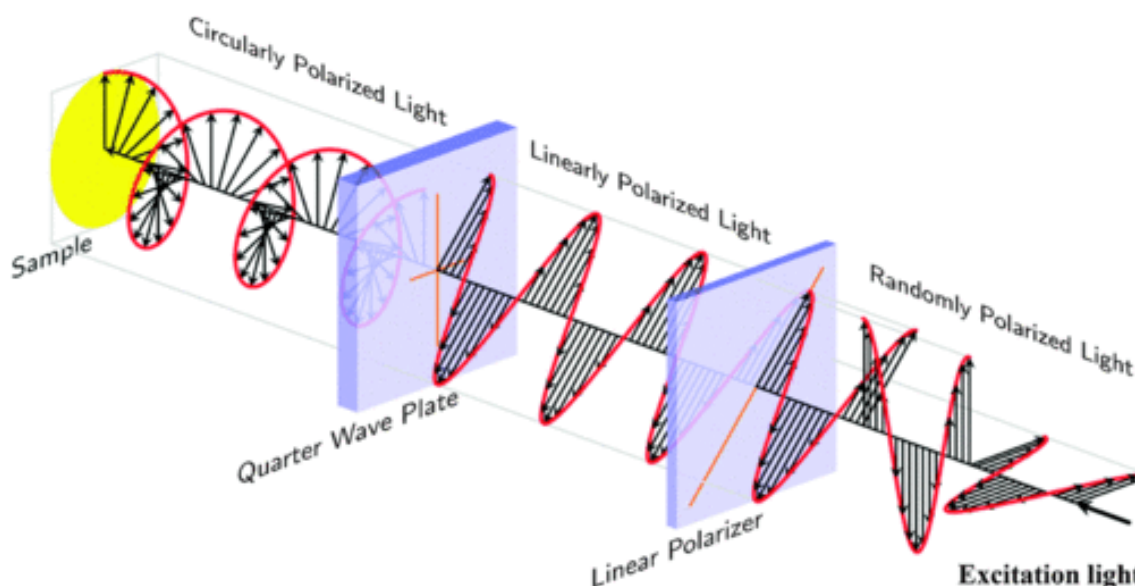


Figure 7 – Three different polarizations of light. The polarization of light defines how the wave component of light (red curve) oscillates relative to the direction in which the light is moving (bottom right to top left in the above figure). Most light, as shown in the bottom and right of the figure, is randomly polarized. This polarization can be changed with different filters, represented by blue squares in the figure. A linear polarizer causes the light to move back and forth along one direction, is shown in the center of the figure. A quarter wave plate results in circularly polarized light, as shown in the top and left of the figure.

It is possible that this circularly polarized light might have been more damaging to right-handed amino acids or more favorable to left-handed ones. Amino acids found in space, for example on meteorites, also exhibit an excess of left-handed molecules. If polarized sunlight gave rise to this imbalance, it could have tipped the scales to the left for life on Earth. Regardless of what established the original inequality, biological processes probably accentuated the imbalance.

Nucleic Acids to Genes

Nucleic acids are used throughout life on Earth to transfer genetic information during cell replication. This genetic information defines the nature and structure of organisms.

Genes that are shared through cell replication and reproduction is why offspring look like their parents. Differences in genetic information can give rise to differences in the traits of an organism. For example, a change in genetic information may result an organism that is taller/shorter or fur color that is lighter/darker.

The process of gene replication—specifically gene mutation, which describes the errors that occur during gene replication—allow for different traits to arise. Some traits may allow an organism to survive better in a given environment. Lighter fur may be useful to blend in with a snowy environment; darker fur may be more helpful in a dark forest.

Organisms with these advantageous traits are more likely to be able to reproduce and pass on their favorable traits. This process of natural selection allows organisms to adapt and evolve to different environments. This ability to grow better equipped to surviving in an environment, also known as Darwinian evolution, is thought to be a key aspect of all living things.

Nucleotides

The nucleic acids found in all known living things on Earth take two forms: ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). Nucleic acids are long, complex chains made up of nucleotides, a specific kind of molecule as diagramed in Figure 8. Nucleotides consist of three components: (1) a central sugar, (2) an interchangeable base, and (3) a phosphate group.

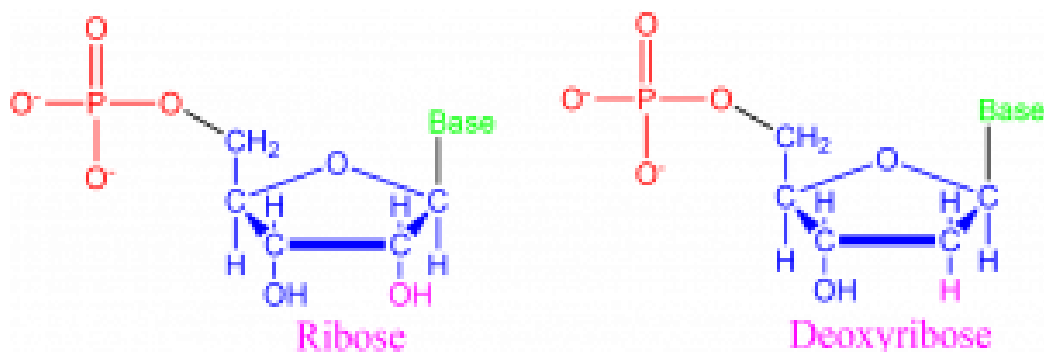


Figure 8 – RNA and DNA consist of nucleotides with three components: a central sugar (blue), a phosphate group (red), and an interchangeable base (green). The difference between the ribose sugar, found in RNA, and the deoxyribose sugar, found in DNA, is highlighted in purple. **What do you think the prefix deoxy- means in this context?**

RNA and DNA differ by the central sugar of the nucleotides that makes them up. RNA uses ribose sugar while DNA uses deoxyribose sugar, which has one less oxygen molecule (note the atoms bonded to the 2nd carbon atom of each nucleotide shown in Figure 8).

The base of a nucleotide changes depending on the specific nucleotide, much like how different amino acids have different side chains. In living things on Earth, five different bases are used. RNA uses adenine, cytosine, guanine, and uracil (ACGU). DNA also uses adenine, cytosine and guanine, but uses thymine in place of uracil (ACGT). These bases bond to each other in specified ways. Cytosine always binds to guanine. Adenine bonds to the thymine in DNA and the uracil in RNA.

The structural formula for the five different bases are drawn in Figure 9. Nucleotide bases fall into two categories: double-ringed purines (A,G) and single-ring pyrimidines (C, T, U). Some of these bases were named after the material where they were first discovered. For instance, guanine was first discovered from guano, a fancy name for bird poop.

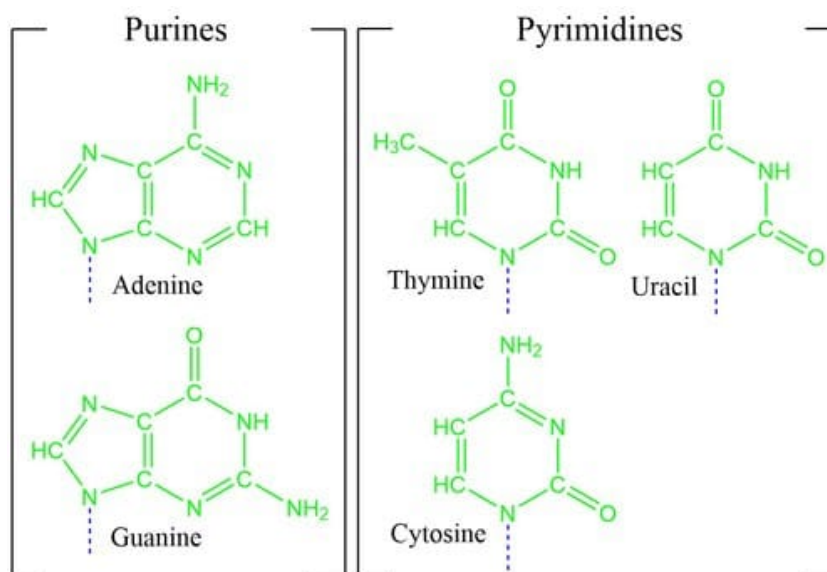


Figure 9 – The five bases known to be used in nucleotides. The bases are separated into purines, with two rings (left), and pyrimidines, with one ring (right). A dotted, blue line indicates where each base binds to the central sugar of a nucleotide. Thymine is used in DNA while uracil is specific to RNA. The other three bases are used in both RNA and DNA. **What is the only difference between thymine and uracil?**

The original version of this chapter contained H5P content. You may want to remove or replace this element.

A phosphate group consists of a phosphorus atom bound to four oxygen atoms. This configuration contains many high energy bonds. The energy stored in this phosphate group allows nucleotides to undergo the reaction that links nucleotides together to form the long-chain nucleic acids that are RNA and DNA. These nucleic acids have a specific structure that was revealed only in the 1950s.

Want to know more: ATP – Cellular Energy Banks

An influx of energy is required to carry out many of the processes necessary for life. You eat to enable your body to create stores of energy to use for these reactions. You eat to create adenosine triphosphate, or ATP.

ATP, shown in Figure 10, is the most widely used energy carrier in all living organisms on Earth. ATP is also a nucleotide consisting of (1) a ribose sugar, (2) an adenine base, and (3) three phosphate groups bonded in a chain. When energy is needed for a chemical reaction, one of the high-energy phosphate to oxygen bonds in the chain is broken. This converts ATP to ADP, adenosine **d**iphosphate, where only two phosphate groups remain.

Phosphate groups are used across living organisms on Earth to supply the energy for necessary reactions. Can you think of reasons why having a common source of energy across reactions would be beneficial? (Hint: would you rather have different charging cables for each of your devices, or one common charging cable?)

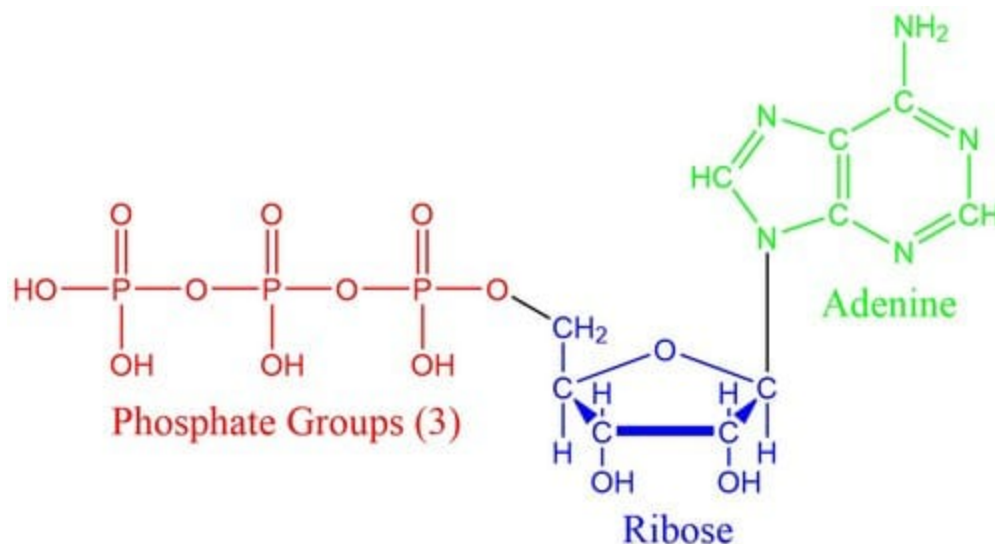


Figure 10 – The chemical structure of adenosine triphosphate, ATP. The three phosphate groups are shown in red. **What bond breaks to release a phosphate group, which consists of one phosphorus (P) atom and four oxygen (O) atoms?**

How do we know: DNA's structure

The exact structure of DNA was revealed through a series of insights that built on one another. When DNA was broken up into its constituent nucleotides, it was discovered that certain bases always appeared in the same proportions. The number of adenine and thymine nucleotides was always equal, and the number of cytosine nucleotides was always equal to the number of guanine nucleotides. In 1949, Erwin Chargoff, an Austro-Hungarian-born American biochemist, sought to explain this observation with the idea of base pairing—the idea that in DNA adenine is always bonded to thymine and cytosine is always bonded to guanine.

Observations of DNA using x-ray crystallography further revealed the structure of the molecule. X-ray crystallography is a complicated technique akin to shining a flashlight into a hall of mirrors and determining where the mirrors are placed based on the way that the light bounces around.

Rosalind Franklin, a British chemist, had a background in physical chemistry that she used to improved on x-ray crystallography techniques in the mid 1900s. Franklin produced unprecedentedly precise x-ray crystallography images (see Figure 11) while working in the lab of Maurice Wilkins. Her most famous photo, known as “Photo 51”, held the key to DNA's structure. This photo was heralded by J.D. Bernal, the father of x-ray crystallography in biochemistry, as “among the most beautiful x-ray photographs of any substance ever taken.”

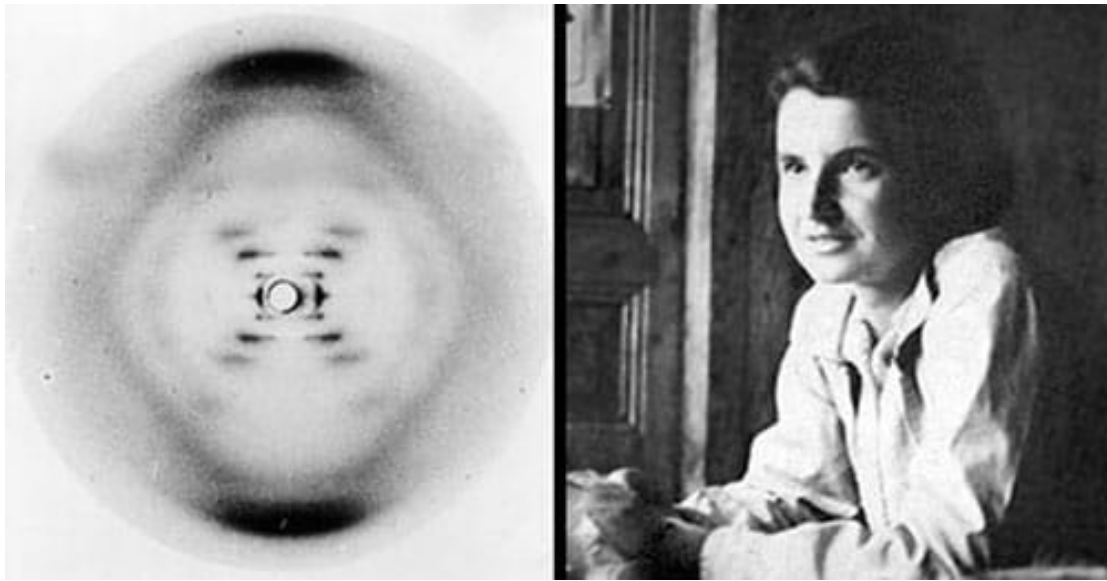


Figure 11 – DNA: picture and pioneer. (left) The x-ray crystallography image of DNA taken by Rosalind Franklin (right) that confirmed the double helix model.

The theoretical biochemists James Watson and Francis Crick used the idea of base pairing and Franklin's images to reveal the double-helix structure of DNA. Watson discovered that the adenine-thymine bond was exactly the same length as the cytosine-guanine bond, which helped him form the picture of each base pair as rungs of a ladder. Crick helped to develop a mathematical model for the pattern that a helical structure would produce with x-ray crystallography.

In 1951, Crick and Watson began to work together. When Maurice Wilkins showed them Rosalind Franklin's Photo 51, they were able to piece together the double-helix model of DNA (see Figure 12).

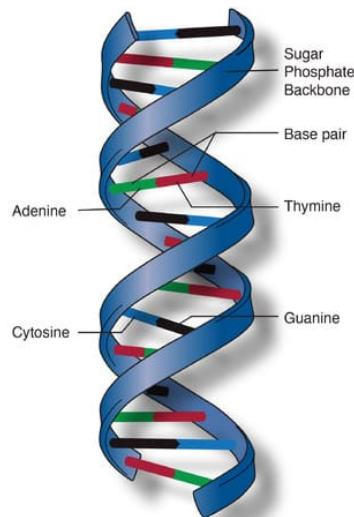


Figure 12 – The double helix of DNA. The base pairs adenine-thymine (green and red respectively) and cytosine-guanine (blue and black respectively) form the rungs of the twisted ladder. The central deoxyribose sugar and phosphate of the nucleotides forms the backbone of the twisted ladder.

The double-helix backbone of DNA is composed of the sugar and phosphate components of nucleotides. The bases stick out from this backbone and bind to their appropriate counterpart through weak hydrogen bonds. In the most common form, the bases appear parallel to each other, like a well-designed stairwell. The double-stranded nature of DNA affords a rigid, stable, and long-lived structure. During DNA replication, each strand is checked against the other to reduce copying errors or accidental mutations.

Translating the Genetic Code

The genetic code in DNA is translated into instructions for how to manufacture proteins with the help of RNA. Messenger RNA (mRNA) transcribes the code from where DNA is located in the cell and carries this information to the ribosome. Ribosomes are the molecules responsible for fabricating proteins in a cell.

A ribosome can read the genetic code from mRNA and translate it to the necessary amino acids to build a protein. Ribosomes themselves are made up in part of RNA, known as ribosomal RNA (rRNA). Each amino acid is specified by different codons, a sequence of three base pairs. Transfer RNA (tRNA) matches up each codon with the appropriate amino acid. These different amino acids are bonded together into a polypeptide chain that can then be folded into the needed proteins.

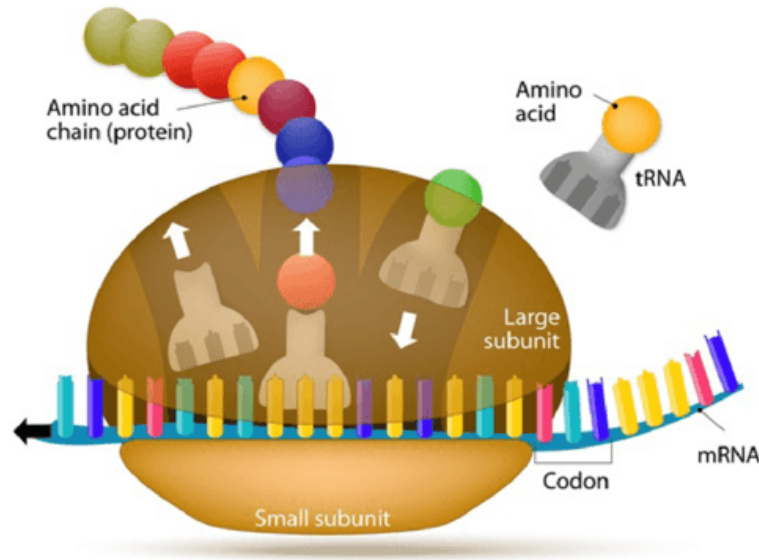


Figure 13 – The process of translating the genetic code to proteins. mRNA, pictured in blue with multicolored bases, carries the genetic information from DNA. This information is interpreted via codons (three base pairs) into amino acids with the help of tRNA carrying the specified amino acid. The rRNA in the ribosome then helps to link together the amino acids into a polypeptide chain ready to be folded into a protein. **What about the tRNA pictured here allows it to correspond a specific codon to a specific amino acid?** Credit:

The Genetic Code

DNA encodes and stores genes, which describe a unit of genetic information. Together, all the genes stored in DNA provide a very lengthy instruction manual for all living creatures on Earth. If you took the entire chain of DNA in one human cell and completely stretched it out, it would measure roughly 2 meters, the average height of an NBA player. If you took all of the DNA from all of the cells in a human body and joined them end-to-end, they would cross the diameter of the Solar System *twice*.

Different segments of DNA are known as chromosomes. An organism's genome is the complete collection of chromosomes. Humans have 23 pairs of chromosomes, encoding roughly 25,000 genes using about 3 billion base pairs. A mosaic of the entire human genome was sequenced between 1990 and 2008. This monumental effort brought together twenty different institutions in six different countries. This remains one of the most impressive collaborative projects in science.

Want to know more: My, What Big Genomes You Have

All the better to encode with...or is it? Intuitively, it might seem that a larger genome would correspond to more complex organisms. A larger genome would allow for more genes, meaning more genetic traits, meaning more complexity. This is wrong.

An example of different genome sizes is given by the table below. Various species are ordered by increasing genome size as defined by the number of base pairs in the organism's complete genome.

Species	Base Pairs	Genes
Virus	170,000	?
E. Coli	4,600,000	3,200

Fruit Fly	180,000,000	13,600
Chicken	1,000,000,000	23,000
Corn	2,500,000,000	59,000
Human	3,000,000,000	25,000
Lily	100,000,000,000	?
Grasshopper	180,000,000,000	?
Amoeba	670,000,000,000	?

Not only does genome size not scale with perceived organism complexity, it also does not scale with the number of genes. Humans have longer genomes than chickens, but we lose out to grasshoppers. Humans also have a longer genome than corn, but less genes.

One reason for this is that most DNA is **noncoding DNA**, which does not translate directly to genes. Noncoding DNA may instead be used to signal the start of a gene, to help with DNA coiling, and/or potentially carry out several other functions that we have yet to discover.

Remarkably, more than 98% of the human genome is non-coding. In contrast, only 20% of the DNA in bacteria is noncoding DNA. The bladderwort plant currently holds the record for most efficient genome with only 3% noncoding DNA.

DNA encodes important instructions for life, but it can become damaged when base pairs or whole segments of DNA are deleted, inverted, duplicated, or moved around. Mutations can be damaging, for example, causing cells to become cancerous. Damage sustained to the phosphate-sugar backbone of DNA is one of the primary causes of mutations. This type of damage is a common result of exposure to UV radiation, such as from the Sun (never skimp on sunscreen).

However, mutations can also occur naturally, resulting in expressed altered genes that give rise to new characteristics. This is the mechanism for Darwinian evolution: beneficial traits arising from mutation will be preferentially selected when mating and propagated through succeeding generations.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Natural mutations arise at measurable rates for different species. This mutation rate allows us to measure the genetic distance between species. This value is obtained by determining the statistical number of mutations required to change one species' genome into another's. For example, deer and giraffes are close in genetic difference, the genome of a deer requires *relatively* few differences to change into the genome of a giraffe when compared to, say, the genome of sunflowers.

Concept Check

If the history of organisms can be traced through their genetic code, is it possible to trace this history all the way back to the beginning of life?

Key Concepts and Summary

Two classes of molecules are common to every form of life on Earth: amino acids and nucleic bases. Understanding the fundamental and relatively simple structure of these molecules allows us to see the more complex patterns of biochemistry. Amino acids are the units for building proteins. While there are hundreds of possible amino acid chains, life on Earth uses only a common set of twenty and some amino acids have been detected with radio telescopes in giant molecular clouds – the star-forming regions of space. The nucleic bases adenine, guanine, thymine and cytosine are components of DNA (and in RNA with a substitute of uracil for thymine). In addition to nucleic bases, the nucleotides in RNA and DNA contain a central sugar and a phosphate group. DNA encodes the genetic information for every organism in a collection of chromosomes. DNA transcription errors result in mutations that occur naturally, but the mutation rate can increase with exposure to ultraviolet light or carcinogenic chemicals.

Review Questions

Exercises

1. It is fairly easy to extract DNA from a strawberry using household materials. See the [step-by-step instructions here](#). In a formal laboratory setting or on your own, follow the steps to extract DNA and take detailed notes of the challenges or successes in your endeavor. What do you observe about the characteristics of the DNA you extracted?

Summary Questions

1. What is an amino acid and how do amino acids form proteins?
2. What functions do proteins have in living things?
3. What is a nucleic acid? How do the two main nucleic acids used by life, RNA and DNA differ?
4. What is the chemical structure of DNA?
5. How are nucleic acids and genes related?
6. What causes genetic mutations?

2.4: Alternative Biochemistries- Speculations about Strange Possibilities

One can only imagine how the field of biology will change the moment extraterrestrial life is discovered for the first time. Will we see shocking similarities? Or will life elsewhere be so different that it is barely recognizable to us?

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss the viability of silicon-based life
- Discuss why liquid water is a useful solvent for life
- Describe some possibilities for alternative nucleic acids

Alternatives to Carbon

While carbon is clearly favored for Earth-based organisms, other hypothetical biochemistries have been considered by biochemists. One of the most frequently imagined alternatives is silicon-based life. Silicon is similar to carbon in that it also has four valence electrons. However, these electrons are in silicon's third electron shell while carbon's valence electrons appear in its second electron shell. The additional shielding of valence electrons by two inner filled shells of electrons is enough to change the nature of silicon bonds. Because of the valence electrons of silicon are farther from the positively charged nucleus, silicon bonds are weaker than carbon bonds. While carbon can effectively form the complex long molecular chains necessary for life, it is rare to find more than three silicon atoms in a single molecule. Even when they do form, compounds with multiple silicon atom have bonds that are easily disrupted by water. The difference in valence shell energy also makes it harder for silicon atoms to form double or triple bonds. Therefore, silicon molecules seldom exhibit chirality.

Because the valence electrons of silicon are more weakly held to the atom, it is easier for oxygen to bond with silicon, but it is also much harder to break these bonds. Once oxygen has bonded with silicon, it takes a lot of energy to pull it away. Biological processes often require recycling elements into different organic compounds. The difficulty in breaking down silicon-oxygen molecules slows the reaction rates. The structure of silicon-oxygen bonds also renders many important organic molecules unfriendly for life. For instance, we exhale carbon dioxide, CO_2 , as part of the process that generates energy. The silicon counterpart of CO_2 , silicon dioxide, is a solid rather than a gas, making it much harder to dispose of than carbon dioxide. In general, silicon tends to form more solid, crystalline structures when compared to its carbon counterparts.

We should not assume that silicon will “never” be a good choice for biochemistry. However, the laws of chemistry seem to strongly favor carbon-based life, especially in temperature and pressure environments found on Earth. The next time your favorite science-fiction movie portrays silicon-based life, remember that it should expel a mouth full of silicon dioxide crystals with every breath.

Water as life's solvent

Water is thought to be essential to life both because where there is life there is water and where there is water there is life. Water possess many unique chemical properties that make it conducive to both life and the organic reactions that govern it.

Water makes up the most of most living systems. The average organism is 70% or more water by weight. Recall that water, or H_2O , consists of one oxygen molecule and two hydrogen molecules. As mentioned before, oxygen is particularly good at attracting electrons. Even within a molecule of water, the electrons preferentially spend time closer to the oxygen molecule than to the hydrogen atoms. Because electrons carry a negative charge, the result is that the oxygen molecule is slightly negative while the hydrogen atoms are left with a slightly positive residual charge.

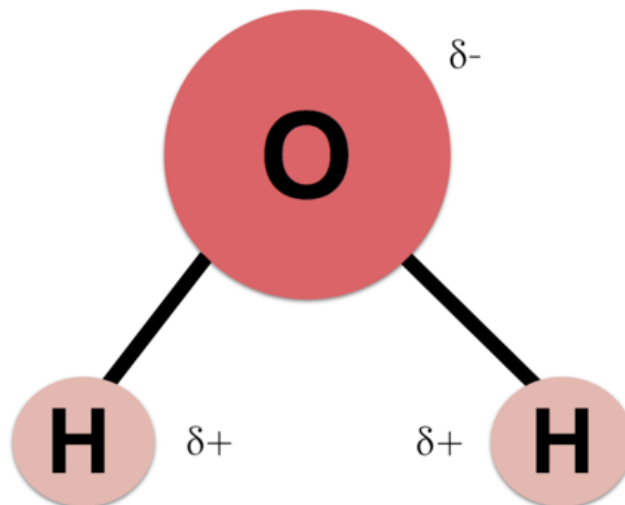


Figure 1 – The water molecule is slightly polar.

The charge differential across the water molecules makes water polar and consequently useful for organic chemistry in several ways. The different charges help group nonpolar molecules together and create effects like self-assembling phospholipid structures. Proteins are also folded with the help of water when amino acids with nonpolar side chains are naturally folded inside more water-friendly side chains.

Because of its polarity, water has a strong bond to neighboring water molecules. Water molecules can form hydrogen bonds with one another. These bonds help water to remain liquid over a greater range of temperature. Water provides a safe, stable environment to nurture the complex reactions required by life.

The difference in charge also allows water to more easily pull compounds apart, making it easier for chemicals (like the salt crystal in Figure 2 below) to dissolve in water. This has earned water the moniker “the universal solvent.” Once in the water, chemicals are free to move around and interact with other chemicals.

Figure 2 – Salt crystal.

A crystal of table salt (NaCl) dissolving in water. The table salt is easily dissolved into negatively charged chlorine ions (green) and the positively charged sodium ions (yellow) in water where the ions are stabilized through interacting with the differently charged ends of water. What section of the water molecule preferentially faces the sodium atom and which end preferentially faces the chlorine atom? Recall that it is opposite charges that attract.

When water freezes into ice, the positively-charged hydrogen atoms align with the negatively-charged oxygen atoms, forming an organized lattice. In this highly-organized configuration, the molecules in ice are more widely spaced than molecules in water, resulting in a lower density that allows ice to float on top of water. This property carries many benefits for biology as ice can therefore serve as a protective barrier for the water underneath. It is capable of blocking out potentially damaging radiation and cooling water to stabilize certain compounds.

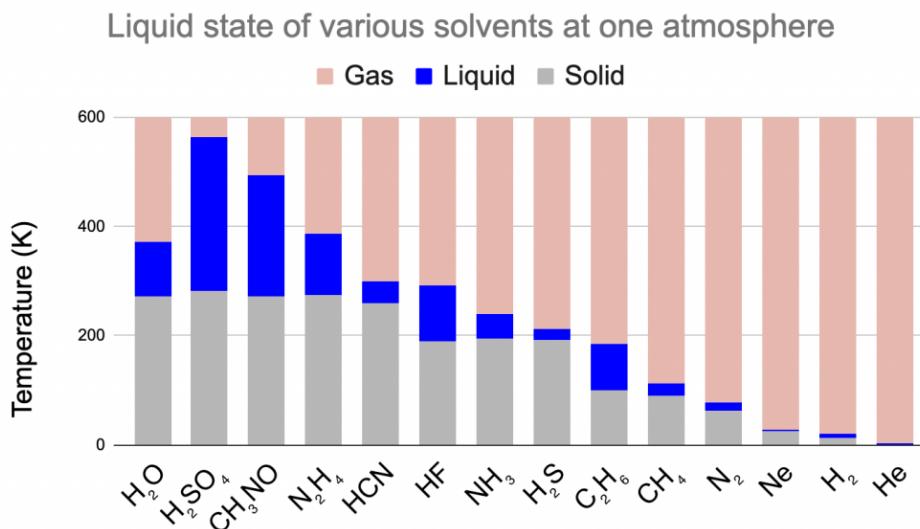


Figure 3 – Phase of solvents at different temperatures.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Could life outside the Earth be based on a different liquid solvent? There are a few considerations to make. If we assume that the liquid state is what is required for biochemistry, then we might consider the range of possible temperatures at which a particular chemical is liquid. This is dependent on the atmospheric pressure with the range increasing as the pressure increases. For convenience sake, the range of temperatures where various chemicals are liquid at standard atmospheric pressure (usually designated with the measurement “one atmosphere”) is given in Figure 3. The compounds which are liquid over the largest range of temperatures are constituted by polar molecules similar to water and generally liquid at temperatures that exceed 200 K.

Scenarios where one of these other chemical compounds would be the dominant liquid on a planet’s surface are subject to the availability of the chemical to the exclusion of others. This is an important consideration as water as a chemical compound is observed to be rather common throughout the universe. The instances where water is not available to act as a solvent in our own solar system are largely due to either the conditions being such so that the material is either permanently frozen solid or is in a perpetual gaseous state sometimes escaping the gravitational pull of the world entirely. In most instances, this results in a dry world with no surface liquid present in any abundance. Other than the Earth, the one exception to this in our own solar system is Saturn’s moon, Titan, which has liquid methane (CH₄) and ethane (C₂H₆) on its surface—this liquid being dominant mostly because the surface temperatures are so low.

Alternatively, subsurface liquids are known to exist in many worlds and it is likely that most of the liquid chemicals in our solar system are contained in such interior oceans. The higher interior pressures beneath solid surfaces provides for environments that are well-suited to preserving liquid state as long as there is some energy source to keep the interiors at high enough temperatures. In many moons, this energy source is derived from tidal stresses. In contrast to larger worlds, they are small enough to have lost the internal heat that was present at their formation and, unlike the Earth, far fewer radioactive elements are present, so some external source of heating is required to keep the interiors warm. The challenges of studying subsurface oceans are similar to the challenges we have on Earth in attempting to access the interior of our own planet compounded by the space travel necessary to reach these distant worlds. Thus, we have yet to directly access any of these subsurface oceans to confirm their composition, but evidence seems to point to many if not most of these subsurface oceans are water.

However, some worlds seem to have interior conditions that might not allow for liquid water. For example, there is evidence that Neptune’s largest moon, Triton, has a reservoir of liquid nitrogen beneath its surface. Such cryoliquids (chemicals that are liquid at very low temperatures) could exist in contexts where temperatures were too cold for liquid water. In many cases, these cryoliquid chemicals are made of nonpolar molecules. At these incredibly low temperatures, chemistry happens at far slower rates and there are some who question whether complex organic chemistry can happen at all, but such bizarre environments are very difficult for us to test carefully since even if we can afford to run experiments at these very cold temperatures, we tend to require human-timescales to run experiments.

Alternative Nucleic Acids

The RNA World Hypothesis sets forth an intriguing possibility that the first proto-life on Earth was derived primarily from RNA and only later did lipids, polypeptides, and DNA become crucial to biochemistry. However, are DNA and RNA the only possibilities for such nucleic acids?

One possibility that briefly excited some in the news media was the claimed discovery in 2010 of a bacteria called GFAJ-1 that appeared to be based on a nucleic acid that had an arsenic backbone rather than the phosphorus backbone of DNA and RNA. Like silicon and carbon, arsenic and phosphorus share similar chemistries. Unfortunately, this claim was disputed by most independent experts who evaluated the claim and a study in 2012 showed that the organism instead of using arsenic actually still relied on phosphorus and was simply arsenic resistant.

More generally, some scientists have speculated and even created entire Xeno Nucleic Acids (XNA) in laboratory environments which use different backbones, sugars, and base pairs. Such study is part of the larger endeavor of **xenobiology** which involves the creation of biologically active chemicals that are not found naturally. Of course, just because such xenobiological chemicals are not found in terrestrial habitats, this does not mean there are no instances of such in the vast universe of possibilities beyond our Earth.

Key Concepts and Summary

Every form of life on Earth uses carbon as the basic atom for biochemistry. We can imagine other types of chemistry for other environments and there has been some speculation about silicon chemistry as an alternative. Like carbon, silicon has four valence electrons, however this is in the third shell rather than the second shell and the silicon bonds are much more fragile than carbon. On Earth, all life requires water as a solvent but on other worlds, there could be non-water solvents. In addition, life elsewhere could use different amino acids and a different genetic code – all of this will serve to make life elsewhere more difficult for us to recognize.

Review Questions

Exercises

1. Investigate XNA possibilities. Below is a list of different nucleic acids which are not used by terrestrial lifeforms. Choose one of them and discuss its chemical properties. Include a diagram in your discussion of how the chemical differs from that of DNA and RNA:
 - Peptide nucleic acids (PNAs)
 - Threose nucleic acid (TNA) or α -(L)-threofuranosyl-(3'-2') nucleic acid
 - Glycol nucleic acid (GNA)
 - 1,5-Anhydrohexitol nucleic acid (HNA)
 - Cyclohexene nucleic acid (CeNA)
 - Morpholino
 - Locked nucleic acids (LNA)
 - Fluoroarabino nucleic acid (FANA)

Each of these artificial nucleic acids differs from naturally occurring DNA or RNA by changes to the molecule's backbone.

Summary Questions

1. Why was silicon-based life considered a possibility by some speculative thinkers? What makes it likely or unlikely to exist?
2. What features of water allow it to be a useful solvent for life?
3. Are there other possible liquid solvents for life? If so, how might they be better or worse than water as a solvent?
4. What parts of DNA might be modifiable to produce alternative nucleic acids that other forms of life might be able to use?

This page titled [2.4: Alternative Biochemistries- Speculations about Strange Possibilities](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

3: ORIGINS OF LIFE

3.1: Definition of Life

3.2: From Biochemistry to Biology

3.3: Emergence of Life on Earth

3.4: Development of Life on Earth

3.5: Extremophiles

3.1: Definition of Life

Biology is the branch of science concerned with the study of living things, or organisms. That definition is pretty straightforward. However, it opens the door to more difficult—and more interesting—questions: What is life? What does it mean to be alive?

Learning Objectives

- Discuss some examples of living things and how we know they're alive
- Describe the attributes that living things have that distinguish them from nonliving things
- Discuss some of the various definitions of life, and why scientists argue over whether they are or are not satisfactory

You are alive, and so am I. The dog I can hear barking is alive, and so is the tree outside my window. However, snow falling from the clouds is not alive. The device you're using to read this article is not alive, and neither is a chair nor a table. The parts of a chair that are made of wood were once alive, but they aren't any longer. If you were to burn the wood in a fire, the fire would not be alive either.

What is it that defines life? How can we tell that one thing is alive and another is not? Most people have an intuitive understanding of what it means for something to be alive. However, it is surprisingly hard to come up with a precise definition of life. Perhaps because of this difficulty, since ancient times there have been “vitalistic” hypotheses—the idea that there must be some quality that life has which makes it different from non-life. This quality has alternatively been called “elan vital” or “life force” or “divine spark.” In spite of many proposals that this is the key distinguishing feature of life, it has consistently eluded measurement. The proponents of these vitalistic hypotheses have claimed to be able to directly observe the thing that distinguishes life from non-life, and have consistently been shown to be incorrect.

A famous example was the case of Franz Mesmer who in the eighteenth century claimed to be able to capture and use what he termed the “magnetism” of living things to cure diseases. Investigations by scientists at the time showed that his claims were based on wishful thinking; when they secretly switched his potions with decidedly non-living vials of water, Franz Mesmer could not tell the difference.

In the nineteenth century, a school of “vitalist chemistry” proposed that the chemistry of living things would always be separate from non-living things. They predicted that it would be impossible to distill chemical compounds produced by living things through artificial or non-living processes. This idea of vitalist chemistry was shown to be incorrect when urea, an organic compound, was synthesized from two inorganic compounds thereby proving that life was not needed to produce an organic chemical.

Even with the debunking of vitalism, proponents continued to argue that there is something special and ineffable about life that science has not yet discovered. The scientific community has largely moved on from this argument and vitalism has been relegated as **pseudoscience** in most circles.

The struggle to clearly define life led to many attempts that are operational descriptions of life—such definitions allow us to separate living things from nonliving ones, but they don't actually pin down what life is. To make this separation, we might try to come up with a list of properties that are, as a group, uniquely characteristic of living organisms. This is the most common solution employed today.

Test Your Intuition

Classify the following as either living or non-living.



How did you decide whether a virus is living or non-living? Note that there is not a generally accepted “yes” or “no” answer to the question of whether a virus is alive.

Properties of life

We can determine properties of life by identifying various traits common to all the living organisms that we know of. Although nonliving things may show some of these characteristic traits, living things consistently show all of them.

Concept Check: What properties do living things all share?

One strategy scientists use to define life is to list the things all life has in common. Make a list of the properties that you think all living things share.

Show Answer

You may have listed **reproduction** or **evolution**. These things are very important for ensuring the survival of a species. A living thing must be able to reproduce or its species won't live on after it dies. A living thing must also be able to evolve to the changing conditions of its environment, or it won't be able to survive to reproduce.

There are many things that nearly all living things share in common, but there are often exceptions which make defining life difficult. For example, your intuition may have been to label the mule in question one as living; however, mules are sterile, and can neither reproduce nor evolve. A mule is one of those exceptions that makes defining life difficult.

Some non-living things have things in common with life. For example, 3D printers can print all sorts of objects, from models of the human brain to life-sized boats. What if a 3D printer printed a copy of itself? Technically the printer would be reproducing.

1. Organization

Living things are highly organized, meaning they contain specialized, coordinated parts. All living organisms are made up of one or more cells, which are considered the fundamental units of life.

Even unicellular organisms are complex! Inside each cell, molecules are organized into cell organelles and coherent structures. Multicellular organisms—such as humans—are made up of many cells. The cells in multicellular organisms may be specialized to do different jobs. Similar cell types can be organized into tissues, such as connective tissue, epithelial tissue, muscle, and nervous tissue. Tissues, in turn, coordinate to create organs, such as the heart or lungs, which carry out specific functions needed by the organism as a whole.

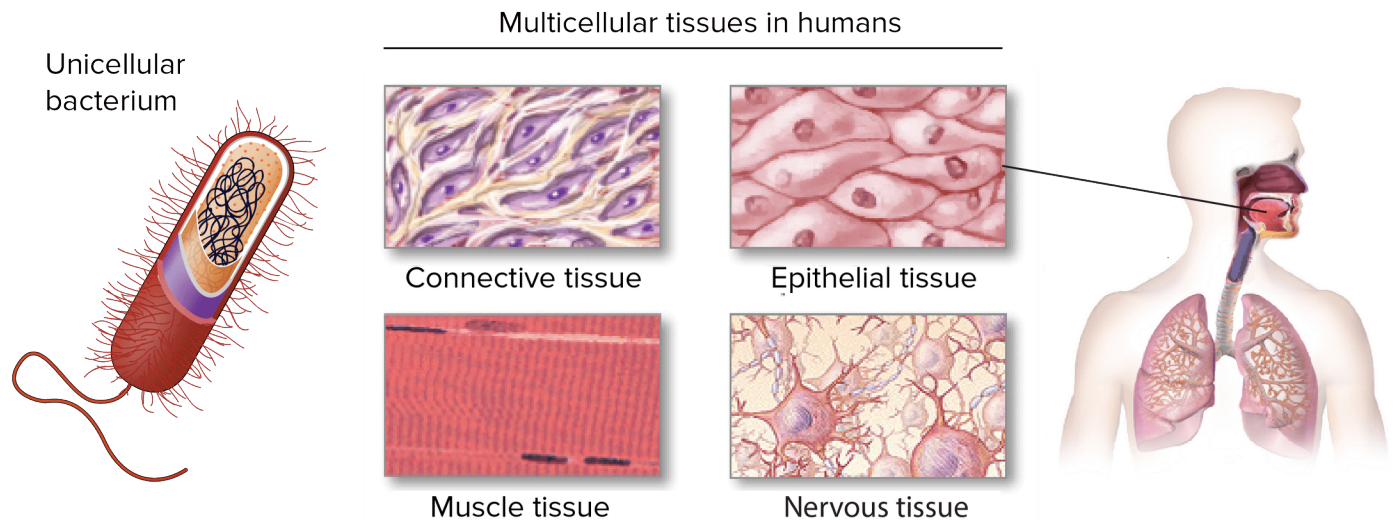


Figure 1 – Left: unicellular bacterium, with the exterior of the cell cut away to show the multiple layers of the cell and the DNA in its interior. Center: multicellular tissues in humans. Small drawings of connective tissue, epithelial tissue, muscle tissue, and nervous tissue. Right: diagram of a human upper body, showing one location where epithelial tissue like that shown in the center panel could be found—the lining of the mouth.

2. Metabolism

All living things require energy to grow, respond to the environment, and reproduce. The series of chemical reactions inside the body that create energy from nutrients are known as metabolism. Life depends on an enormous number of interlocking chemical reactions. These reactions make it possible for organisms to do work, such as moving around to catch prey, growing, reproducing, and maintaining the structure of their bodies. Living things must use energy and consume nutrients to carry out the chemical reactions that sustain life.

3. Homeostasis

Homeostasis alludes to the maintenance of a stable, internal environment, even in the face of a changing external environment. Living organisms regulate their internal environment to maintain the relatively narrow range of conditions needed for cell function. For instance, your body temperature needs to be kept relatively close to 98.6°F (37°C). This stability can be achieved either by sweating to cool off or shivering to warm up.

Concept Check: Homeostasis

Can you think of an example of something non-living that maintains an internal temperature in spite of the temperature of its surroundings?

Show Answer

Stars produce energy in their interiors and maintain a steady temperature for billions of years in spite of being situated in cold, frigid space. Hot springs are also warm regardless of the temperature of the environment above them. They get their temperature from geothermal heat below.

Check out this 7 second video of a hot spring in Lassen Volcanic National Park in California:

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=84>

4. Growth and Change

Living organisms undergo regulated growth. Individual cells become larger in size, and multicellular organisms accumulate more

and more cells through cell division. You yourself started out as a single cell and now have tens of trillions of cells in your body!

Concept Check: Growth and Change

All living things have the ability to grow and change. A seed grows into a tree, a baby grows into an adult, and even a tiny bacterium grows in size until it splits into two. Can you think of some non-living things that can grow and change?

Show Answer

Crystals, glaciers, rivers, and mountains grow in size and change over thousands of years. However, the causes of growth and change to these things are based on physics rather than biology, e.g., temperature-pressure conditions, climate, and plate tectonics.

Check out this 30 second time lapse of a crystal growing:



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=84>

5. Reproduction

Living organisms are capable of reproduction, which creates new organisms. Reproduction can be either **asexual**, involving a single parent organism, or **sexual**, requiring two parents. Single-celled organisms, like the dividing bacterium shown in the left panel of the image at right, can reproduce themselves simply by splitting in two!

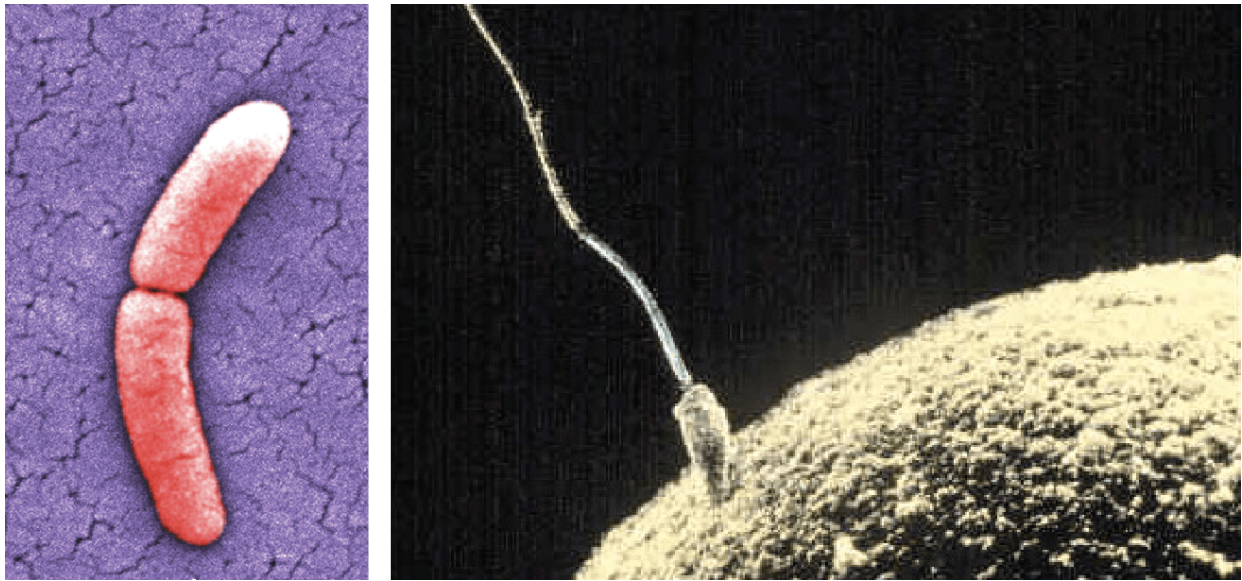


Figure 2 – (left) Salmonella bacterium dividing into two bacteria.(right) image of sperm and egg meeting during fertilization.

In sexual reproduction, two parent organisms produce sperm and egg cells containing half of their genetic information, and these cells fuse to form a new individual with a full genetic set. This process, called fertilization, is illustrated in the image above on the right.

6. Response

Living organisms show “irritability,” meaning that they respond to stimuli or changes in their environment. For instance, people pull their hand away—fast!—from a flame; many plants turn toward the sun; and unicellular organisms may migrate toward a source of nutrients or away from a noxious chemical.

See a plant (*mimosa*) respond to touch:



Concept Check: Response to changes in environment

All living things are capable of responding to some changes in their environment. Changes in temperature, humidity, sound, light, and physical contact with another living thing can cause a response, such as growth, movement, dormancy, or aggression. Can you think of a few non-living things that respond to changes in the environment?

Show Answer

A pond changes with the weather. It freezes over in the winter, dries up during a drought, and becomes full during periods of rain.

Fire also responds to changes in the environment. A forest fire is intensified if it's windy, but if you put water on the fire, it will extinguish the flames.

7. Evolution

Populations of living organisms undergo **evolution**, meaning that the genetic makeup of a population changes over time. Living organisms must therefore have a system for storing genetic material, such as in DNA and/or RNA, and a pathway through which this genetic material can change.

In some cases, evolution involves **natural selection**, in which a heritable trait, such as darker fur color or narrower beak shape, lets organisms survive to reproduce better in a particular environment. Over generations, a heritable trait that provides a reproductive advantage may become more and more common in a population, making the population better suited to its environment. This process is called **adaptation**.

Separating living and non-living things

How well do the properties above allow us to determine whether or not something is alive? Let's revisit the living and nonliving things we saw in the introduction as a test.

The living things we listed in the introduction—humans, dogs, and trees—easily fulfill all seven criteria of life. We, along with our canine friends and the plants in our yards, are made of cells, metabolize, maintain homeostasis, grow, and respond. Humans, dogs, and trees are also capable of reproducing, and their populations undergo biological evolution.

Living things may keep some properties of life when they become nonliving, but lose others. For instance, if you looked at the wood in a chair under a microscope, you might see traces of the cells that used to make up the living tree. However, the wood, having been made into a chair, can no longer grow, metabolize, maintain homeostasis, respond, or reproduce. A wooden chair is therefore not alive.

Nonliving things may exhibit some of the criteria for life. For instance, crystals of snow are organized—though they don't have cells—and can grow, but they don't meet the other criteria of life. Similarly, a fire can grow, reproduce by creating new fires, respond to stimuli, and can arguably even be said to carry out chemical reactions that constitute a metabolism. However, fire is not organized, does not maintain homeostasis, and lacks the genetic information required for evolution.

Indeed, it seems biochemistry is a key component of living things that allows us to make a clear distinction between living and nonliving things. The four main types of chemicals living things utilize for this are proteins, nucleic acids, carbohydrates, and lipids. We will spend considerable time investigating the details of these chemicals to understand the contexts where they arise and speculate on whether and how they might exist in environments well outside the Earth.

Concept check: Can a robot be considered alive?

Consider the properties listed above and assess whether a robot fulfills each criteria or not.

Show Answer

Let's test out our list of properties on a trickier case. Imagine a very sophisticated robot, such as R2D2 or C3PO from the Star Wars movies. Such a robot would show organization—without cells—respond to stimuli, and even have a metabolism of sorts, using energy to power the circuitry of its “nervous system”. It might even maintain homeostasis, with an internal fan or heater that turned on if the temperature changed.

A typical robot would not grow, reproduce, or be part of an evolving population, and thus wouldn't be considered alive. However, what if a robot were programmed to add pieces onto itself? To build more robots? To build more robots with variations in the “DNA” of their programs? As these ideas show, a sophisticated enough computer or robot—far beyond what we have today!—could start to bend the definition of life.

What counts as life is still being defined.

Even with the biochemical definition of life in hand, the question of what it means to be alive remains unresolved. Living organisms have many different properties related to being alive, and it can be hard to distinguish the exact set that best constitutes an operational definition of life. Different thinkers have developed different lists of the pertinent properties of life. For instance, some lists might include movement as a defining characteristic, while others might specify that living things carry their genetic information in the form of DNA. Still others might require that life is carbon-based, meaning the molecules that make up a living organism must incorporate carbon.



Figure 3 – Image of a mule on a farm. The mule looks similar to a donkey and is clearly a living animal, despite the fact that it cannot reproduce.

It's also true that the list above is not foolproof. For instance, a mule, pictured below, is the offspring of a female horse and a male donkey. Mules look and act similar to donkeys and are easy to categorize as “living.” However, mules are unable to reproduce, thereby failing to fulfill one of our seven criteria of life listed above. Does this make a mule non-living?

Viruses are another debated category. Viruses are tiny structures made up of proteins and nucleic acid genetic material that can only reproduce inside host cells. They have many of the properties of unicellular life—they are organized structures, respond to stimuli, and evolve. However, they cannot reproduce without a host. They also aren't capable of carrying out their own metabolism, and it is not clear that they maintain homeostasis.

For these reasons, biologists do not generally consider viruses to be alive, but not everyone agrees with this conclusion. Whether viruses count as life remains a topic of debate. Some even simpler molecules, such as self-replicating proteins—like the “prions” that cause mad cow disease—and self-replicating RNA enzymes, also have some of the properties of life, like viruses, but not all. Should these simpler molecules be considered living or non-living?

Moreover, we are basing our definitions of life on the characteristics of life we see on earth. If extraterrestrial life exists, it may or may not share these same characteristics. Indeed, NASA's working definition of life states that:

“Life is a self-sustaining system capable of Darwinian evolution.”

Let's break down this definition into two parts: 1) Self-sustaining chemical system, and 2) capable of Darwinian evolution.

Self-sustaining chemical system: A system can be made up of many parts. As individual humans, we cannot live forever, but as a species (i.e., *Homo sapiens*), we can live indefinitely. Our method of self-sustaining our species is known as sexual reproduction.

Concept check: Reproduction

Let's return to the selections you looked at earlier in assessing what is alive.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

For each selection, ask if it can reproduce copies of itself without a continued source of energy. If they can reproduce, identify how. For example, mules cannot sustain themselves as a species as they cannot reproduce.

Show Answer

Dogs are self-sustaining. Similar to humans, dogs can sustain themselves indefinitely via sexual reproduction.

Trees are self-sustaining. Trees reproduce both sexually and asexually, and can therefore sustain their species indefinitely.

A **3D printer** is self-sustaining. If a 3D printer was programmed to take raw chemicals from the environment and print a copy of itself, it could sustain itself until the chemicals run out.

A **virus** is self-sustaining. There are many species of viruses. Each species could potentially be considered self-sustaining as they can reproduce and live indefinitely as long as living cells are present in their environment.

Fire is not self-sustaining. A fire dies out as soon as it runs out of fuel (e.g., wood). Even big forest fires only last for about five weeks. There is no method a fire has to sustain itself in the face of its rapidly depleting fuel source.

A **crystal** is not self-sustaining. Individual crystals are very strong because the atoms they are made out of are arranged in a compact, organized manner. Their atomic structure helps them exist in the environment for billions of years. However, they cannot copy or reproduce themselves as a chemical system.

The other half of the NASA definition is:

Capable of Darwinian evolution: Darwin published his theory of evolution in a book titled, *On the Origin of Species* in 1859. The general concept is that when living things reproduce, they do so imperfectly (e.g. we are not exact clones of our parents). This allows variation amongst a group of similar living things (i.e. a species). Variations can lead to competitive advantages amongst individual living things, which are passed on to their children, evolving the species to be better suited to the environment.

At its core, being capable of Darwinian evolution means being capable of copying yourself imperfectly.

What do you think?

The list given here provides a reasonable, though not foolproof, set of properties to help distinguish between things that are alive and those that are not. How would you define life? Would you add something to the list of properties above, subtract something, or use an entirely different definition? Can you think of exceptions or special cases that aren't covered by the list?

Key Concepts and Summary

The definition of Life is central to the search for life. We want to know the attributes of our quest, but an all-encompassing definition of life currently eludes us. Life on Earth is organized as one or more cells that carry out chemical reactions (biochemistry) for metabolism, growth, and reproduction. The hallmark of life is that populations of organisms undergo evolution through natural selection, with genetic modifications to adapt to the environment. However, evolution will be difficult to monitor or detect on distant worlds.

This article is remixed from <https://www.khanacademy.org/science/biology/intro-to-biology/what-is-biology/a/what-is-life> under a CC BY-NC-SA 4.0 license.

This page titled [3.1: Definition of Life](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

3.2: From Biochemistry to Biology

How did the first life emerge on Earth? At some point, abiogenesis — the transition from chemicals to living cells — occurred. Since there is no direct evidence of this first life, scientists must rely on experiments that try to recreate the conditions on the early Earth to see how easily the building blocks of life can assemble.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe the Miller-Urey experiment and how it contributes to our understanding of how life formed on Earth
- Discuss the main ideas of the RNA World hypothesis for how life formed on Earth
- Discuss how synthetic biology informs the study of the emergence of life on Earth
- Describe how vesicles can form spontaneously from prebiotic chemistry
- Discuss how the first protocells formed and could plausibly replicate

The Building Blocks of Life

What were the conditions on the early Earth like? After the Earth formed and the period of heavy bombardment ended, chemical reactions occurred between simple molecules, leading to more complex molecular structures. How far can natural processes go toward the formation of the first biological cell? How can the non-living suddenly become alive? The study of **prebiotic chemistry** has yielded some important insight into these questions.

The Miller-Urey Experiment

Once biochemists understood some of the important chemistry in living organisms, a new field of synthetic biology was born. Biochemists tried to replicate the pathway for the evolution – or the complexification – of simple chemistry. Could chemicals on the early Earth provide a pathway to the formation of the first living cells?

In 1953, Stanley Miller and Harold Urey published the results of their now famous experiment: they spontaneously produced amino acids from simple elements under conditions that they believe emulated early Earth. Given the importance of amino acids and proteins for life, this experiment was viewed as an important step toward understanding the origin of life.

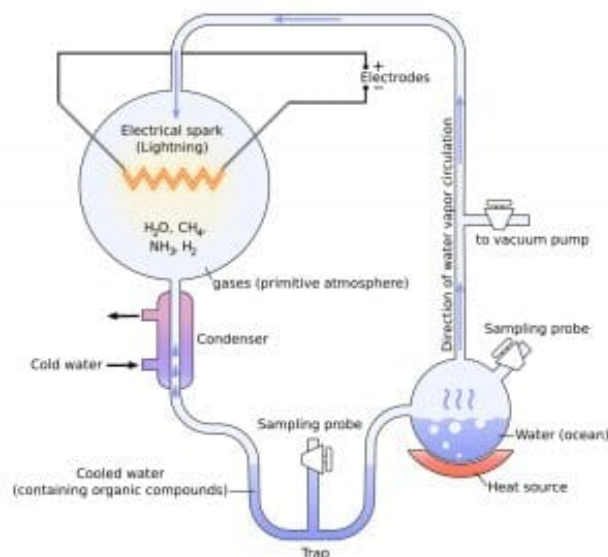


Figure 1 – The set up of the Miller-Urey experiment. Gases characteristic of the early Earth atmosphere (water, methane, ammonia, and hydrogen) filled one bulb and was subject to lightning while the other bulb was filled with water and heated to simulate the ocean and induce evaporation. **What is the reasoning behind having two locations for sampling probes, one directly connected to the ocean and one in the u-bend trap?** Credit: Wikipedia, CC BY-SA

The experimental setup, depicted in Figure 1 above, consisted of two large flasks. One flask contained gases thought to be present in the primitive atmosphere (methane, ammonia, and hydrogen) and the other contained water, representing the ocean. Electrodes were inserted into the atmosphere flask to produce electrical sparks that would emulate lightning.

The ocean flask was heated, producing water vapor that traveled throughout the setup and interacted with gases in the atmosphere and simulated lightning. The enriched vapor then traveled through a condenser and cooled to a liquid state in a u-bend at the bottom of the setup where samples could be collected. Within a day of running the experiment, the liquid at the bottom of this u-bend had turned pink. The inaugural run of the experiment lasted for a week. From these results, Miller was able to identify five different amino acids that had been created from this simple set up of atmospheric gases, lightning, and water alone. As expected, the chirality of these amino acids included both left- and right-handed molecules, even though only left-handed amino acids are incorporated into biology as we know it today. Recent reanalysis of sealed vials have detected more than 20 different amino acids that were actually formed during that one-week experiment.

Since 1953, there have been many additional experiments that aim to synthesize the building blocks of life. Longer chains of amino acids and polypeptides have been synthesized. An impressive experiment was carried out by John Sutherland's lab in 2009. Sutherland began with a compound called 2-amino-oxazole, and using phosphorus as a catalyst, he was able to synthesize some of the nucleotides used in RNA and DNA.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Want to know more: The Murchison Meteorite – amino acids in space

On September 28, 1969, two months after Apollo 11 landed on the moon, a huge meteorite soared through the sky and landed near Murchison, Victoria, in Australia. The Murchison Meteorite is one of the most valuable and well-studied meteorites. Importantly, the incoming meteorite was highly visible. It flew through the atmosphere as a fireball and caused a noticeable tremor when it landed. This allowed for rapid recovery of the meteorite, preventing terrestrial contamination. A piece of the Murchison meteorite is shown below.



Figure 2 – A segment of the Murchison Meteorite. The Murchison Meteorite is classified as a carbonaceous, or carbon-rich, chondrite, a particularly interesting class of meteorites due to the central role carbon plays in biology. Chondrules are round grains present in meteoroids or asteroids that formed as round, molten droplets. Can you spot the round chondrules visible in this image of the meteorite?

In total, more than 100 kg (the average weight of an NFL linebacker) of the Murchison Meteorite was recovered over an area of 13 square kilometers. This meteorite contained more than 70 amino acids, many of which are not known to exist naturally on Earth. There is a slight over-abundance of left-handed amino acids, which suggests that there may be a cosmic origin to the dominance of left-handed amino acids that are used by life on Earth. Every amino acid discovered on the Murchison Meteorite has since been successively synthesized using a Miller-Urey-like setup with gases, water, and simulated lightning, demonstrating that these amino acids can spontaneously form under fairly simple conditions. The Murchison Meteorite is proof that amino acids can easily form in space. Amino acids: so easy to synthesize that a rock can do it!

Even more complex organic molecules have been detected in interstellar space, a testament to how easily complex, carbon-based structures can form and remain stable. Some of these molecules have carbon ring-structures called polycyclic aromatic hydrocarbons and are similar to the organic molecules found in biological organisms. This confirms that carbon is a good base for forming complex molecules structures, including amino acids. While this gives a hint that carbon-based life may also exist on other planets, there are still many steps we don't yet understand about how to go from prebiotic chemistry to life.

Abiogenesis

Keep in mind that just because we discover processes that can create the chemicals we have identified as vital to living things does not answer the question of how life emerges. The transition from such chemicals to lifeforms and biotic systems is a process generally called **abiogenesis**. Many mysteries remain in the study of this question.

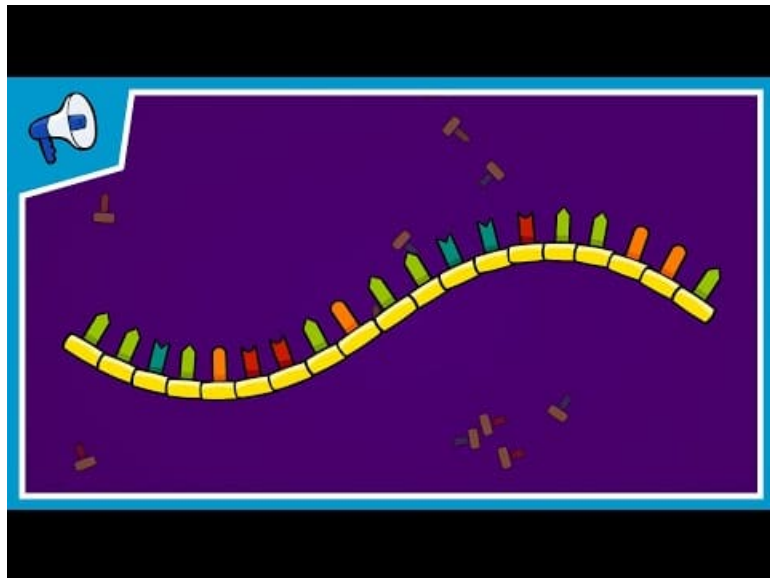
Once amino acids are formed, they can link together through peptide bonds to form chains (see the section on [“amino acids to proteins”](#) for a review). These polypeptide chains can fold into proteins, which perform functions essential for life. Thus, we see a natural increase in the chemical complexity of simple molecules, leading from monomers to polymers. Just as amino acids polymerize to form proteins, another crucial step in the origin of life involves the polymerization of nucleotides to form RNA (see [“nucleic acids to genes”](#)). Let's explore how the RNA world hypothesis provides insights into the early development of self-replicating systems and the foundation of genetic information.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

RNA World

Proteins are needed to catalyze chemical reactions critical to the survival of cells. But, it is difficult to imagine how proteins could have been the precursor to living cells because DNA is required to manufacture proteins, presenting a “chicken-and-egg” problem. A breakthrough came when Sidney Altman and Thomas Cech discovered a class of RNA molecules called **ribozymes**

that could catalyze their own replication. Ribozymes show that RNA, which can encode genetic information, can also act as an enzyme. This discovery was awarded the [Nobel prize in Chemistry in 1989](#) and supported a hypothesis called the Early RNA world, where ancient life used RNA for storing genetic information and catalyzing chemical reactions. This hypothesis had been suggested in the 1960's by Carl Woese, Frances Crick and Leslie Orgel.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=88>

After watching this video, answer the following questions:

1. What are three reasons why RNA is thought to have given rise to the first life?
2. Explain how mutations contribute to the evolution of RNA strands.
3. How can a single strand of RNA become a ribozyme? Draw a sketch of this process.

Video source: [@StatedClearly](#)

According to this hypothesis, the instability of RNA promoted mutations and natural selection eventually evolved a more stable, double-stranded DNA molecule as ribozymes were phased out. A fascinating “smoking gun” for this hypothesis is the fact that the ribosome, which assembles proteins in cells today, is a ribozyme! While the current day ribosome incorporates some proteins, none of the proteins are anywhere near the active site where chemical reactions take place. They appear to exist largely for structural support for the ribosome.

While the capabilities of RNA seem to make it the perfect candidate to explain the origin of life, RNA is a far more complicated structure and not as easy to make as amino acids. Miller-Urey experiments have been capable of synthesizing a series of smaller, very reactive molecules. When enough of these molecules are made, detectable amounts of purine and pyrimidine bases, which are essential components of nucleotides needed for RNA, can be detected. Components of nucleotides have also been discovered on meteorites like the [Murchison Meteorite](#). However, no complete extraterrestrial nucleotides or nucleic acid chains have been discovered yet.

A breakthrough came in 2009 from British chemist John Sutherland along with Matthew Powner and Beatrice Gerland (read more [here](#) about their experiment and results). Sutherland's group determined a chemically efficient pathway for nucleotides to form that is plausible in a prebiotic environment. Rather than form each component of the nucleotide individually, which would require separate and unlikely chemical environments, they proposed a method that formed and attached a purine and a ribose sugar in the same reaction. Phosphate is used to help catalyze the reaction and incorporated into the nucleotide later. This work showed that the building blocks of RNA could form naturally on the early Earth.

Knowing that RNA nucleotides can form under prebiotic conditions, how can these nucleotides assemble into self-replicating RNA molecules? In 2009, Drs. Tracey Lincoln and Gerald Joyce created an RNA enzyme that was capable of self-sustained replication indefinitely (read their [paper in Science](#)).

Dr. Gerald Joyce discusses research in his lab on the self-replication of RNA.

After watching the video, consider the following discussion questions:

1. Why are scientists interested in creating self-replicating systems?
2. According to the timeline of events on Earth that Dr. Joyce shows, when did abiogenesis most likely occur on Earth?
3. What is the “Replicator” that Dr. Joyce refers to?
4. What biological process do these synthetic systems mimic? (cell division)
5. Summarize the first method that Dr. Joyce discusses on how to build a replicating system using just prebiotic chemistry.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=88>

Video source: @MoleCluesTV

Understanding how nucleotides can assemble into self-replicating RNA molecules explains part of the puzzle of how biological life began. But how did these complex molecules organize into structures resembling the first living cells? Next, we examine the spontaneous formation of protocells, or vesicles, from fatty acids, providing a protective environment where these RNA molecules could further evolve and function.

Encapsulation and Protocells

Now that we have complex RNA molecules that can make copies of themselves, what's the next step for forming life? Encapsulation, the action of capturing and surrounding something within a container or membrane (which is a wall made out of chemicals) is suggested by astrobiologists as being an important process in forming primitive life. Cells today are bound and regulated by membranes, which are composed of phospholipids. It is not clear how the first phospholipids formed, but it is possible that the earliest membranes in protocells were first composed of a less complicated lipid connected to hydrocarbon chains.

Protocells can be thought of as simple cell-like structures that form spontaneously from molecules such as fatty acids that were present in the prebiotic environment. For a protocell to form, it requires that it can encapsulate all the molecules needed for that cell to function. However, if you've ever heard the old saying “oil and water” don't mix, you'll know that certain molecules don't generally dissolve in water, making them hard to encapsulate within a protocell. So, how do we get different types of molecules to all cooperate together in one protocell?

A type of molecule called an amphiphile may be the solution to our “non-mixing molecules” problem. Amphiphiles are molecules that have a water “loving” head (i.e. part of the molecule that can dissolve in water) and an oil “loving” tail (i.e. part of the molecule can dissolve in oil). This allows it to interact with both oils and waters, helping them co-exist to form protocells. Detergents such

as dish soaps are amphiphilic. The hydrophilic head of dish soap will bond with water molecules and the hydrophobic tail will bond with the oils (for example, grease on a pan), therefore allowing the grease to mix with the water and be removed from the pan.

The first protocells developed from even simpler compartments which are composed of a type of molecule called lipids. An early form of protocells, called vesicles, can form naturally and can randomly encapsulate different molecules at different concentrations. Vesicles can be made in a lab!



The formation of a vesicle in an environment that goes through wet and dry cycles. An example is a tidal pool or volcanic hot springs. Credit: Ben Pearce, public domain

Vesicles can form in a natural environment due to what is known as wet/dry cycling (Figure 2). When an environment is dry, the lipids form sheets (Step 1), however when they get wet again (like after it rains) the sheets can slowly bud off (Steps 2-4) and eventually form vesicles (Step 5) and encapsulate whatever material was between the lipid sheets. This material can include molecules important to the origin of the first life!

Clays like montmorillonite can also catalyze the formation of longer RNA chains by providing a surface upon which lipid molecules can become concentrated, react and polymerize. Chains of up to 50-60 nucleotides have been formed experimentally in this way.

Key Concepts and Summary

The biochemistry for life first emerged on Earth 3-4 billion years ago. Since the fossil evidence of this chemistry has long been erased, some scientists have tried to recreate the process in their labs. One of the most famous experiments was carried out by Miller and Urey in the 1950's and resulted in spontaneous synthesis of five different amino acids. More studies point to an early RNA world that evolved a molecular code – a precursor of our genetic code – to adapt to environmental conditions.

Review Questions

Summary Questions

1. What was the Miller-Urey experiment? How was it set up?
2. What did the results of the Miller-Urey experiment imply about the formation of the building blocks of life on the early Earth?
3. What does abiogenesis mean? Is there direct evidence for abiogenesis on Earth? Explain your answer.
4. What are ribozymes?
5. How does the RNA World hypothesis explain how ribozymes could be a pathway to the first life on Earth?
6. How has experimental work on synthetic self-replicating systems informed our ideas on the first life on Earth?
7. How can protocells naturally form on the early Earth, without the need for proteins?
8. How can clay minerals, like montmorillonite, facilitate the process of forming protocells?

Exercises

1. Check out the [Exploring Origins Project](#), created in collaboration with Jack Szostak's lab in 2006.

3.3: Emergence of Life on Earth

Once life formed on Earth, over time it became more complex...eventually leading to humans capable of pondering their own existence. All life on Earth descends from the very first instance of life. Natural selection drove species to form and adapt to new environments and ecological niches. The [rise of oxygen in the atmosphere about 2.5 billion years ago](#) caused a proliferation of new life forms that had more efficient metabolic pathways. We will look at some of the milestones in the formation of more complex life throughout Earth's 4.5 billion year history.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe the three different domains of the tree of life
- Explain the difference between prokaryotes and eukaryotes
- Explain how endosymbiosis leads to more complex cells
- Describe how cyanobacteria increased the amount of oxygen in the Earth's early atmosphere
- Explain how aerobic respiration is more efficient than anaerobic respiration

The Last Universal Common Ancestor

The Last Universal Common Ancestor (LUCA) on Earth is a concept, rather than an actual organism. Any universal characteristics of life on Earth are universal either because they are inherited or because they are truly fundamental to life in general. Without a second example of life we are unable to understand how common these features will be on other worlds.

LUCA represents the earliest shared qualities of ancestral life and likely appeared on Earth between 3.5 to 4 Gya and seeded our planet with life. All of the features of life today would have been inherited from LUCA. From what we know about terrestrial biology, this means that LUCA would have been carbon-based, dependent on water, incorporated left-handed amino acids, and used ATP for energy transport. LUCA would have used DNA or RNA to encode genes and translate them into proteins. Significantly, the codons that translate for specific amino acids are the same in every known organism. This code must have been passed down from a common ancestor from which every other species has since evolved. LUCA might not have even been as sophisticated as a single celled organism.

The Tree of Life

The first instance of life was likely very simple and single-celled. Today, life is fantastically varied and complex. We separate the life we see today into three domains based on shared cellular structure and genetic material. The three domains of life, archaea, bacteria, and eukaryotes, are a diversification of the Last Universal Common Ancestor (LUCA).

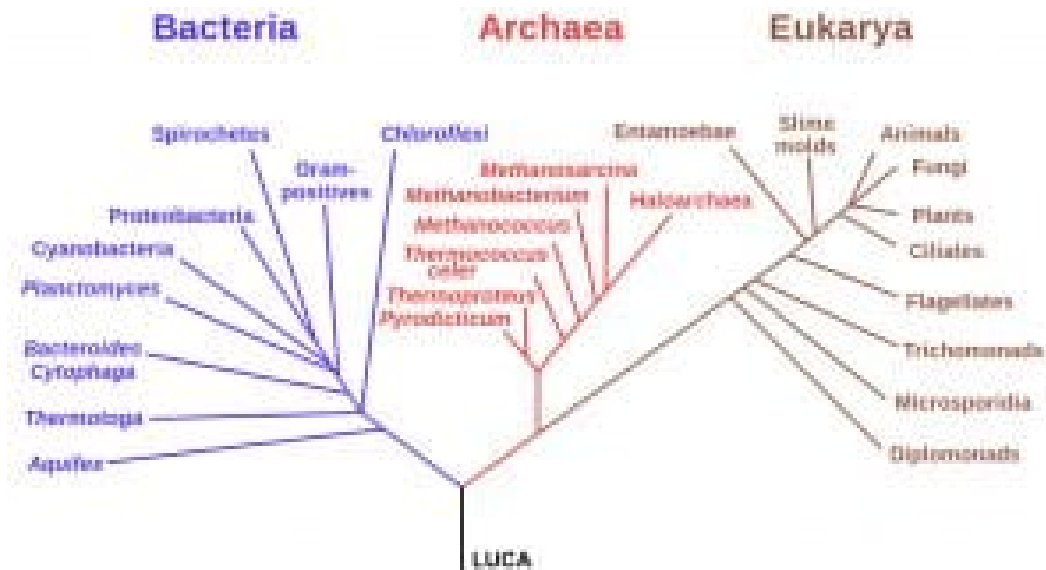


Figure 1 – The phylogenetic Tree of Life. The Tree of Life shows genetic relationship between all life on Earth. The stem represents the concept of the Last Universal Common Ancestor (LUCA).

- **Bacteria** make up the largest domain with both the greatest number of individual species and a biomass that exceeds the combined biomass of all plants and animals. Bacteria were one of the earliest forms of life on our planet and they are found in most habitats. Bacteria can rapidly recombine their genes with other bacteria to allow for genetic innovations, such as resistance to antibiotics.
- **Archaea** are genetically distinct from bacteria with their own, separate domain. Archaea are known to generate energy in a variety of ways and have been found in some of the most extreme environments on Earth. **Halophiles**, for example, exist in extremely salty conditions.
- **Eukaryotes** are distinct from **prokaryotes** in containing a central nucleus enclosed in a membrane and also contain other membrane-bound organelles. Examples of eukaryote organelles include chloroplast, the site of photosynthesis in plants and some algae, and mitochondria, where energy is generated in a cell. Eukaryotes can be uni- or multi-cellular, allowing for larger and more complex organisms.

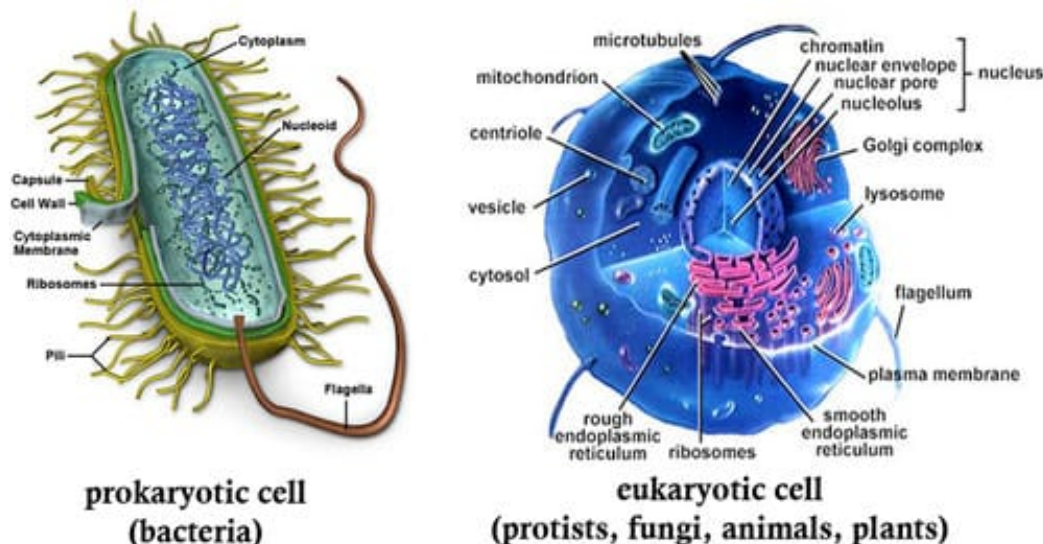


Figure 2 – A comparison of eukaryotic cells to prokaryotic cells. Note how the nucleoid, or DNA, of the prokaryotic cell is free floating while the DNA of the eukaryotic cell is contained within a nuclear membrane. While both have certain organelles, such as ribosomes, only eukaryotes possess membrane-bound organelles like mitochondria.

Both bacteria and archaea are prokaryotes: single-celled microbes that do not contain membrane-bound organelles. However, the membranes in archaea incorporate a different type of lipid than either bacteria or eukaryotes. Genetic analysis reveals that

archaea are closer to eukaryotes in an evolutionary sense than they are to bacteria. Archaea and eukaryotes use many of the same enzymes for DNA translation.

The Diversification of Life

Endosymbiosis

Eukaryotic cells have gained membrane-bound organelles and increased complexity through the process of endosymbiosis. Endosymbiosis is a process whereby primitive organisms benefited by living inside other organisms. Chloroplasts and mitochondria are examples of highly complicated organelles in eukaryotes that have their own membrane. They retained the DNA, messenger RNA, transfer RNA, and ribosomes of their bacterial ancestors before they were symbiotically incorporated into larger eukaryotic cells. The larger cell presumably provided protection and easy access to organic molecules while the chloroplast and mitochondria provided energy to the larger cell. This beneficial relationship led to the creation of larger, more efficient cells.

Want to know more: Mother's Mitochondria

With sexual reproduction, the offspring ends up with a combination of the mother's and father's DNA. However, the DNA found in mitochondria is exclusively the mother's DNA. When cells are replicated, the mitochondria split themselves as needed and so maintain a self-consistent set of DNA. With mammals, the egg destroys most of the mitochondria in sperm when they merge. In addition, most of the sperm's mitochondria is positioned in the tail to provide energy and does not make it into the egg.

This preservation of maternal mitochondrial DNA is often used to trace ancestries. Because it is contained only in the mitochondria and is infrequently used, mitochondrial DNA also suffers fewer mutations. It is therefore also helpful in determining the ancestry of different species and how they fit onto the tree of life.

Cyanobacteria

The rise of oxygen likely began with organisms known as cyanobacteria, an early type of photosynthetic bacteria thought to be the first organism to produce oxygen as a byproduct. Photosynthesis is the process by which organisms can harness the energy of the sun to generate energy for their own use. Cyanobacteria are the only known prokaryotes to produce oxygen, and this adaptation brought about the destruction of many other organisms.

Want to know more: Oxidation Reactions and the Free Radicals

Oxidation describes a chemical process in which a molecule, atom, or ion loses an electron. Oxygen is a particularly good oxidizing agent because its nucleus strongly attracts electrons to fill its valence shell. Oxidation often produces free radicals, which are very reactive. They can attack and break apart bonds in other molecules, inciting a chain reaction that is damaging for biochemical reactions.

Antioxidants are compounds that inhibit oxidation and thereby prevent the formation of free radicals. Plants and animals have many natural antioxidant systems in place to guard against this or use antioxidant vitamins such as vitamin A, vitamin C, and vitamin E. A certain amount of antioxidants is required in a well-balanced diet. However, clinical studies have been unable to prove benefits of artificially increasing antioxidant intake or antioxidant supplements.

The rise in atmospheric oxygen was far from smooth and steady (see Figure 3 below). Cyanobacteria slowly became more abundant, but there was still a significant delay in the build up of oxygen. There are several processes that would have hindered the rise in oxygen. Oxygen would have reacted with various chemicals, mainly iron, in the oceans, and these reactions would trap oxygen, prevent it from building up the atmosphere.

image

Figure 3 – The amount of oxygen in the atmosphere as a function of billions of years ago (Gyr). While the start of photosynthesis corresponds to an increase in oxygen, there is a notable plateau before oxygen jumps up to current day levels.

Oxygen may also have been taken up by microbes in metabolic pathways that generate energy. Organisms that used the oxygen

to oxidize ammonia appear to have been plentiful at the time and could have been effective in the reduction of free-floating oxygen.

Other organisms produced methane as a byproduct that could have acted as a sink for atmospheric oxygen. However, organisms that are known to produce methane require nickel to carry out the necessary reactions, and concentrations of nickel were dropping. The decrease of nickel would mean less excreted methane, and would allow oxygen to begin accumulating.

The increase of oxygen in the atmosphere was poisonous for anaerobic organisms. Most life before the Great Oxidation Event was anaerobic, so the rise of atmospheric oxygen surely resulted in one of the most significant extinction events in Earth's history. The fossil record shows a mass extinction of anaerobic life around 2.4-1.6 billion years ago, coincident with the rise of aerobic life.

Concept Check

Why didn't oxygen build up in the Earth's atmosphere immediately after cyanobacteria began producing it (around 3.2 billion years ago)? Give two reasons and explain each.

Anaerobic vs Aerobic Life

The rise of atmospheric oxygen coincided with the appearance of far more complicated life forms. There is every reason to believe that the rise in oxygen would have been responsible since aerobic metabolisms are more efficient.

Respiration, in biological terms, describes the process by which organisms convert nutrients into usable energy by forming ATP bonds. Respiration begins with glycolysis, wherein glucose, a sugar, is broken down to form two molecules of a compound called pyruvate along with two molecules of ATP. Without oxygen, organisms have no choice but to undergo anaerobic respiration, or fermentation. Anaerobic respiration of yeast is what makes bread rise and beer bubbly. In this scenario, pyruvate is shuttled down a pathway which produces just two molecules of ATP.

Side Note: A byproduct of anaerobic respiration is a molecule called lactic acid. Lactic acid should be very familiar to any athlete or any student who has been very late to class and had to make a run for it. When our bodies overexert themselves, we begin to use up more oxygen than we can take in. In order to produce the energy needed to keep running, the body switches to anaerobic respiration, which leads to a build up of lactic acid. Lactic acid can damage muscle cells and hinder recovery. The process of producing this lactate is also the cause of next-day muscle soreness.

Concept Check

What are the advantages of aerobic respiration for more complex life?

In the presence of oxygen, aerobic respiration becomes possible. With the help of oxygen, pyruvate can be broken down and enters a more complicated pathway known as the Krebs cycle or the citric acid cycle (note that organisms exhibit an enormous variety of metabolic pathways; the citric acid cycle merely represents one of the more common and well understood pathways). Through the citric acid cycle, organisms can produce from 30-36 ATP from just one molecule of glucose. Though more complicated to assemble, and therefore likely taking longer to evolve, this process can be up to 18-fold more efficient than anaerobic respiration. With more energy, it is possible to carry out more biochemical processes. This may have allowed cells to become increasingly complex and trend towards the more varied, multicellular life we see today.

Want to know more: Viruses

Viruses are even more abundant than bacteria. A virus consists of three functional parts: (1) genetic material, (2) protein coat, and typically (3) an envelope of lipids outside the protein coat. For genetic material, viruses have been discovered to use both DNA and RNA. The protein coat, also known as the capsid, encases and protects this genetic material. The envelope of lipids adds an additional layer of protection.

However, the question of whether viruses can be considered life is hotly debated. Viruses have their own genetic material and

are even capable of evolving through natural selection. Viruses survive and replicate by infecting a host cell. After attaching to a cell, the virus injects its DNA into the host cell. Now, the virus is capable of taking over the host cell's replication mechanisms. In doing so, the virus can now create copies of itself until it has exhausted the cell's resources. Millions of viruses can be made in this period before the cell dies and viruses escape to infect new host cells. Because viruses require the metabolism of a host cell to produce energy and reproduce, they can not be called self-sustaining. In this way, viruses fail the NASA definition of life.

Despite their differences, evidence of viruses appear wherever life does, suggesting that viruses and life evolved together. The history of viruses is traced through their DNA or RNA and has given rise to three theories on how viruses came to be.

Regressive evolution theory proposes that viruses were once components of small, parasitic cells. Similar to the idea of endosymbiosis, viruses may have begun as small structures within larger cells that over time became separated. In fact, there exist today some bacteria that, like viruses, can only reproduce in a host cell. As they evolved, these small parasitic structures regressed further from cell-like characteristics, becoming the viruses we see today. However, there is no evidence of types of cells today that could serve as an intermediary between early and present day viruses. Even the smallest cellular parasites fail to really resemble viruses at all.

Another theory is the **escaped gene theory**, which proposes that viruses got their start as DNA or RNA that escaped from the genome of a larger organism. Surprisingly mobile DNA has recently come to play a large role in biology. Plasmids (Figure below) are circular units of DNA and separate from the genome of an organism. They are most commonly found in bacteria and have been known to move between cells. Scientists have also recently discovered transposons, or "jumping genes," which are large segments of DNA that can move around within a cell's genome. While either of these mechanisms could have provided the genetic material for viruses, it remains unclear where the complicated capsids enclosing these genes arose from.

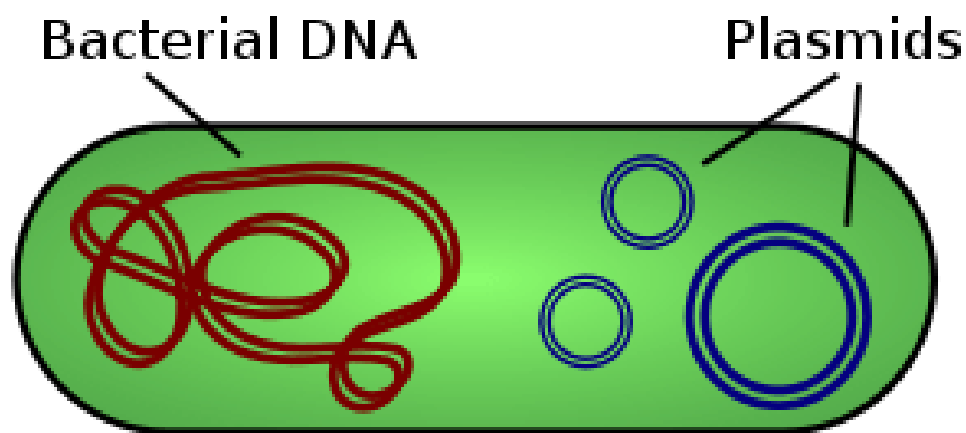


Figure 4 – Plasmids, pictured as small, blue loops. Being much smaller than chromosomal DNA, plasmids can be transcribed more easily and move from one organism to another. Credit:

The **pre-cellular origin theory** proposes that viruses formed from complex proteins that spontaneously arose even before RNA. Viruses could have evolved alongside the first cells or even before. However, this fails to account for the dependency viruses have on host cells. If viruses need the mechanism of a host cell to survive, how did they manage before cells even evolved?

The exact origin of viruses, much like the origin of life, remains an open question. This vein of research is being pushed forward through analysis of viral and host DNA sequences. What we can say is that genetic comparisons show that the origins of viruses may have predated life splitting into the three different domains. Perhaps the precursors to viruses evolved from the self-replicating molecules that dominated the RNA world. Similar to RNA, viruses are capable of self-assembling in host cells.

Viruses are now well enough understood to be used as tools in biomedical research. In a process called gene therapy, researchers use viruses to inserting genes into specific cells, offering possible treatments for diseases like cystic fibrosis. Some viruses will seek out and destroy cancer cells, while leaving healthy cells alone, allowing for a highly targeted and effective treatments.

Key Concepts and Summary

All life on Earth shares common characteristic: it is carbon-based, dependent on water, has left-handed amino acids, uses ATP for energy storage and transport and has an inherited genetic code for building proteins to carry out processes in the living cell. These shared attributes are distilled into the concept of the Last Universal Common Ancestor at the root of the Tree of

Life. There are three phylogenic branches on the Tree of Life: bacteria, archaea, and eukaryotes. The most complex organisms are eukaryotes, which evolved in part by the process of endosymbiosis to incorporate compartmentalized prokaryotic cells. The first cells were anaerobic – using chemical gradients for energy – and mutated into cells that used the energy of the sun for aerobic respiration. Aerobic respiration is 18 times more efficient in producing ATP than anerobic processes. Photosynthesizing cyanobacteria produce oxygen as a byproduct and are thought to be responsible for the rise of an oxygen-rich atmosphere on Earth.

Review Questions

1. What are the three domains on the tree of life and how do scientists distinguish between them?
2. What are the main differences between prokaryotes and eukaryotes?
3. Give an example of endosymbiosis and explain what advantages and disadvantages were gained by organisms that evolved from such situations.
4. How did cyanobacteria cause one of Earth's mass extinctions?
5. What causes aerobic metabolism to be more efficient than anaerobic metabolism?

Exercises

1. Four examples of eukaryotic organelles that evolved from endosymbiosis include (1) choloroplasts, (2) mitochondria, (3) nitroplasts, and (4) diazoplasts. Conduct the necessary research to construct a table that compares these four examples in separate rows with three different columns: (1) the kinds of organisms that contains these organelles, (2) the approximate timeframes for when the incorporation of these organelles occurred, and (3) explanations of the function of these organelles.

3.4: Development of Life on Earth

Billions of years ago the Earth was a very different place. Life seems to have emerged almost as soon as conditions were favorable after the planet had cooled and the environment allowed for the prebiotic chemistry to happen. As life proliferated, it had a profound effect on our planet affecting the environment, climate, and composition of our atmosphere while leaving behind a fossil record of its presence. In this chapter, we will explore the development of life and what evidence we have for how these changes occurred across geologic time.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss the evidence for where and when the first life on Earth may have arisen
- Describe hydrothermal vent communities
- Explain how oxygen became a key component of the Earth's atmosphere and life processes.
- Describe the different geological timeframes and how they relate to the development of life on Earth

Where did life arise on Earth?

It is unclear whether the transition from chemistry to biology occurred on the land surface of Earth or in the oceans. Those intermediate species on our planets – entities that are more than chemistry, but less than life – no longer seem to exist.

In the mid-1960s, Alexander Graham Cairns-Smith proposed that complex prebiotic molecules may have first organized around the regular crystal patterns found in minerals in clay. Cairns-Smith hypothesized that electrostatic forces in mineral crystals would help to concentrate and align specific molecules to their surface. With the help of this clay, precursors of RNA could have assembled, triggering the start of an RNA world. Tide pools have also been proposed as the site where chemistry developed the complexity needed for life. The cycles of evaporation and subsequent addition of water in tide pools might have concentrated the organic materials necessary to form life. Higher concentrations increase the likelihood that these organic materials will react and form RNA or lipids.

It is possible that, far from the rocky surface of Earth, life originated deep in the oceans when our planet was a frozen world. During the phases where Earth was covered by a thick mantle of ice, the ice may have protected organic compounds beneath the ice from the damage of impacting asteroids or ultraviolet radiation from the sun.

Deep-sea hydrothermal vents are hot spots that release gas from the Earth's interior. DNA sequencing suggests that LUCA was a thermophilic microbe and deep sea vents might have been ideal locations for nurturing that early life. At first glance, this seems an inhospitable place for life. These vents are located in some of the darkest, highest pressure environments on our world. However, the gas that is released from hydrothermal vents creates a natural chemical gradient that can be used by living cells to generate the energy needed for metabolism. In deep hydrothermal vents, this chemical gradient flows around the rocks surrounding the vent and creates nooks where organic compounds can collect and reach higher concentrations. Minerals that are capable of acting as catalysts have been found in deep sea vents. Therefore, hydrothermal vents seem to provide the needed ingredients: high concentrations of organic compounds, natural catalysts, and a powerful energy source in the absence of sunlight. Today, extremely varied ecosystems are found around these vents.

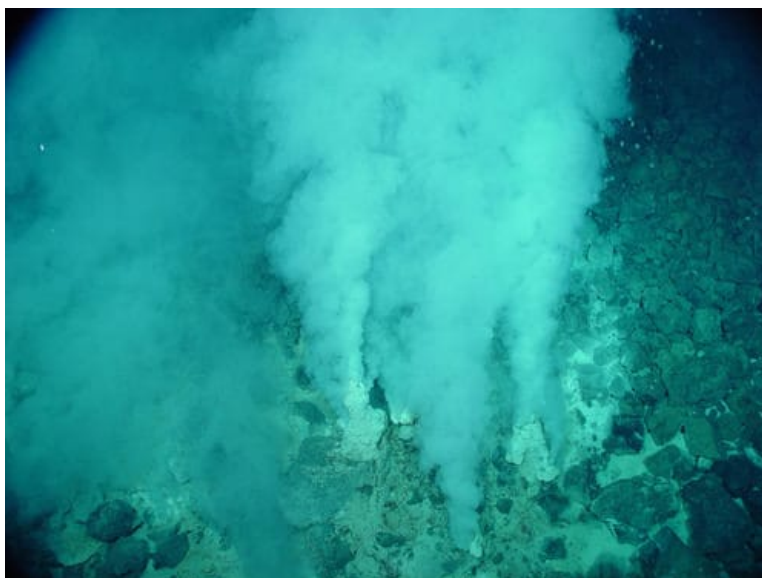


Figure 1 – A deep-sea hydrothermal vent. These vents release a cocktail of varied chemicals and ions that contain the starting materials for life and create a helpful chemical gradient that can be used as a source of natural energy.

When did life arise on Earth?

The evidence for timing the initial rise of life on our planet is ambiguous. The most ancient evidence has been destroyed as the first forms of life would have been single-celled organisms that did not leave behind fossil remains. However, suggestive evidence from biology, chemistry, and geology supports estimates that life arose on Earth between 3 to 4 billion years ago. Since the Earth formed about 4.56 billion years ago, the early appearance of life hints that the evolution from chemistry to life may be statistically probable. If so, this greatly increases the odds that life has also evolved elsewhere.

Carbon isotopes

The earliest line of evidence for timing the rise of life comes from carbon isotopes. Carbon naturally occurs in three isotopes, always with 6 protons, but with either 6, 7, or 8 neutrons, which are annotated ^{12}C , ^{13}C , and ^{14}C respectively. The superscripts indicate the atomic mass number, or the total number of protons and neutrons in an element. Of these isotopes, ^{14}C will undergo radioactive decay with a half life of 5730 years. However, ^{12}C and ^{13}C do not decay and therefore, the ratio of ^{12}C to ^{13}C is constant over time on the Earth. However, life preferentially incorporates the lighter ^{12}C , rejecting ^{13}C . Therefore, a low ratio of ^{13}C to ^{12}C offers a tentative suggestion that organic material existed and that isotopic imbalance can be incorporated into the metamorphic structure of the rock.

Why does life prefer to use ^{12}C ? While the chemical reactivity of an atom is predominantly driven by the electron configuration, more massive isotopes tend to have slightly slower reaction rates. That slight edge in the speed of reaction rates is enough of an advantage to favor the uptake of carbon-12 over carbon-13 in organic biochemistry.

Zircon crystals are commonly used to assess ages in very ancient geological records because they are very durable minerals, resistant to both heat and corrosion. Trapped minerals can be preserved when zircons form. A group of scientists from the University of California, Los Angeles studied 10,000 zircons gathered from western Australia. Of these ten thousand zircons, one contained a graphite inclusion, a compound composed entirely of carbon atoms.

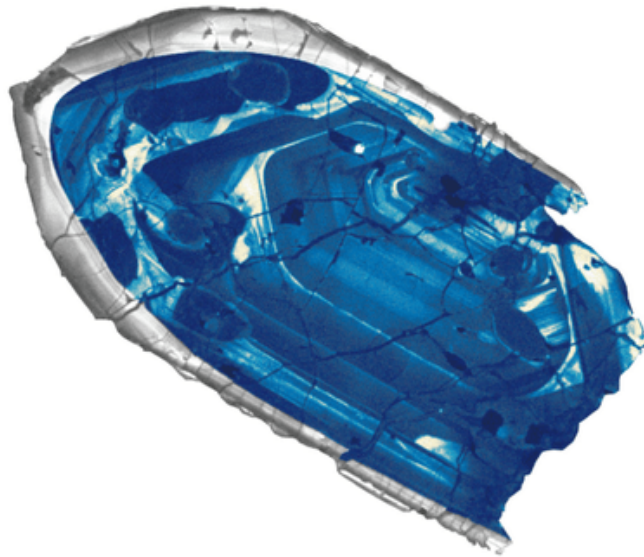


Figure 2 – A hearty zircon crystal. Zircon crystals are not only capable of surviving through great stressors, but they are also good at keeping elements trapped inside them. Why are both of these properties important for accurate radiometric dating?

Radiometric dating showed that this zircon crystal was 4.1 Gyr old and the carbon inclusion exhibited a larger ratio of ^{12}C to ^{13}C . [Was this the chemical stain of primordial life](#), dating back to 4.1 Gya? On Earth today, this would be a good indicator of organic material. But, interpreting this result from so long ago is more controversial.

Stromatolites

Ancient stromatolites offer a more secure timeline for the emergence of life. These fossilized structures are found in shallow waters and look like modern structures formed by cyanobacteria. The oldest stromatolites have been found in South Africa and Australia and date back to the early Archean Period between 3.2 – 3.5 Gya. By the end of the Archean and throughout the Proterozoic geological periods, stromatolites appear to have been abundant and formed the first reefs.

Figure 3 – Fossilized stromatolites. This distinctive banding pattern resembles structures produced by living organisms today. This suggests that similar organisms existed at the time these fossils were formed.

Stromatolites today are composed of layers of microbial mats of photosynthesizing cyanobacteria, as shown in the YouTube video below. Photosynthesis depletes carbon dioxide in the water, precipitating calcium carbonate deposits, which along with other sandy sediments, is trapped in the sticky bacterial film. The bacterial colonies grow upward towards better exposure to the sun, and over time the layers of bacterial film and mineral precipitates build up to form distinctive layered stromatolites. Similar stromatolite structures are formed today.

Stromatolites



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=94>

Video credit: @jamief94086

We guess that ancient stromatolites were formed through biological processes, since we observe this behavior in microbes today. The processes required to make these structures seem to require microbes that are already quite complex and capable of photosynthesis. Actual first life then, likely to be a more primitive organism, would have existed even earlier than these ~3.5 Gyr old structures.

Microfossils

The most direct evidence for life comes in the form of microfossils, fossils preserving micro-organisms that may have been among the first living creatures on Earth. The most convincing microfossils date back to approximately 3 billion years ago. However, these are difficult to identify both because rocks undergo erosion over time and because the structures that resemble microfossils might actually be formed by nonbiological processes. Many supposed microfossils turn out to be false positives after careful chemical analysis. Though the discovery of a microfossil is an unmistakable mark of life, it is easy to be fooled. The Figure below gives you a good idea of the scant information content in microfossils.

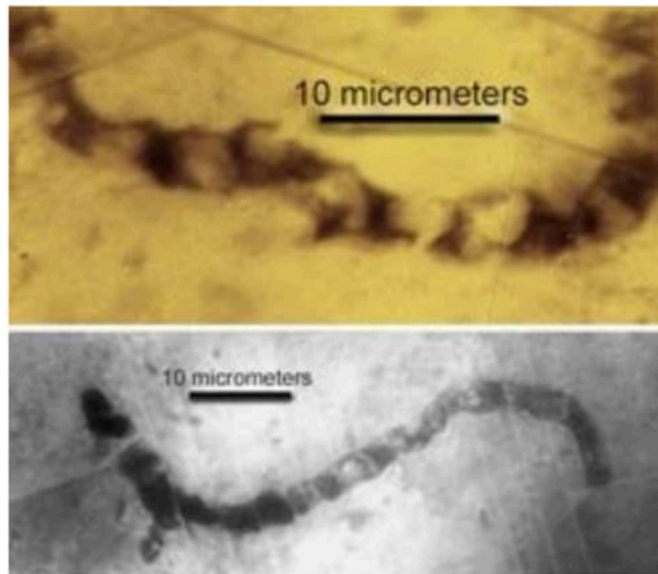


Figure 4 – An example of possible microfossils. A micrometer is a millionth of a meter. Human hair is on average 80 micrometers thick.

The Age of Life

Evolution and Extinction

On geological time scales, time is organized into eons, eras, and periods. There are four main eons subdivided into different eras. The eras are in turn split into different periods. The current eon, the Phanerozoic Eon (541 Mya through today), is known as “The Age of Life.” The Cambrian Period is the earliest period of the Phanerozoic eon, lasting from 540 to 500 Mya. The events leading up to this period set the stage for an explosion of complex life. Changing atmospheric oxygen concentration allowed for more efficient aerobic metabolism and increasingly complex life forms.

Concept Check

What defines the divisions between different eras or periods in the Phanerozoic?

The transitions between different geologic periods are defined by changes in the fossil record with many of the most dramatic changes occurring during mass extinction events, defined as a period of time when the rate of species going extinct is increasing with respect to the rate at which new species arise. Throughout the Phanerozoic era, there were 15 documented major extinction events. Five of these were major mass extinction events where more than half of all species on Earth at the time were lost. While mass extinctions are devastating, they do serve to encourage diversity. Had the dinosaurs not been eliminated 65 Mya, there may never have been such a successful emergence of mammals. The frequency of life in the Universe is a function of how often it arises, how often it survives, and how long it endures after evolving. In this way, the frequency and intensity of mass extinctions control the frequency of life.

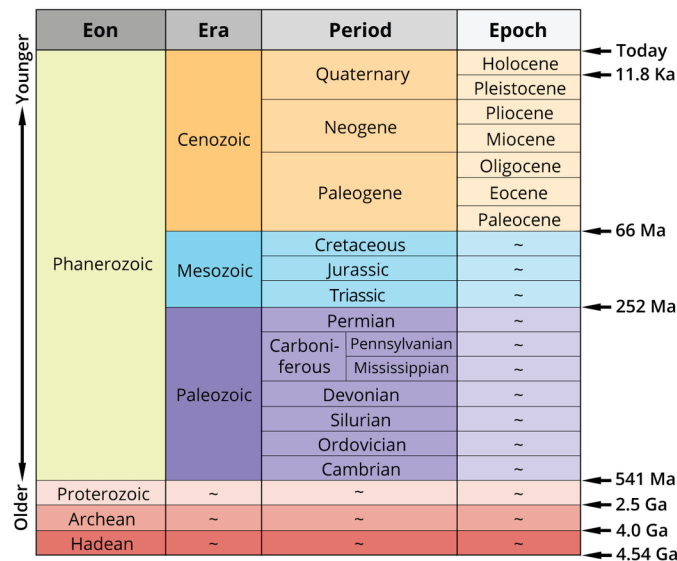


Figure 5 – The geological time scale. The time labels on the right are in units of “annum” (a), where Ga stands for Giga annum or a billion years.

Precambrian Life

The Proterozoic Eon began 2.5 billion years ago and set the stage for Phanerozoic life. From the geologic record, we know that oxygen levels were low before the Proterozoic. Therefore, microbes from the start of this era would have primarily been anaerobic chemotrophs, producing energy through chemical pathways.

Want to know more: Metabolic Lifestyles

Organisms can be identified according to the source of carbon they use for metabolism as well as their energy source. The prefixes auto- (“self”) and hetero- (“other”) refer to the origins of the carbon sources various organisms can use. Organisms that convert inorganic carbon dioxide (CO₂) into organic carbon compounds are **autotrophs**. Plants and cyanobacteria are well-known examples of autotrophs. Conversely, **heterotrophs** rely on more complex organic carbon compounds as nutrients; these are provided to them initially by autotrophs. Many organisms, ranging from humans to many prokaryotes, including the well-studied *Escherichia coli*, are heterotrophic. All pathogens are heterotrophic because their carbon source is their host.

Organisms can also be identified by the energy source they use. All energy is derived from the transfer of electrons, but the source of electrons differs between various types of organisms. The prefixes photo- (“light”) and chemo- (“chemical”) refer to the energy sources that various organisms use. Those that get their energy for electron transfer from light are **phototrophs**, whereas **chemotrophs** obtain energy for electron transfer by breaking chemical bonds. There are two types of chemotrophs: organotrophs and lithotrophs. Organotrophs, including humans, fungi, and many prokaryotes, are chemotrophs that obtain energy from organic compounds. Lithotrophs (“litho” means “rock”) are chemotrophs that get energy from inorganic compounds, including hydrogen sulfide (H₂S) and reduced iron. Lithotrophy is unique to the microbial world.

The strategies used to obtain both carbon and energy can be combined for the classification of organisms according to nutritional type. Most organisms are chemoheterotrophs because they use organic molecules as both their electron and carbon sources. Table 1 below summarizes this and the other classifications.

Table 1: Classifications of Organisms by Energy and Carbon Source

Classifications		Energy Source	Carbon Source	Examples
Chemotrophs	Chemoautotrophs	Chemical	Inorganic	Hydrogen-, sulfur-, iron-, nitrogen-, and carbon monoxide-oxidizing bacteria

Classifications		Energy Source	Carbon Source	Examples
	Chemoheterotrophs	Chemical	Organic compounds	All animals, most fungi, protozoa, and bacteria
Phototrophs	Photoautotrophs	Light	Inorganic	All plants, algae, cyanobacteria, and green and purple sulfur bacteria
	Photoheterotrophs	Light	Organic compounds	Green and purple nonsulfur bacteria, heliobacteria

Source: [LibreTexts Biology](#)

The Proterozoic Eon was likely a time of rapid diversification of life. Today, there are systems of proteins that run checks during the replication of DNA to minimize transcription errors. However, these types of proteins are less common in prokaryotic microbes, which were common during the Proterozoic Era. A higher rate of mutations would have accelerated natural selection and driven a greater diversification of organisms. One of those mutations enabled photosynthesis, a more efficient process of respiration, resulting in an explosion of complex multi-cellular organisms that appeared at the start of the geologic era known as the Cambrian Explosion.

The Cambrian Explosion

At the time of the Cambrian explosion, there was an exponential increase in the number and complexity of organisms. Complex organisms likely preferred living in the oceans at first, where there was easy mobility and easy access to nutrients. Even very shallow levels of water provide protection from damaging UV radiation.

The earliest evidence of complex, land-based life comes from early land plants that begin to appear in the fossil record around 475 Mya. DNA evidence suggests that these plants evolved from algae, a general term for aquatic, photosynthetic organisms. It is a likely transition given that many breeds of algae prefer growing in shallow water to begin with. Algae adapted protections, such as thicker cell walls in order to survive dry spells, that would have been beneficial in the transition from water to land.

Figure 6 – Fossils of the earliest land plants. Early land plants were small, simple, and likely remained close to water. Having yet to develop vascular systems, they more likely resembled moss than what we traditionally associate with plants today.

With other organisms still constrained to the water, land plants flourished and grew large in size. The lack of competition further allowed for increasing complexity: the development of root systems and tubes to transport water and nutrients throughout the plant. Photosynthesizing land-based plants had abundant energy from the Sun.

The Carboniferous Period

Forests began to cover the lands and dead organic matter began to pile up and decompose among the plants. Millions of years later, this organic matter would be compressed into coal, earning this period the name “carboniferous,” which means “coal carrying” in Latin.

The Carboniferous period began roughly 360 Mya and lasted about 60 million years. It was marked by an era of giantism. The fossil record from early in this period is rich in ocean and fresh water organisms. Land-based organisms appeared in the middle to later part of the period. The first terrestrial organisms were supersized insects and amphibians. By the middle of the Carboniferous period, adult amphibians were up to 6 meters long and developed the scaly skin typical of lizards by the end of the period. The Arthropleura, which resembles a 2.6-meter long millipede, proliferated during the late Carboniferous period and remains the largest-known land invertebrate ever discovered.

The Carboniferous period also gave rise to the largest-known flying insect: the Bolsover dragonfly. The Bolsover dragonfly looks nearly identical to the dragonflies common today, but with a wingspan of 30 cm – six times larger than a modern dragonfly. With a wingspan this large, the Bolsover dragonfly would not be able to fly in today’s atmosphere, with 21% oxygen. This has been used to suggest that atmospheric oxygen levels may have been as high as 30% during the Carboniferous period. Increased levels of oxygen as well as a generally moister environment would explain the large size of organisms at the time. Respiration was likely more efficient so that organisms could produce more energy and sustain their larger sizes.



Figure 7 – A fossil of the Bolsover dragonfly, also known as the Meganeura. With a wing span of 30 cm (or nearly a foot), please read the “mega” in meganeura with extra emphasis. Credit: <https://www.geologypage.com/2020/03/...dragonfly.html>

As life continued to evolve and diversify, ecosystems began to resemble the types of settings we are familiar with today. This penchant for giantism continued into the Mesozoic era, which spanned from 252 to 66 Mya and is known as “the Age of the Dinosaurs.” The Cenozoic era that followed saw the emergence of mammals. The end of each period during the Phanerozoic Eon was marked by mass extinction events that completely upended the biological landscape of the time.

Great Oxidation Event

The Great Oxidation Event is alternatively called the Oxygen Catastrophe depending on whether you ask aerobes or anaerobes. It marks a drastic change in the Earth’s atmospheric composition that occurred about 2.5 Gya.

An increase in atmospheric oxygen is seen in the geological record as a sudden onset of oxidation of iron occurred. Iron oxidation coincides with the geological evidence for glaciation 2.5 billion years ago. Oxygen is not a greenhouse gas, but it is a highly reactive species that would have interacted with the primitive methane atmosphere produced by volcanic outgassing. Methane is a powerful greenhouse gas. As oxygen levels rise, oxygen would react with methane to form carbon dioxide and water. This mechanism for removing methane would have produced the significant cooling and widespread glaciation. An outcrop of ancient Canadian rocks shows evidence of a glaciation event between different rock layers. Below (and therefore older than) this glaciation event, the rocks are consistent with low levels of atmospheric oxygen. Younger rock above the glaciation layer show significantly higher oxidation. Radiometric dating of the rock show that the increase in atmospheric oxygen, or the Great Oxidation Event, occurred 2.45 billion years ago.

The addition of oxygen to the atmosphere is believed to be the result of photosynthesis from cyanobacteria. Oxygen did not accumulate immediately because of the enormous number of sinks, including reduced gases and minerals that would have overwhelmed its production. Those sinks for oxygen are now largely saturated; respiration of anaerobic organisms (including us) and decay of organic material take up most of the oxygen produced today.

As the atmosphere cooled, a positive feedback loop ensued. Water vapor, which is another important greenhouse gas, would have condensed out of the atmosphere. The Earth eventually froze over and became what is known as “snowball Earth.” Volcanic activity would have continued during this phase, releasing internal heat along with carbon dioxide into the air. With a sufficient amount of CO₂, the greenhouse warming would have melted the ice, allowing the Earth to recover from its frozen state. Assuming present rates of volcanism, the necessary build up of CO₂ would take 10 million years. Evidence in the rock record suggests that one snowball Earth event occurred around 2.2 billion years ago and another series of snowball Earths that happened just prior to the Cambrian Explosion.

Figure 8 – An artist rendition of how the Snowball Earth may have appeared. The extent of glaciation would have been worldwide, causing a thick layer of ice from pole to pole. Credit:

Key Concepts and Summary

The deep oceans may have been one of the most stable environments on Earth, immune to the wild swings of conditions on the surface of the planet. This has led to speculation that the first life may have emerged in deep-sea thermal vents with strong chemical gradients to energize biochemistry. The evidence for life has been captured in the geologic record and unsurprisingly is more ambiguous the farther back in time we go. There is circumstantial evidence for life in the ratio of carbon isotopes in outcroppings of rocks and in zircon crystals that are about four billion years old. Fossilized layered structures date back to about 3.2 billion years ago and are reminiscent of stromatolite mats that are abundant today. Microfossils that are 2 – 3 billion years old and have structures that appear to be imprints of microbial life. After the Cambrian explosion 540 million years ago, the geologic record becomes clear, recording the rise and fall of at different spatial depths that correspond to different look-back times.

Review Questions

Summary Questions

Exercises

3.5: Extremophiles

At a chemical level, life consists of many types of molecules that interact with one another to carry out the processes of life. Life also needs an environment in which those complicated molecules are stable (don't break down before they can do their jobs) and their interactions are possible. Your own biochemistry works properly only within a very narrow range of about 10°C in body temperature and two-tenths of a unit in blood pH (pH is a numerical measure of acidity, or the amount of free hydrogen ions). Life overall must also have limits to the conditions in which it can properly work but, as we will see, they are much broader than human limits. Let's have a look at some of the conditions that can challenge life and the organisms that have managed to carve out a niche at the far reaches of possibility.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe some of the most extreme conditions found on Earth
- Discuss different types of extremophiles that have been found on Earth
- Discuss other worlds where we might expect to find extremophiles
- Explain why tardigrades are useful in understanding extreme life

Extreme Conditions

Extremophiles are organisms that live in habitats that seem extreme to humans (the suffix *-phile* means “lover of”). The ability of organisms on Earth to live in a wide range of environments is the secret to the survival of life on this planet and extremophiles help us to understand the limits to life. On the early Earth, the large swings in climate would have cycled between hot and freezing conditions. Before microbes invented photosynthesis, life adapted to take advantage of many different metabolic pathways. There is evidence in the geologic record that oxygen levels in the atmosphere fluctuated wildly before a precipitous rise 2.4 billion years ago. Through every [mass extinction](#), there were some niches of life that survived. Even our flammable, oxygen-rich atmosphere constitutes an extreme environment.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=99>

Link to Part 2 of video: <https://nasaclips.arc.nasa.gov/video/ourworld/our-world-where-do-we-find-extremophiles>

Video Credit: @NASAeClips

Being “extreme” is in the eye of the beholder. Most of the organisms that humans identify as extremophiles are small prokaryotes in the archaea or bacteria domain on the Tree of Life. They are known to survive in a wide range of extraordinary environments.

Temperature

Both high and low temperatures can cause a problem for life. As a large organism, you are able to maintain an almost constant body temperature whether it is colder or warmer in the environment around you. But this is not possible at the tiny size of microorganisms; whatever the temperature in the outside world is also the temperature of the microbe, and its biochemistry must be able to function at that temperature. High temperatures are the enemy of complexity—increasing thermal energy tends to break apart big molecules into smaller and smaller bits, and life needs to stabilize the molecules with stronger bonds and special proteins. But this approach has its limits.

Thermophiles

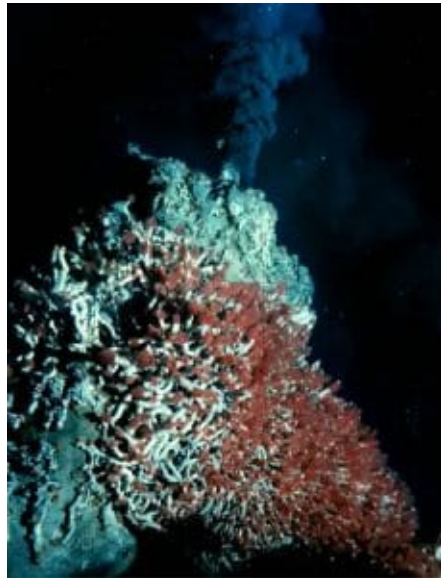


Figure 1 – A black smoker community on the seafloor comprised of giant red tubeworms and hundreds of squat lobsters. Credit: public domain

Thermophiles (temperature-loving organisms) live in high temperature environments, some at temperatures of 235°F (113°C). High-temperature environments like hydrothermal vents on the deep seafloor (Figure 1) and hot springs on land surfaces (Figure 2) often offer abundant sources of chemical energy and therefore drive the evolution of organisms that can tolerate high temperatures. Bacteria feeding on this chemical energy form the base of a food chain that can support thriving communities of animals—in the case shown in Figure 1, a dense patch of red and white tubeworms growing around the base of the vent. What appears to be black smoke is actually superheated water filled with minerals of metal sulfide.

High temperatures are challenging for life because the energy causes proteins to denature, or unfold. Chemical reactions proceed more quickly, and temperatures above 100°C can denature nucleic acids, causing DNA to lose its helical shape or break apart the structure of biomolecules. The fluidity and permeability of cell membranes are affected by heat. High temperatures decrease the solubility of carbon dioxide and oxygen in water, a problem for aquatic aerobic life. Thermophiles have metabolisms that take advantage of the higher chemical reactivity enabled by high temperatures. They make use of special temperature-resistant proteins and protective mechanisms for shielding DNA. Thermophile coping mechanisms include an altered ratio of saturated hydrocarbons in cell membranes.

The young Earth, with a tremendous amount of internal energy from accretion and differentiation, may have harbored many hot springs and ocean floor hydrothermal vents. Young Earth would have been a paradise for thermophiles, and the archaea and bacteria that are most deeply rooted in the phylogenetic tree of life are thermophiles. The range of metabolisms for thermophiles may be the result of the diversity of chemistry at deep sea hydrothermal vents.

Figure 2 – The Great Prismatic Spring of Yellowstone National Park. This hot spring provides a fascinating study of thermophiles. The rings of different colors represent different thermophiles that survive best at different temperatures. The closer to the water, which reaches a balmy 92°F, the greater the temperature these microbes can withstand.

Currently, the high temperature record holder is a methane-producing microorganism that can grow at 122°C, where the pressure also is so high that water still does not boil. That's amazing when you think about it. We cook our food—meaning, we alter the chemistry and structure of its biomolecules—by boiling it at a temperature of 100°C. In fact, food begins to cook at much lower temperatures than this. And yet, there are organisms whose biochemistry remains intact and operates just fine at temperatures 20 degrees higher.

Psychrophiles

Cold can also be a problem, in part because it slows down metabolism to very low levels, but also because it can cause physical changes in biomolecules. Psychrophiles are organisms that can withstand extreme cold. They are found in liquid brine inclusions in ice cores, or living under rocks in the extreme deserts of the world. This class of extremophiles includes bacteria and eukaryotes.

Concept Check

On Earth, psychrophiles can be found in glaciers, sea ice, deep ocean water and Arctic soils. Can you find specific locations where psychrophiles have been found? For example, psychrophiles have been found in subglacial waters of the Vatnajökull ice cap, Iceland.

In our solar system, what other worlds — planets or moons — would you expect to find psychrophiles? Be specific as to the location on that world where you would expect to find them and explain why.

Cells must be resilient to freezing, which can form ice shards that would pierce and destroy ordinary cells. Cell membranes—the molecular envelopes that surround cells and allow their exchange of chemicals with the world outside—are basically made of fatlike molecules. And just as fat congeals when it cools, membranes crystallize, changing how they function in the exchange of materials in and out of the cell. One defense of a psychrophile is to lower the freezing point with solutes in cytoplasm that essentially act as an antifreeze. To increase fluidity of cell membranes, psychrophiles have evolved a different ratio of unsaturated to saturated fats. At cold temperatures, proteins are more rigid, so enzymes are used to lower the activation energy for biochemical reactions. Thus far, the coldest temperature at which any microbe has been shown to reproduce is about -25 °C in Arctic permafrost.

pH



Figure 3 – Spain's Rio Tinto. With a pH close to 2, Rio Tinto is literally a river of acid. Acid-loving microorganisms (acidophiles) not only thrive in these waters, their metabolic activities help generate the acid in the first place. Credit: wikimedia

Acidophiles and alkaliphiles thrive in environments with very low or high pH, respectively. Conditions that are very acidic or alkaline can be problematic for life because many of our important molecules, like proteins and DNA, are broken down under such conditions. The most acid-tolerant organisms (acidophiles) are capable of living at pH values near zero—about ten million times

more acidic than your blood. Acidophiles have been found in conditions as acidic as battery acid. They have been able to adapt mechanisms for keeping the acid out so that the cell cytoplasm can have a neutral pH. Figure 3 shows Rio Tinto in Spain where acidophiles thrive; the rusty red color that gives the river its name comes from high levels of iron dissolved in the waters

At the other extreme, some alkaliphiles can grow at pH levels of about 13, which is almost a million times more alkaline than your blood. This is also comparable to the pH of household drain cleaner, which does its job by breaking down the chemical structure of things like hair clogs.

Salt

Halophiles can only survive in an environment with high concentrations of salt, an environment that would dehydrate most other organisms. High levels of salts in the environment causes a problem for life because the salt blocks some cellular functions. Humans recognized this centuries ago and began to salt-cure food to keep it from spoiling—meaning, to keep it from being colonized by microorganisms. Yet some microbes have evolved to grow in water that is saturated in sodium chloride (table salt)—about ten times as salty as seawater. On Earth, halophiles are found in salty lakes such as the Great Salt Lake in Utah or Lake Retba (Lac Rose or “Pink Lake”) in Senegal.

High Radiation

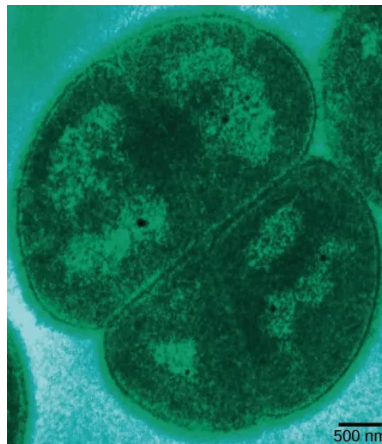


Figure 4 – *Deinococcus radiodurans* is a prokaryote that can tolerate very high doses of ionizing radiation. It has developed DNA repair mechanisms that allow it to reconstruct its chromosome even if it has been broken into hundreds of pieces by radiation or heat. Credit: OpenStax Biology

Life has even been found next to nuclear waste storage sites in the presence of enough radiation to grant super powers. Radiation can be very damaging to DNA and cause cancerous cells or tumors. These radioresistant organisms have evolved mechanisms to protect their DNA even in the presence of over 1000 times the radiation a typical organism can withstand. There is even an organism, *Deinococcus radiodurans*, that can tolerate ionizing radiation (such as that released by radioactive elements) a thousand times more intense than you would be able to withstand. It is also very good at surviving extreme desiccation (drying out) and a variety of metals that would be toxic to humans.

Life is everywhere on Earth

Many other adaptations to environmental “extremes” are also known. Endoliths are extremophiles that live inside of rocks. On Earth, endoliths have been found living inside rocks in deserts in Antarctica, Chile, and Namibia, as well as regions deep below the Earth’s surface, including inside a gold mine in South Africa that is more than 2 miles underground. Barophiles thrive in high-pressure environments. Very high pressures can literally squeeze life’s biomolecules, causing them to adopt more compact forms that do not work very well. But we still find life—not just microbial, but even animal life—at the bottoms of our ocean trenches, where pressures are more than 1000 times atmospheric pressure. There also still exist anaerobes that do not require oxygen and live deep underground. Xerophiles require very little water and are found in the soil of the world’s largest deserts.

From many such examples, we can conclude that life is capable of tolerating a wide range of environmental extremes—so much so that we have to work hard to identify places where life can’t exist. A few such places are known—for example, the waters of

hydrothermal vents at over 300°C appear too hot to support any life—and finding these places helps define the possibility for life elsewhere. The study of extremophiles over the last few decades has expanded our sense of the range of conditions life can survive and, in doing so, has made many scientists more optimistic about the possibility that life might exist beyond Earth. Extremophiles are a great boon to astrobiology both in demonstrating the wide variety of life that is possible, and by expanding the definition of what is “habitable.”

Earth Analogs in the Solar System

There is no other place in our solar system that today has the same hospitable conditions as Earth, with liquid water oceans on its surface today. However, we can learn from the extreme conditions on Earth that host life and identify similar environments on other planets or moons. Regions similar to frozen lakes and dusty deserts where we find extremophiles on Earth can be pinpointed elsewhere in the solar system, where a search for similar extremophiles in these regions holds promise.

Mars

Today, Mars is a barren world with a thin atmosphere dominated by CO₂. Photos of the surface Mars taken by the rovers show a landscape that looks similar to some deserts on Earth (Figure X). Can you tell which picture is of Earth and which is of Mars (before you read the caption)?



Figure X – A Mars analog on Earth. The photo on the left is the Jezero crater on Mars. On the right is a picture of the Red Stone Jurassic fossil delta in the Atacama Desert in Chile on Earth. Credit: NASA/JPL-Caltech; Atacama photo: Armando Azua-Bustos/Provided

The Atacama desert is the driest place on Earth so represents a reasonable analog for some places on Mars, such as the Jezero Crater. Life has been found within the extreme location of the Atacama desert, including halophiles, endoliths and radioresistant bacteria.

Europa and Enceladus

Jupiter’s moon Europa and Saturn’s moon Enceladus are both believed to harbor liquid water oceans beneath their icy surfaces. Several places on Earth serve as analogs for these cold, dark, high-pressure environments. Lakes buried beneath Earth’s surface, such as [Lake Vostok in Antarctica](#), have similar conditions to what we would expect for the oceans of Europa and Enceladus; searches for life within the overlaying ice or in the water of Lake Vostok and other subglacial lakes provide some clues as to what we could reasonably expect to find there. A rich variety of Hydrothermal vent communities, such as the “Lost City” at the Mid-Atlantic Ridge on the floor of the Atlantic Ocean, including alkaliphiles and microbes that use chemosynthesis to generate energy.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Tardigrades



Figure 5 – Tardigrades are hardy creatures that can survive in almost any environment.

Tardigrades, also known as water bears or moss piglets, are not technically extremophiles but are nevertheless extreme in their own right as well as strangely adorable. They are a type of micro-animal, only 0.5 mm in length on average, with eight-legs, typically found in water, and among the most resilient creatures known. They are different from extremophiles, which are adapted to live in terrifyingly harsh conditions, because they do best in average conditions. However, they are capable of surviving the most extreme conditions on this world, including low pressure environments in space.

The video below nicely describes how tardigrades can exist in a state known as anhydrobiosis (can you guess what this word means just based on the roots *hydro* and *biosis*?).

Meet the tardigrade, the toughest animal on Earth



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=99>

Video Credit: @TEDEd, Lesson by Thomas Boothby, animation by Boniato Studio.

In 2007, samples of tardigrades were taken into low Earth orbit and exposed to the vacuum and radiation of space for 10 days.

After the reanimation of the tardigrades back on Earth, most of the tardigrades began living as normal after 30 minutes (though most of the sample did suffer later health effects). Their incredible resilience allowed water bears to survive through five great extinction events on Earth. Their capability to survive in space gives new life to the theory of panspermia. Though not technically extremophiles, tardigrades certainly win a prize for resilience.

In April 2019, an Israeli spacecraft called Beresheet almost made it to the moon. The privately funded mission was the first stage of a privately funded project to transfer DNA to the moon to build a repository for rebuilding life in the event of catastrophic mass extinction. Among the passengers on the spacecraft: dehydrated tardigrades. It is unlikely that these creatures could survive on the moon without liquid water.

Side Note: Considering the fact that tardigrades can survive such extreme conditions, it may be surprising to know just how easy they are to collect and look at under a microscope. A team at LaGuardia Community College in the highly urban area of Queens, NY were able to collect tardigrades from moss and lichens on street trees and on the shore of a toxic waterway (a SuperFund site called Newtown Creek).

Key Concepts and Summary

Life on Earth has been identified in even the most extreme environments: very high (or low) acidity, temperatures, or pressures and even in high-radiation environments. The study of these extremophiles and the limits of extreme environments yields insight into the range of conditions for habitability that might exist on other worlds. The one requirement for life on Earth is the presence of water, but tardigrades (which are found on every continent on Earth) even have a strategy for dealing with extended periods of desiccation.

Review Questions

Summary Questions

1. Which environments on Earth represent “extreme” conditions for life to survive in?
2. What types of organisms can thrive in these extreme environments?
3. For each of the following extremophiles, explain what extreme condition they can survive in and where they are found on Earth: thermophiles, psychrophiles, acidophiles, alkaliphiles, halophiles, radioresistant, endoliths, xerophiles and barophiles.
4. What are some other places in our solar system with extreme conditions that could support extremophiles?
5. What are tardigrades? Which extreme conditions can they survive in?

Exercises

4: THE SEARCH FOR LIFE IN OUR SOLAR SYSTEM

- 4.1: Solar System Formation
- 4.2: Earth Characteristics
- 4.3: Searching for Life In Our Solar System
- 4.4: The Search for Life on Mars
- 4.5: The Jovian Moons

This page titled [4: The Search for Life in Our Solar System](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

4.1: Solar System Formation

Much of astrobiology is motivated by a desire to understand the origin of things: to find at least partial answers to age-old questions of where the universe, the Sun, planets, the first life on Earth, and we ourselves came from. On Earth, chemicals on the early surface at some point made the transition from non-living to living — the process of abiogenesis. We can take a step back and ask how the solar system formed and why the terrestrial and jovian planets ended up with their initial compositions. Then we can see if the model for how our solar system formed can describe exoplanetary systems as well.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain how stars are formed in giant molecular clouds.
- List the main properties of the planets in our solar system.
- Describe the main steps in forming the solar nebula.
- Discuss how the solar nebula theory explains all of the properties observed in our solar system
- Discuss evidence that supports the nebular theory for forming solar systems.

Star Formation

At the heart of our solar system is a star. Stars of any type can host a planetary system and the most common type of stars in our galaxy are small, cool red dwarfs. The ultimate fate of a star, from birth to death, are determined by its initial mass. Where does this mass of material that forms a star come from? The answer is from huge clouds of gas and dust called Giant Molecular Clouds (GMCs).

Spiral galaxies like the Milky Way contain about 1000 GMCs, most located in the spiral arms of the galaxy. GMCs are among the largest objects in galaxies, with physical dimensions spanning 10 light years up to 1,000 light years. GMCs are cold enough for molecules to remain stable and are made of hydrogen (mostly H_2), helium and other molecules as well as dust particles made of carbon, iron and silicates. All of the main building blocks for life — carbon, hydrogen, oxygen and nitrogen (CHONs) — were present in the cloud that collapsed to form our solar system. Ultimately, this concentration of gas and dust collapses to form hundreds of thousands of new stars and planets.



Figure 1 – NGC 3324, a star forming region with giant molecular clouds in the Carina Nebula. This picture was taken in infrared light by the James Webb Space Telescope. The field of view shown in this image is approximately 16 light-years across. **Only a small fraction of the GMC is used for star formation. How do you think the GMC dissipates? Hint: Think about the evolution of the highest mass stars in the cloud.**

Side Note. Roughly half of the stars in our galaxy emerge from a GMC gravitationally bound to another star. Such stars are called “binary” stars and they can have orbital periods as short as an hour or as long as thousands of years. The O and B type stars are exceedingly rare. However, roughly 75% of these massive stars are gravitationally bound in binary star systems. In contrast, about 60% of stars like our Sun have a gravitationally-bound stellar buddy and fewer than 40% of the smallest stars – the low mass M dwarfs – have binary stellar companions.

A GMC is stable when the force of gravity, which tries to collapse the cloud, is balanced by the pressure from random motions of molecules in the cloud. Gravity is a force that pulls objects together and only depends on the mass of the objects and how far away from each other they are; the closer the objects, the stronger the gravity. A shock wave from a nearby star exploding as a supernova or some other event that compresses the density of the cloud can cause the cloud to collapse and fragment into smaller clumps. These smaller clumps are the seeds of solar systems, and it is the mass of these individual clumps that determines the fate of the star that forms at its center.

As a clump contracts, the density and the collision rate between particles increases so that the core of the cloud begins to heat up. At this point, the core of the cloud is a protostar with a temperature of a few thousand degrees, embedded in an obscuring shroud of cooler gas. As the clump contracts, it spins up and flattens and the object becomes a protoplanetary disk, or protoplanetary disk.

Want to know more – star formation in spiral arms of the Milky Way

It is interesting to map out the location of star formation. Stars are distributed nearly uniformly in the disks of spiral galaxies, but you would never know it looking at the composite image of the Whirlpool galaxy. The bright spiral arms are regions of enhanced density where new star formation is triggered. Because the massive stars are the most luminous, they light up the arms during their relatively short lives and are never seen far from the molecular clouds. The youngest stars are born in clusters located along the spiral arms of the galaxy and the giant molecular clouds are seen as dark dust lanes along the trailing edge of the spiral arms. The spiral density waves that cause turbulence and large scale coherent structure are observed only in flattened galaxies but the forcing mechanism that causes the spiral density waves is complicated and still controversial.



Figure 2. The Whirlpool Galaxy presents a view of dust lanes and massive stars tracing out spiral arms. Would you guess that there are more stars located along the spiral arms? If so, you would be wrong... so what is going on? The spiral structure is a region of higher gas density, where bright, young stars are forming.

The Whirlpool Galaxy presents a view of dust lanes and massive stars tracing out spiral arms. Would you guess that there are more stars along the spiral arms? If so, you would be wrong... so what is going on? The spiral structure is a region of higher gas density, where bright, young stars are forming.

Solar System Observations

Any theory of solar system formation must be able to explain all of the properties of existing solar systems. This includes not only our solar system but the properties of exoplanetary systems in our galaxy. We will stay focused on first explaining the properties of our own solar system, via the solar nebula model. The video below was made when astronomers had only detected about 500 exoplanets (now there are more than 5000) and summarizes the properties of objects in our solar system.

Video credit: @NatGeo



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=109>

After watching the video clip above or doing some research, make a list of similarities in the properties of the planets in our solar system. For example, do all planets orbit around the Sun in the same direction?

Now list any unusual patterns that you observe. Are comets and asteroids found at random distances or in specific locations?

Show Answer

1. There are two types of planets: small, rocky planets (terrestrial) and large gas giants (jovians)
2. Terrestrial planets are located in the inner part of the solar system; jovian planets in the outer part of the solar system
3. Terrestrials are denser than jovians
4. All of the planets revolve around the Sun in the same direction and the rotation ("spin") of the Sun has the same direction as the orbital revolution of planets.
5. The orbital planes of the planets are inclined by only a few degrees with respect to each other. In other words, planets all revolve around the Sun in approximately the plane of the Sun's own rotation.
6. All planets rotate on their axes with different tilts
7. The solar system contains large numbers of asteroids and comets that are at specific distances
8. All jovian planets have moons and rings; no terrestrial planets have rings and only Earth and Mars have moons.

We can organize the properties into three main categories: motion constraints, size and chemical constraints, and distance constraints. We call them *constraints* because they place restrictions on our theories; unless a theory can explain the observed facts, it will not survive in the competitive marketplace of ideas that characterizes the endeavor of science. Let's take a look at these constraints one by one.

There are many regularities to the motions in the solar system. All eight planets revolve around the Sun in the same direction and approximately in the plane of the Sun's own rotation (the blue orbits in Figure 3). If looking down at the solar system from a vantage point above it, this direction is counterclockwise. With the exception of the comets and other objects orbiting beyond Neptune, such as dwarf planets, the motions of the system members define a flattened disk or pancake shape.

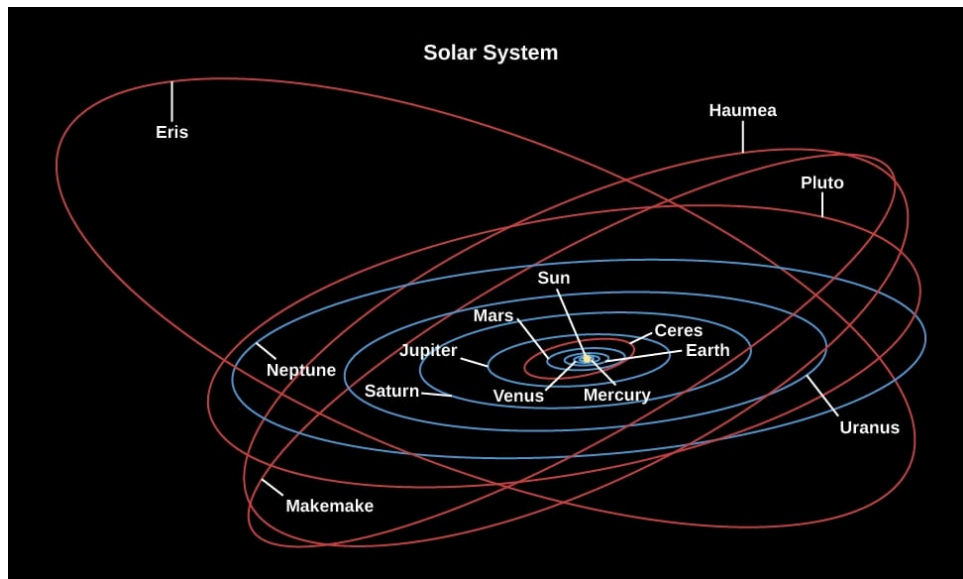


Figure 3 – Orbits of the Planets. All eight major planets orbit the Sun in the same direction and in roughly the same plane (blue lines). The five currently known dwarf planets are also shown: Eris, Haumea, Pluto, Ceres, and Makemake (red lines). ***In what way is Pluto's orbit more similar to the dwarf planets than the regular solar system planets?***

Side Note: Pluto was reclassified from a planet to a dwarf planet in 2006. The reason for Pluto's "demotion" is that it has not cleared the neighborhood around its orbit, meaning it is orbiting the Sun with a lot of other material in the Kuiper Belt.

Most of the planets rotate in the same direction as they revolve, with the exception of Venus which rotates clockwise. Some planets have no tilt to their rotation axes while other are tilted at seemingly random angles: the Earth's axis is tilted by 23.5°, for example. Figure 4 summarizes the different tilts.

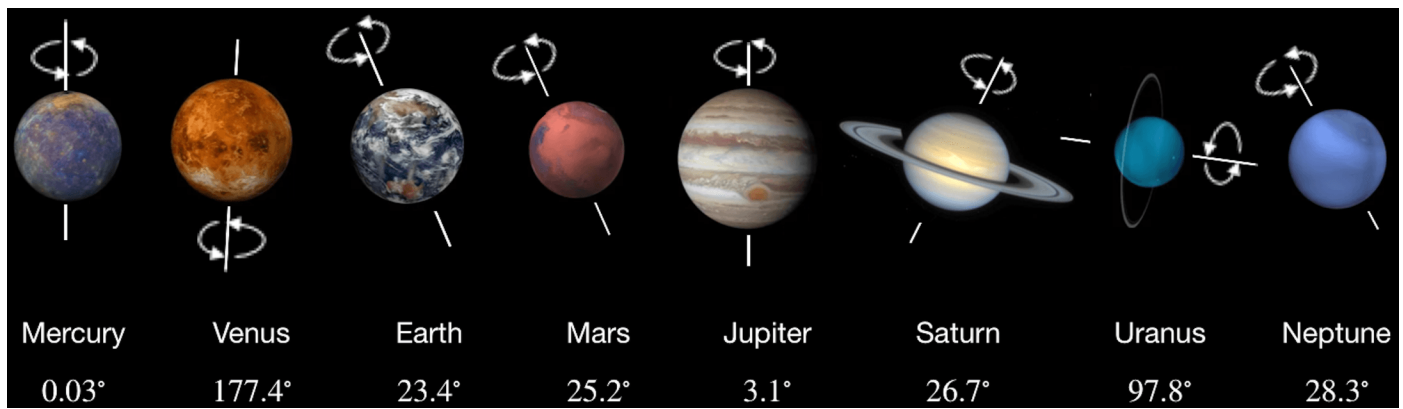


Figure 4 – Planetary tilts, orbital and rotation periods. **Which two planets in our solar system have the most extreme orbital tilts?**

The original version of this chapter contained H5P content. You may want to remove or replace this element.

There are two main types of planets: rocky terrestrials (Mercury, Venus, Earth and Mars) and gas giants (Jupiter, Saturn, Uranus and Neptune). The terrestrials are smaller in both mass and radius than the jovians. By chemical constraints, we mean that the different types of objects in the solar system — terrestrial planets, jovian planets, asteroids, and comets — all have different chemical compositions. The four terrestrial planets are rocky worlds with a solid crust made of rocks containing mostly silicate minerals (e.g., granite and slate). Although the jovian planets have rocky cores, they are vastly dominated by the very light gases hydrogen (H) and helium (He). These chemical differences are reflected in the relatively high densities of the terrestrial planets compared with the jovians. There are sub-trends within the trends as well that need to be explained. For example, the Moon and Earth have some similarities in their chemical compositions but some differences, too.

The terrestrial and jovian planets are at different locations in the solar system: the terrestrial planets are in the inner solar system while the jovians are in the outer solar system. In general, asteroids are found in a belt in between Mars and Jupiter, and comets are found out past Neptune in the Kuiper Belt.

In the next section, we describe the solar nebular theory for how our solar system formed, and explain how each of the constraints described above are successfully explained by this theory.

The Solar Nebula Model

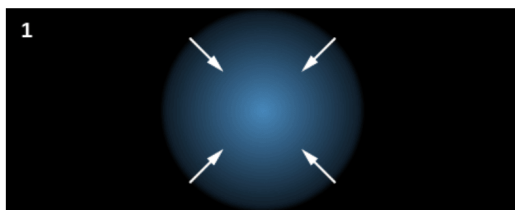
The cloud of gas and dust that collapsed to become our solar system is called the **solar nebula**. Our solar system was formed from this cloud beginning 4.6 billion years ago. The figure below shows an artist's sketch of material in the solar nebula orbiting the protosun: many collisions occurred during this time and eventually material clumped together into larger objects: gas and dust grains stick together via first electrostatic forces and then gravity to form clumps, these clumps accrete more material through their increased gravitational pull and become **planetesimals**, and these planetesimals further grew to become **protoplanets** and eventually planets.



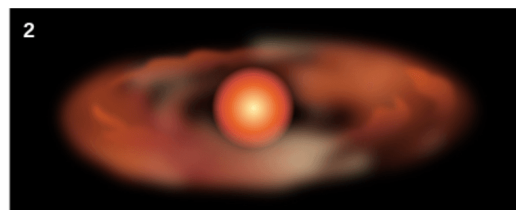
Figure 5 – Solar Nebula. This artist's conception of the solar nebula shows the flattened cloud of gas and dust from which our planetary system formed. Icy and rocky planetesimals (precursors of the planets) can be seen in the foreground. The bright center is where the Sun is forming.

Think about conservation of angular momentum and collision between particles to explain why most particles in the protoplanetary disk ultimately rotate in the same direction

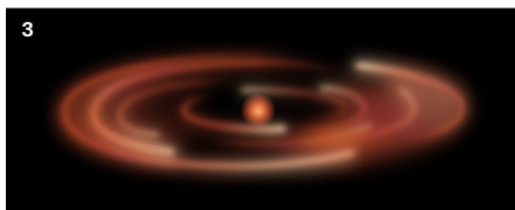
As the solar nebula collapsed under its own gravity, material fell toward the center, where things became more and more concentrated and hot. Increasing temperatures in the shrinking nebula vaporized most of the solid material that was originally present. At the same time, the collapsing nebula began to rotate faster through the conservation of angular momentum, which tells us that a rotating object will spin faster as it gets smaller (or spin more slowly as it gets larger). Like a figure skater pulling their arms in to spin faster, the shrinking cloud spun more quickly as time went on. Now, think about how a round object spins. Close to the poles, the spin rate is slow, and it gets faster as you get closer to the equator. In the same way, near the poles of the nebula, where orbits were slow, the nebular material fell directly into the center. Faster moving material, on the other hand, collapsed into a flat disk revolving around the central object (Figure 6).



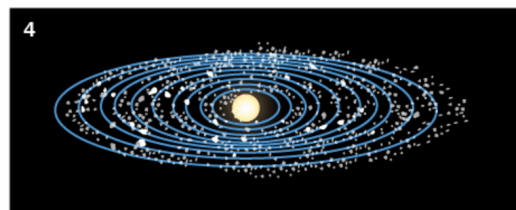
The solar nebula contracts.



As the nebula shrinks, its motion causes it to flatten.



The nebula is a disk of matter with a concentration near the center.



Formation of the protosun. Solid particles condense as the nebula cools, giving rise to the planetesimals, which are the building blocks of the planets.

Figure 6 – Steps in Forming the Solar System. This illustration shows the steps in the formation of the solar system from the solar nebula. As the nebula shrinks, its rotation causes it to flatten into a disk. Much of the material is concentrated in the hot center, which will ultimately become a star. Away from the center, solid particles can condense as the nebula cools, giving rise to planetesimals, the building blocks of the planets and moons. **Explain how this process links the timing of planet formation to star formation.**

It eventually gets hot enough at the center of the disk (10 million K) for the protosun to begin fusing hydrogen into helium; this marks the beginning of the Sun's life as a star. In the last stage (4) shown in Figure 6, material is clumping together within the

lanes that formed in the solar nebula, and planetesimals accrete more material to become protoplanets and then the planets.

How do we know?

Protoplanetary disks had long been theorized to exist in star forming regions and evidence of their existence began with the detection of a circumstellar disk around the star β Pictoris in 1984 with the Infrared Astronomy Satellite. The images of the protoplanetary disks in the Orion Nebula in the figure below were taken with the Hubble Space Telescope (HST), which launched in 1990. The bright points at the center of the proplyds in Figure 7 are young protostars that are surrounded by dusty disks. Many other proplyds were observed in the Orion Nebula by the same team using HST, and this was a very significant finding in 1992. This is the same year that the first exoplanet was detected orbiting a pulsar and there was still much speculation on how common other solar systems were in our galaxy. Finding a large number of proplyds in just one nearby star forming region strongly suggested that the process of planet formation could be common, and perhaps other planets could host life.

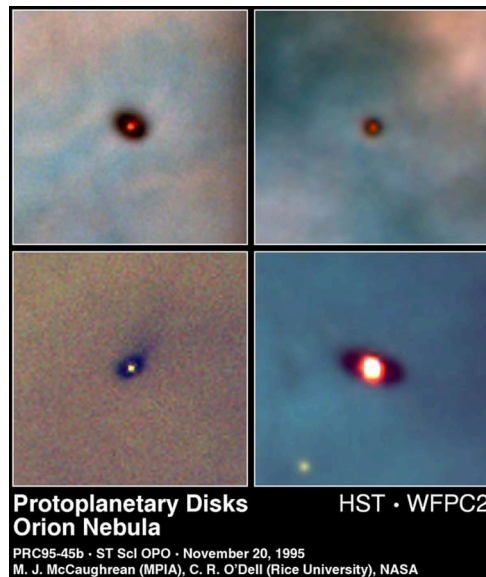


Figure 7 – Hubble Space Telescope images of four protoplanetary disks around young stars in the Orion nebula, located 1,500 light-years away. Gas and dust disks can be directly seen in visible light by Hubble. Match these images to the theoretical stages for planet formation (Figure 6).

The temperature within the disk decreases with increasing distance from the Sun, much as the planets' temperatures vary with position today. As the disk cooled, the gases interacted chemically to produce compounds; eventually these compounds condensed into liquid droplets or solid grains. This is similar to the process by which raindrops on Earth condense from moist air as it rises over a mountain.

Closer to the disk, it was too hot for hydrogen compounds to solidify into ices; at these higher temperatures, only rocks and metals could condense into solids. Eventually, at a certain distance away from the star, it is cool enough for the hydrogen compounds to also condense into ice. This idea is shown in Figure 8 below, where the line beyond which it is cold enough for hydrogen compounds to freeze is indicated as the **frost line** or **snow line**. The location of the frost line depends on the mass of the star that is forming at the center. For a star like the Sun, the frost line is 2.7 AU from the center. It will be closer for a low mass star and further out for a more massive star.

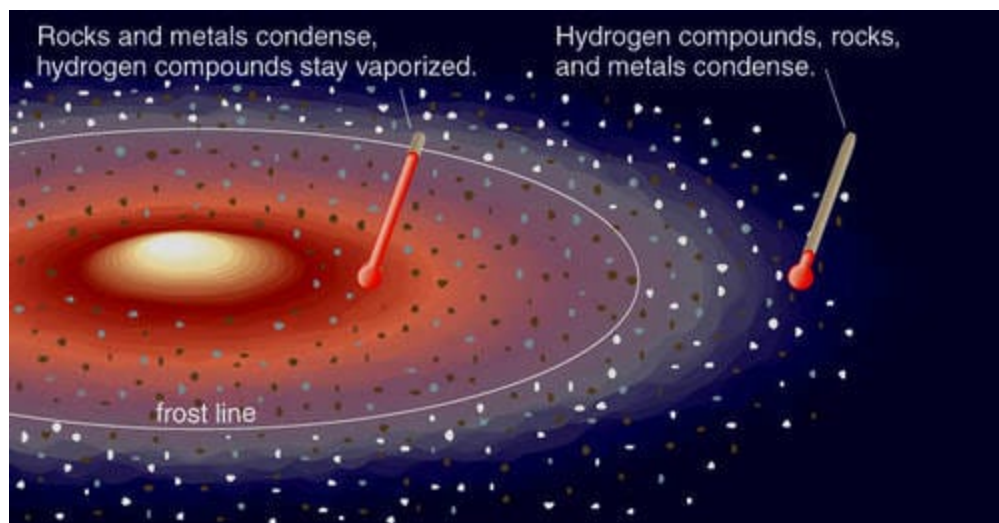


Figure 8 – This cartoon sketch of the protoplanetary disk illustrates that rocks and metals condense in the inner part of the disk where the temperatures are high. **What is the approximate distance to the frost line in our solar system?**

This very intuitive idea — that temperature decreases as you get further from the star — goes a long way in explaining the chemical differences as well as size and density differences between the terrestrial and jovian planets. The jovian planets are much more massive than the terrestrial planets, while the terrestrial planets have a higher density than the jovians. The terrestrial planets are rocky and have average densities of 5.0 g/cm^3 , while the jovian planets have an average density of 1.2 g/cm^3 (for reference, at STP, hydrogen gas has a density of 0.089 g/cm^3 , water has a density of 1.0 g/cm^3 , silicon has a density of 2.3 g/cm^3 , and iron has a density of 7.9 g/cm^3).

Let's start with the mass differences. The terrestrial planets are made of rock and metal, and the jovians are made of rock, metal, and ices. Ice has mass, so the jovians grew more massive as they accumulated ice on top of the rock. The higher mass from the ice meant that gravity was stronger for these planets and hence more material could be pulled in. More material means more mass and stronger gravity...this self-reinforcing process continued until the jovians had pulled in all of the hydrogen and helium gas. This explains the mass differences.

As for density, since the additional ice and gases that the jovians pulled in have very low densities, this brought the overall densities of the jovians down to values that are much lower than rock. The density of Jupiter is 1.3 g/cm^3 and this is the average density for the entire planet. While Jupiter has a dense metal core, the planet is vastly dominated by the light gases helium and hydrogen so the average density is much lower than the values for metals and rock. Without these additional ice and gas layers, the terrestrial densities remain higher, with values matching those of rocks and metals.

The solar nebula model explains why planets should be expected to orbit in the same direction as the spin of the star: the planets and star all formed from the same spinning disk of material. Therefore, the Sun rotates in the same direction as the spinning disk and the planets all orbit the Sun in the same direction. The planetary orbits would also stay in the plane of the flattened disk, which explains why the orbits of the planets are coplanar (recall Figure 3).

The different locations of the terrestrials and jovians are also explained by the temperature in the disk where they formed. Asteroids and comets are left over planetesimals that were not swept up into a planet. Their compositions are also related to the conditions where they formed in the disk. Asteroids formed inside the frost line and are made of rocky materials. Comets formed outside the frost line, and are consequently rocks covered in a layer of ice.

The unusual features that we observe, such as the sideways tilt of Uranus or Venus' upside-down rotation, can be explained by the solar nebula model as the result of collisions during the planet-building phase. What about all the moons and rings around the jovians? This relates back to the high masses of the jovian planets. Since these planets are so massive, their strong gravity pulls in some of the dust and gas from the solar nebula, thus becoming "mini solar nebulas" as they form. This swirling disk of material around the jovians is the material that forms their moons and rings. The terrestrial planets were not massive enough to pull in material from the solar nebula, so the terrestrial moons are thought to be the result of early collisions with planetesimals. Our Moon, for example, is believed to have formed from the debris of a collision between the early Earth and another large body (Figure 9). Evidence to support this formation scenario for the Moon comes from lunar rocks brought back to Earth by astronauts on the Apollo missions between 1969 and 1972. These rocks confirmed that the composition of the Moon is similar to that of the

Earth's crust and mantle.

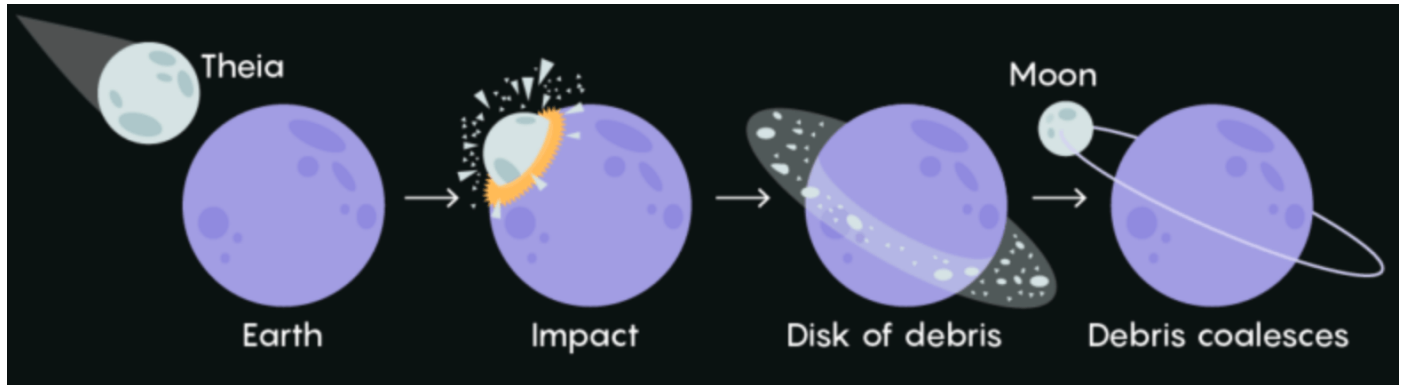


Figure 9 – The giant impact hypothesis theory for the formation of the Moon. **What does this model for formation of the moon imply about the timing of the collision between Earth and Theia?**

How do we know?

Dramatic evidence supporting the separation of material into distinct lanes or rings within the solar nebula came from this image shown to the right (Figure 10), which was taken in 2014 with the Atacama Large Millimeter Array (ALMA) of radio telescopes in the Chilean desert. This picture of the young star HL Tauri looks like an artist's sketch of a solar system forming; in fact, it looks quite similar to how panel 3 of Figure 6 would look if sketched face-on rather than edge-on.

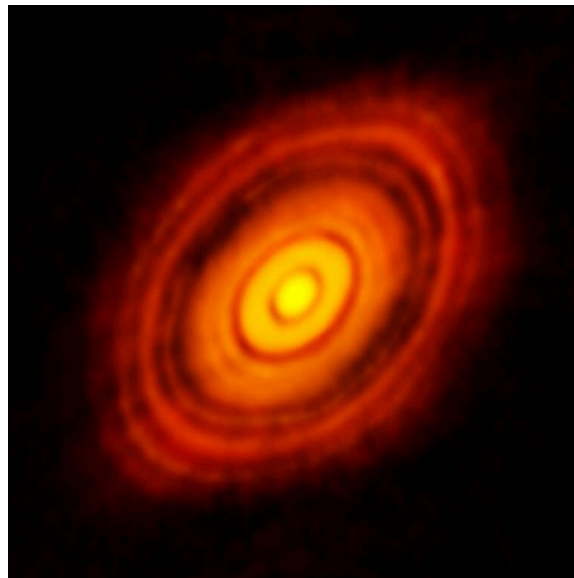


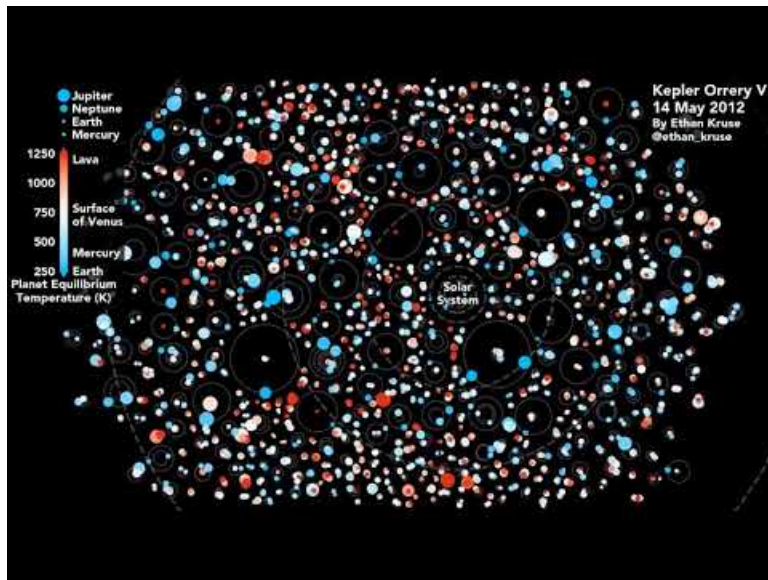
Figure 10 – The protoplanetary disc surrounding the young star HL Tauri, taken by ALMA. **Where in the HL Tau disk might planets be forming? What is your reasoning?**

Comparison with other solar systems

Over 4,000 exoplanetary systems have been detected to date. Do they resemble our solar system, with four inner rocky planets and four outer gas giants. In a word, No. The animation below shows what the multi-planet systems detected by the Kepler Mission look like. Only the planets are shown and not the host stars and the planets are color-coded according to their temperature: the hottest “lava” planets are red and comparatively cooler planets, like Earth, Mars, and the gas giants, are colored blue. Planets that are analogues of Venus, with its hot rocky surface that could melt lead, are colored white. Watch the animation a few times and pause it to really take in how different all of these planetary systems are from our own.

Animation of multi-planet systems.

Video credits: Ethan Kruse/NASA Goddard, CC BY 3.0



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=109>

There are some biases for the planetary systems shown in this animation, as the transit method that Kepler uses to detect planets will preferentially find planets that are orbiting close to their host star. We can still make the general conclusion that solar systems show an immense amount of variation. As we explore in *Exoplanets: Statistics and Discoveries*, the vast majority of exoplanets detected to date are different from the terrestrial and jovian planets in our solar system. In some cases, gas giants like Jupiter are found orbiting extremely close to their host stars. These “hot Jupiters” seem inconsistent with a formation scenario in which they formed so close to their stars. Computer simulations are being used to explore if perhaps these planets formed where expected but somehow migrated closer in to their star.

From the evidence we see of protoplanets in our galaxy, we have confidence that the main idea of the solar nebula model — that solar systems form when a cloud of gas and dust collapses and forms a swirling disk of material with a protostar at its center — is correct.

Solar System Stability

Planets are kept in their orbits primarily through gravitational interactions between the planet and their host star. However, planets also feel a much smaller but non-negligible gravitational pull from the other planets in the system. Though the planets in our Solar System are relatively well separated, this is not always the case in other Solar Systems, and may not even always be the case for our own system. The closer two bodies are to each other, the stronger their gravitational pull will be on each other.

Close encounters between planets can affect their orbits’ eccentricities, semi-major axes, and relative inclination. These effects are accentuated in the case of resonant systems. **Resonance** describes the situation in which two planets have periods that form an integer ratio. For example, if one planet has a period of two years and a second planet has a period of four years, these two planets are said to be in resonance. Today, we can use computer simulations to determine what will happen to different systems of planets as they evolve through time that will expose these types of interactions.

Planets may also move around during formation. This may push smaller, more-likely rocky planets into the host star, or send debris like asteroids and comets into the path of other planets. Such motion in our Solar System is thought to have caused the period of late heavy bombardment, which left its mark on our Moon. This type of mechanism may have caused the strange position of hot Jupiters, which are commonly found and much closer in than where they could have formed.

Figure 11 – A proposed model of planet migration in the early Solar System. The different a, b, c, and d lines represent different snapshots in time. The x-axis is distance from the Sun in AU. Smaller circles represent planetesimals before they have formed into full planets. The movement of Jupiter and Saturn in this model appears to have affected the formation of the close-in terrestrial planets, which are favorites for habitability. **How does the current position of the asteroid belt figure into this model? What about the Kuiper belt?**

In 2008, a paper by Batygin and Laughlin showed that on million-year time scales, planetary orbits may evolve in random, but ultimately stable ways. However, on longer timescales, orbits can rapidly and spontaneously evolve into chaotic states hard for even the best computer simulations to predict. Subtle interactions between the different planets and objects with mass in the solar system are capable of enacting significant change over the dynamical state of a system of planets. Close encounters will amplify these changes, rearrange orbits, and eject small bodies out of the Solar System. Resonant affects could even depopulate entire areas in a planetary system. The architecture and stability of a planetary system are important contributors to a planet's potential habitability.

Key Concepts and Summary

Stars form when high density regions in giant molecular clouds begin to gravitationally collapse. As the protostar is collapsing, conservation of angular momentum forces it to spin faster and a disk of gas and dust settles around the equator of the star. This is the protoplanetary disk, where clumps of material collide and grow to form planetesimals, which collide and grow to form planets. In our solar system, there are two types of planets that formed: smaller rocky planets with thin atmospheres and gas giant planets. The solar nebula model describes formation of the solar system and describes the main features that we observe: the rocky planets orbit more closely to the Sun and gas giants formed and orbit beyond the ice line. The process of planet formation takes almost a million years, far beyond the lifetime of human astronomers. However, the discovery of young stars with protoplanetary disks and the structure of exoplanetary systems have helped to confirm and extend our understanding of star and planet formation.

Review Questions

Summary Questions

1. What are giant molecular clouds? How do star systems form within GMCs?
2. List the main observed properties of the planets in our solar system.
3. What is the solar nebular theory?
4. What properties are different for terrestrial and jovian planets?
5. How does the solar nebular theory explain each of the properties observed in our solar system?
6. How do comets and asteroids differ from planets? Where are comets and asteroids found in our solar system?
7. What evidence do we have that the solar nebula model explains the formation of other solar systems in our galaxy?
8. How did Earth's moon form? What evidence supports this claim?
9. Do other solar systems have the same layout as our own? What evidence supports your answer?

Exercises

4.2: Earth Characteristics

The Earth is filled with a rich diversity of life. What makes the Earth such a habitable oasis for life? The planets Venus, Earth, and Mars all started out with atmospheres, and yet they evolved into vastly different worlds today. Certain features unique to the Earth, such as plate tectonics, have enabled it to be a dynamic world with oceans and a climate hospitable for life. By identifying the characteristics that make Earth hospitable, we can gain insight into what properties to look for on exoplanets that may point to the possibility of life.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss the Earth's density and surface gravity and how it compares to the other planets in our solar system
- Explain the differentiation of Earth and its main interior layers
- Discuss how heat is generated inside a planet and how it is lost to cooling.
- Discuss the importance of plate tectonics for Earth's habitability
- Explain how Earth's magnetosphere protects life on Earth



Figure 1 – The Blue Marble. This image of Earth from space, taken by the Apollo 17 astronauts, is known as the “Blue Marble.” This is one of the rare images of a full Earth taken during the Apollo program; most images show only part of Earth's disk in sunlight. **Can you estimate the fraction of Earth covered by water from this image? Can you describe the relative thickness of the atmosphere?**

The size of a world and distance it orbits its host star can take us very far in assessing whether the world could host life. We say more in the chapter on Habitable Zones about how the distance of the Earth from the Sun keeps the Earth at the right temperature for liquid water — and makes Venus too hot today for liquid water — and we will focus here on the bulk properties of the Earth.

Bulk Properties of the Earth

The most basic measurements of the Earth are its size, mass, and density. Many features of the Earth have been understood since antiquity. Aristotle argued that the Earth had to be round because the Earth cast a circular shadow over the moon during a lunar eclipse. [Eratosthenes of Cyrene](#) obtained a clever measurement of the radius of the Earth that was within 20% of the value that we can measure today. To a good approximation, the Earth is spherical in shape and the volume of the spherical Earth is given by $\frac{4}{3} \pi R^3$, where R is the radius of the Earth.

Side Note: The experiments of the ancient Greeks show how much can be determined with basic high-school mathematics.

Before you decide that something is unknowable, pause and think about how you could make a reasonable estimate, even within a factor of 10 (an “order of magnitude”). This habit of making estimates will make you a powerful problem-solver in science, business, and life.

Armed with knowledge about the volume of Earth, we can use the density of rocks to estimate the mass of Earth. Density is simply the ratio of mass to volume, $\rho = M/V$, and it tells us how tightly material is packed into an object of a given size. A rock that is about a cubic centimeter in size weighs about 3 grams, so the density of that rock is $\rho = 3 \text{ g/cm}^3$. We can then turn the density equation around to estimate the mass of the Earth: $M = \rho V$. This estimate is off by a factor of about two since the interior of the Earth is compressed and has a higher density than surface rocks. However, this “order of magnitude” estimate is an excellent approximation – a factor of two is a great start when nothing is known before.

In the early 18th century (and with the benefit of insights from Isaac Newton about the laws of gravity) Henry Cavendish “weighed the Earth” in his lab by cleverly measuring the acceleration of gravity on Earth. Gravity is a property of mass and Cavendish calculated a bulk density for the Earth of 5.48 g/cm^3 , very close to the currently known value of 5.515 g/cm^3 .

The density of a planet can easily be calculated if the radius and mass are known, quantities that can be measured even for planets orbiting other stars. The density of a planet gives immense insight into that world. All terrestrial planets in our solar system have higher densities than the jovian gas giants. The average density of the terrestrial planets is 5.0 g/cm^3 , while the average density for the jovian gas giants is 1.2 g/cm^3 . Exoplanets with densities that fall in between these values are commonly found and point toward intermediate types of planets that are not found in our solar system. For example, the “mini-Neptune” K2-18 b has an estimated density of 2.6 g/cm^3 .

Worked Example: Calculating the density of a planet

Looking up the mass and radius of the Earth, and using the density equation we find:

$$\rho = M/V = M / \left(\frac{4}{3} \pi R^3 \right) = (5.97 \times 10^{24} \text{ kg}) / \left(\frac{4}{3} \pi (6.38 \times 10^6 \text{ m})^3 \right) = 5500 \text{ kg/m}^3$$

This tells us that an average cubic meter of the Earth contains 5500 kg, or 5.5 metric tons of matter. A cubic meter is about the size of a washing machine. If 5.5 tons seems heavier than you would have guessed, it is because the core density of Earth is greater than the dirt and rocks on the surface of our planet.

It can be more convenient to express densities in units of g/cm^3 (grams per cubic centimeter). In these units the density of water is 1 g/cm^3 and makes an easy reference point to remember. In these units, $\rho_{\text{Earth}} = 5.5 \text{ g/cm}^3$

Extra: Saturn has the lowest density in our solar system, with a value of $\rho_{\text{Saturn}} = 0.7 \text{ g/cm}^3$. How many times denser is the Earth than Saturn?

Show Answer

7.9 — the Earth is almost 8 times more dense than Saturn

Another property that also depends on the mass and radius of a planet is **surface gravity**. The equation for finding the surface gravity (g) of a world is: $g = G M/R^2$. Here, G is the gravitational constant and we can easily make comparisons of the surface gravities of different worlds by taking their ratio relative to Earth. Table 1 lists the densities for the planets in our solar system and their surface gravities relative to the surface gravity of Earth. The surface gravity helps to assess if rocky planets can hold onto their atmosphere.

Table 1 – Comparison of the density (ρ) and surface gravity (g) for planets in our solar system.

	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune
$\rho \text{ (g/cm}^3\text{)}$	5.4	5.2	5.5	3.9	1.3	0.7	1.3	1.6
$g \text{ (relative to Earth)}$	0.38	0.91	1	0.38	2.4	0.92	0.89	1.1

The size of a world is also directly related to how fast it cools down. All planets start off in a molten state and over time the energy is radiated away into space. The **cooling time** of an object depends on the ratio of the surface area to the volume ($\frac{SA}{V}$) of a planet. Since the surface area of a sphere is $SA = 4\pi R^2$ and the volume of a sphere is $V = \frac{4}{3}\pi R^3$, this means that the cooling time is proportional to $\frac{1}{R}$. This means that there is an *inverse* relationship between the cooling time for a world and its radius, and a smaller world cools off faster than a world with a larger radius. How quickly a world loses its heat (putting aside for a moment the question of what creates the heat inside of a world) can be understood with a simple example from everyday life. Consider a large, hot baked potato and a small piece cut off the larger potato: which cools first? Intuition tells us it is the smaller piece, and this intuition is correct. In our solar system, the worlds with the smallest size indeed have cooled the fastest: the asteroid Ceres, the Moon, and Mercury have all entirely lost their internal heat and are *geologically dead*.

The fact that smaller worlds lose their internal heat (or cool off) more quickly than larger worlds is a fundamental reason why the Earth is hospitable to life. The Earth still has heat inside that can power geological activities such as plate tectonics that stabilize our atmosphere.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Earth's Interior

How do we know about the interior structure of Earth, since we live on the surface? We can only drill a few kilometers before the temperature and pressure become too high to continue. The deepest manmade hole on Earth is the aptly named the Kola Superdeep Borehole, which reaches down to 12.3 km (7.6 miles).

Almost everything else we know about the interior of the Earth comes from seismic waves. Seismic waves carry the energy from events like earthquakes or volcanic eruptions. Seismic waves compress the elastic interior of the Earth like sound waves compress the air. The speed of seismic waves depends on the depth and density of medium. The fastest moving P-waves compress material in the direction that they travel. They are the first to arrive at seismic stations and travel through the core of the Earth. Sheering S-waves are slower and do not propagate through liquid material; they are blocked by the liquid core of the Earth. Slower still, R-waves ripple along the surface of the Earth. They arrive last at the seismic stations but have the potential to do the most damage.

Seismic waves reveal the bulk density structure of the Earth, as depicted in Figure 2 below. The core is about half the radius of the Earth and is comprised of an inner core that is about the size of the moon, surrounded by a molten iron core. The liquid core is also an important factor in the Earth's habitability, as it enables the Earth to have a protective magnetosphere surrounding it. The thick rocky mantle wraps around the core and extends for the other half of the Earth radius. The outer shell is a thin crust of low density rock called the lithosphere.

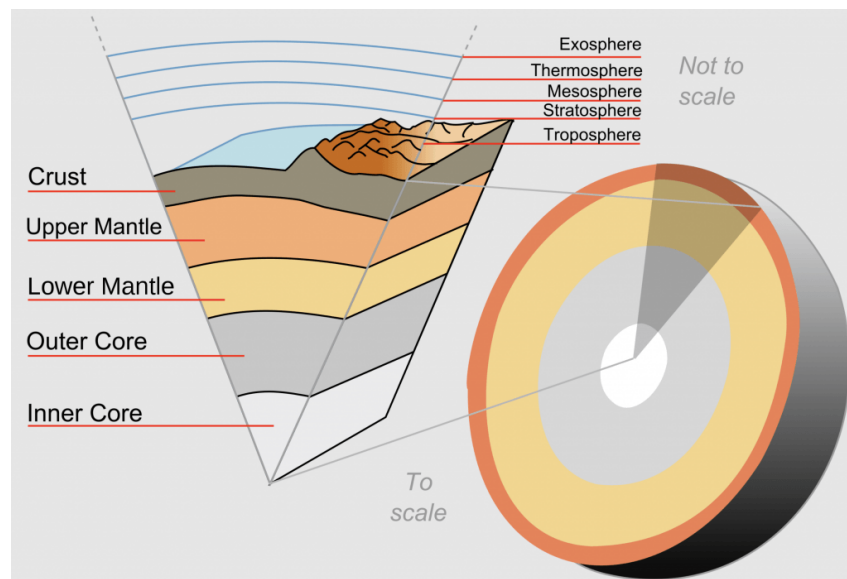


Figure 2 – Diagram showing interior layers of the Earth. **What layer of the Earth generates a global magnetic field? By what process does this occur?**

All terrestrial planets in our solar system appear to have a similar layered structure inside, with a dense metal core at the center surrounded by the mantle and crust. This layered structure occurs because the planet cools over time and as it cools, the process of **differentiation** occurs, where the densest materials (like iron) sink to the center and the lightest materials float to the surface. A simple real world example shows this to be an intuitive process. If you shake oil and water together in a container, they will initially be in a mixture. However, over time, the denser water sinks to the bottom of the container and the lighter oil rises to the top.

One feature that distinguishes Earth from the other terrestrial planets is its crust. The Earth's crust comprises two types: continental crust and oceanic crust. These are the tectonics plates that literally float and move around on top of the mantle, specifically on top of the asthenosphere. Earth is the only terrestrial planet that has oceanic crust so this gives us a clue that plate tectonics may be an important consideration when assessing the habitability of an exoplanet.

Sources of Planetary Heat

The Earth is a geologically active planet. Plate tectonics has already been mentioned as one of the unique properties that the Earth has compared to the other terrestrial planets in our solar system and a source of heat is required to drive this geological activity. There are a few possible origins for heat inside of a planet: **accretion energy**, **differentiation** and **radioactive decay**.

Planets form by accreting planetesimals that collide and stick together to form larger bodies. These collisions generate significant amounts of heat or accretion energy as the planet is assembled. However, all of the heat from accretion energy is injected during the early stages of planet formation.

As material sinks during **differentiation**, energy is conserved as gravitational potential energy is converted first to energy of motion (kinetic energy) and then to heat (thermal energy).

Over time, as differentiation is complete and accretion is much less frequent, **radioactivity** becomes and remains the main source of interior heat for a planet (we note here that the large jovian planets still have heat from their formation in their voluminous interiors). Radioactivity is a natural process in which heavy nuclei spontaneously break apart; when the nuclei break apart into smaller ones, energy is emitted. All planets in our solar system contained the same chemical mixture from the solar nebula at early times, and these include some heavy radioactive elements such as uranium (^{238}U) and potassium (^{40}K).

Radioactivity continues to supply internal heat to the Earth to fuel geological activity, yet Mercury and the Moon are geologically dead. The difference here goes back to the cooling rate of a planet. Smaller planets lose their heat faster than larger ones. Therefore, the size of a world is an important indicator of the potential for geological activity. Before describing types of geological activity, let's consider exactly how heat escapes from inside of a planet.

Heat Loss From a Planet

The interior heat of a world can be lost via several mechanisms: **convection**, **conduction**, and **radiation**. When heat from the core moves to the lower mantle boundary, convection cells are set up inside of the mantle. Convection is the familiar process wherein hot material, such as air, rises as cooler material sinks. Convection, in which energy is transported from a warm region, such as the interior of Earth, to a cooler region, such as the upper mantle, is a process we encounter often in astronomy—in stars as well as planets. You can see convection in action when boiling a pot of water on a stove. As the water heats up from below and starts to boil, convection cells are set up in the water, as seen in Figure 3. Once the heat reaches the top of the mantle, it then moves through the crust via conduction. Finally, the heat makes it to the Earth's surface and is radiated away into space (or, to be specific, in the case of planets with atmospheres, like the Earth and Venus, the heat must travel through this layer first).

Mechanisms of Heat Transfer

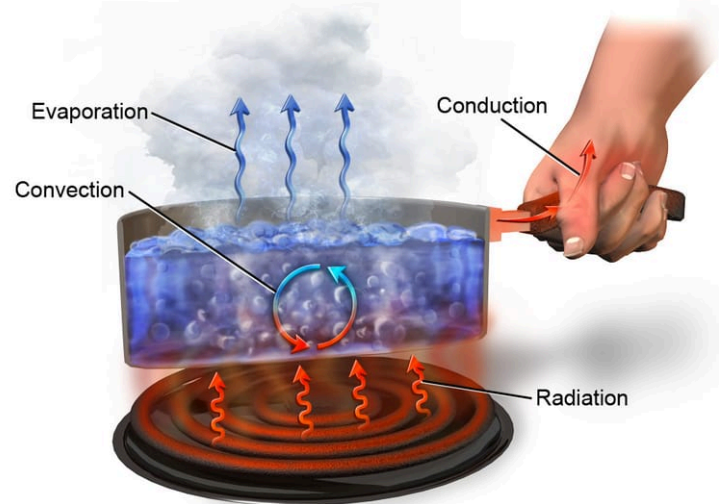


Figure 3 – Heat transfer. For a metal pot of boiling water, heat travels from the source (the stove burner) to the pot, where convection moves the heated water from the bottom to the top. Heat escapes as the water evaporates. **What is the primary mechanism of heat transfer in the molten iron core of Earth? What is the primary heat transfer mechanism on the surface of Earth?**

Now that the mechanisms for generating heat inside a planet, as well as how that heat gets out of a planet, are understood, we can move on to discuss the different types of geological activity that can be driven by this internal heat.

Plate Tectonics

The size of the Earth, specifically its radius, is fundamentally the reason why the Earth is still geologically active: there is enough internal heat inside our relatively large planet to drive geological activity. A smaller terrestrial planet, such as Mercury, is too small to retain enough heat to sustain any geological activity: Mercury cooled off fast and is now geologically inactive.

An important property of Earth is that it has plate tectonics. **Plate tectonics** explains how slow motions within the mantle of Earth move large segments of the crust, resulting in a gradual “drifting” of the continents as well as the formation of mountains and other large-scale geological features. The power to move the plates is provided by slow convection of the mantle, a process by which heat escapes from the interior through the upward flow of warmer material and the slow sinking of cooler material.

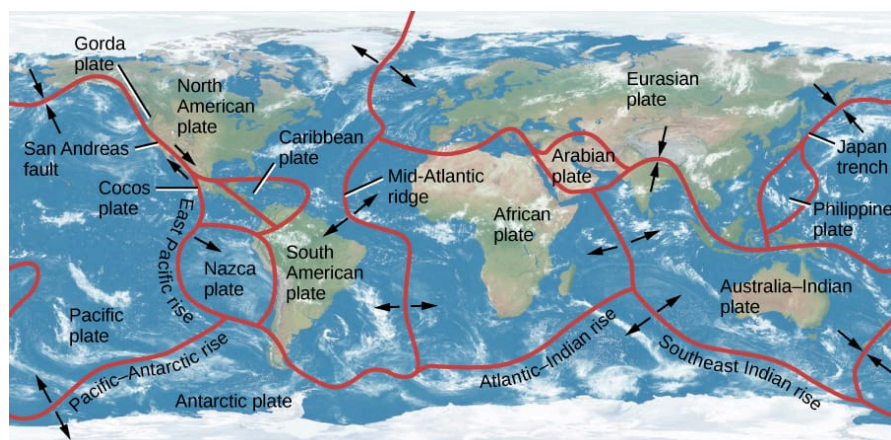


Figure 4 – Earth's tectonic plates. This map shows the major plates into which the crust of Earth is divided. Arrows indicate the motion of the plates at average speeds of 4 to 5 centimeters per year, similar to the rate at which your hair grows.

While Venus and Mars are likely to also have convection in their mantles, Earth is the only planet in our solar system known to have plate tectonics. Geological evidence suggests that plate tectonics began operating on the Earth about 3.8 billion years ago. The upper rocky layer of the Earth is divided into about twelve major tectonic plates that float on top of the convecting mantle, shown in Figure 4. The distribution of these plates affect global climate because continental crust has a higher albedo (reflects

more light) than ocean water. Continental crust was thought to have a maximum extent between 1.6 – 2.7 billion years ago, causing ice ages and high rates for burial of organic material in the early and late Proterozoic eon.

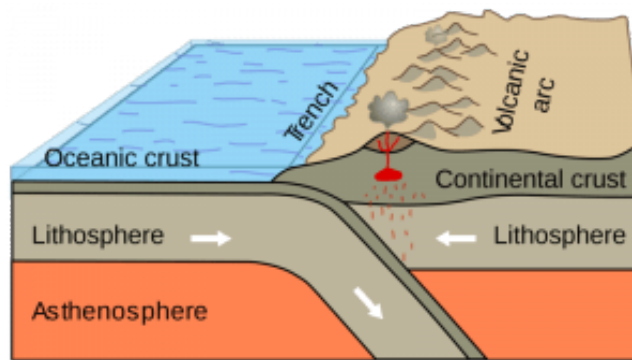


Figure 5 – A subduction zone, where an oceanic plate beneath a continental plate.

Plate tectonics provide a mechanism for circulating material between the surface of the planet and Earth's interior. Plates pull apart from each other along rift zones, such as the Mid-Atlantic ridge, driven by upwelling currents in the mantle. Subduction zones form at boundaries where one edge of an approaching plate moves under another. These recycling zones are critical for life on our planet. They are part of a **negative feedback loop** that stabilizes our climate thanks to chemical interactions between surface rocks and the atmosphere. Rain removes carbon dioxide from the atmosphere and stores it as calcium carbonate in surface rocks through a weathering process. Carbon-enriched rocks are then transported into the deep mantle along the subduction zones, removing excess CO₂ from the atmosphere. The flow of material goes the other way as well; some elements are transported from the mantle to the surface of the planet.



Figure 6 – Volcanic outgassing. **What elements are released from volcanoes and what is the effect upon the atmosphere?**

Volcanoes are also formed through plate tectonics. When volcanoes erupt, gasses trapped inside of the Earth are released. This process of volcanic **outgassing** supplies the raw materials for the Earth's atmosphere (Figure 6). The primary materials that are outgassed include water vapor, CO₂, N₂, and SO₂. Some of these gasses are greenhouse gases and contribute to maintaining the energy balance of the Earth. Plate tectonics on Earth allows carbon to be cycled from the atmosphere, down to carbonate rocks on the ocean floor, and back into the atmosphere through volcanism. This **carbon cycle** is an essential part of maintaining a hospitable atmosphere on Earth. Other planets in our solar system experience tectonic stresses, but plate tectonics – with the active movement of crustal and oceanic plates – is unique to Earth. When exploring exoplanets that may be conducive to life, the presence of plate tectonics would be a favorable condition, at least based on our experience on Earth.

Earth's Magnetosphere

The core of the Earth has a temperature similar to the surface of the Sun. This heat ionizes the molten iron, producing charged particles. As the liquid metal inside Earth circulates, the charged particles set up an electric current. When many charged particles are moving together like that—in the laboratory or on the scale of an entire planet—they produce a magnetic field. Earth's magnetic field extends thousands of miles above the surface of the planet.

Our planet behaves in some ways as if a giant bar magnet were inside it, aligned approximately with the rotational poles of Earth. When charged particles from the surface of the Sun (the “solar wind”) are directed toward us, the Earth's magnetic field deflects them and shields the planet. This region, called the magnetosphere, is defined as the zone within which Earth's magnetic field dominates over the weak interplanetary magnetic field that extends outward from the Sun (Figure 7). The magnetosphere plays an important role for habitability, by preventing stripping of our atmosphere.

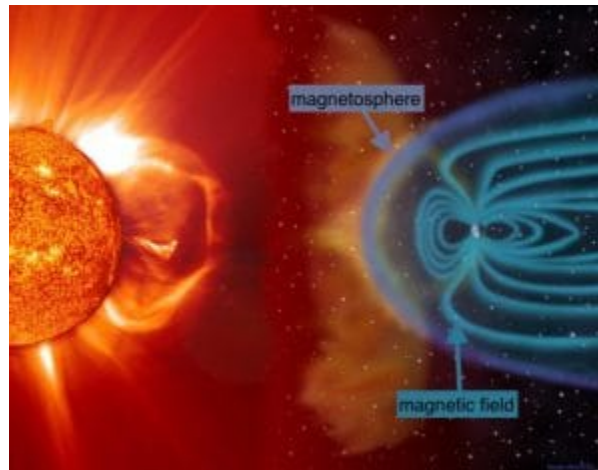


Figure 7 – Earth's magnetic field and magnetosphere. Earth's magnetosphere blocks dangerous particles from the Sun — called the solar wind — from reaching Earth's surface.

For a planet to have a magnetic field, it must have a liquid metallic layer and it must also be spinning fast enough to create a turbulent flow. Jupiter meets both of those conditions — it has a liquid conducting layer made of liquid metallic hydrogen in its outer core and it rotates once every 11 hours — and has the strongest magnetic fields in our solar system. Although Venus has a liquid metallic outer core like the Earth, it rotates too slowly on its axis (once every 243 days) to generate a magnetic field. When assessing the habitability of an exoplanet, we want to determine whether it has the conditions to set up a protective magnetosphere.

There is evidence in Earth's geologic record from magnetic minerals that many magnetic pole reversals have occurred in the past and some scientists believe that the magnetic fields are beginning to collapse and flip again. Life has existed on our planet for billions of years, and no correlation has been found between mass extinctions and magnetic pole reversals. The risk to life during a magnetic pole reversal may be modest, but the risk to power grids that modern humans depend upon will be more substantial.

Earth's Habitability

We have looked at some of the properties of Earth that contribute to making it a habitable world. We began this chapter by stating that the Earth's size and distance together can explain some of the unique properties that make Earth habitable. The Earth has a large enough size to have heat inside of it, initially from accretion and differentiation, but mainly from radioactivity today. As this heat escapes from the Earth, it is carried from the lower to the upper mantle through convection. This heat at the upper mantle powers plate tectonics, which is responsible for maintaining a continuous cycle between the Earth's atmosphere and the crust, thus setting up a stable climate. One of the Earth's differentiated layers is the liquid outer core, and this layer combined with Earth's rotation rate sets up a protective magnetosphere around our planet.

Key Concepts and Summary

The Earth formed through the collision of planetesimals in our solar system when the Sun was forming more than 4.5 Gya. The

planetesimal collisions deposited frictional accretion energy in the interior of our planet and heavier elements in the molten young Earth began to settle toward the core, differentiating into layers. The frictional processes of differentiation delivered an additional boost of energy to the interior. Today, Earth has a solid iron core surrounded by a molten liquid outer metal core. Turbulence in the outer core is coupled with rotation of the planet to generate a global magnetic field that shields the atmosphere from being eroded by the fast-moving particles in the solar wind. The energy of accretion and differentiation were one-time processes that deposited internal heat to the planet long ago. Now, an additional source of energy comes from the radioactive decay of heavy elements. The internal heat beneath the lithosphere drives motion of the floating continental plates. Planets lose internal heat over time (with smaller planets cooling faster).

Review Questions

Summary Questions

1. What two properties of a world determine its density?
2. How does the density of the Earth compare with the other planets in our solar system?
3. How does surface gravity differ from density?
4. Can a planet with a low surface gravity hold on to an atmosphere?
5. Where did heat inside the Earth come from initially? What is the main source of heat from inside the Earth today?
6. How can heat escape from inside a planet to the surface?
7. What property of a planet determines how quickly it will lose its interior heat?
8. Which planet cools faster: Earth or Mars? Explain.
9. How does plate tectonics contribute to Earth's habitability?
10. Can a planet without plate tectonics be habitable?
11. How does Earth's magnetosphere protect life?

Exercises

4.3: Searching for Life In Our Solar System

The Earth is the only place in our solar system that we know has life. Earth has some specific characteristics that make it habitable, such as plate tectonics and a surface gravity that prevents the atmosphere from escaping. What about the other planets in our solar system? By looking at the properties of each planet, and the Moon, we can assess if life could potentially exist in these environments.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss why the Moon and Mercury are very unlikely places for any life
- Discuss the possibility of life in the clouds of Venus
- Describe the main properties of the jovian gas giants and any possibilities for life
- Discuss different ideas on panspermia

Mercury and the Moon

Mercury and the Moon look very similar in photos, as seen in Figure 1 – both have surfaces that are heavily cratered and do not appear to have any clouds. The number of craters on a world is related to how geologically active that world is – active volcanoes, for example, will “pave” over craters and create a smooth surface. Mercury and the Moon, thus, appear to be geologically inactive.

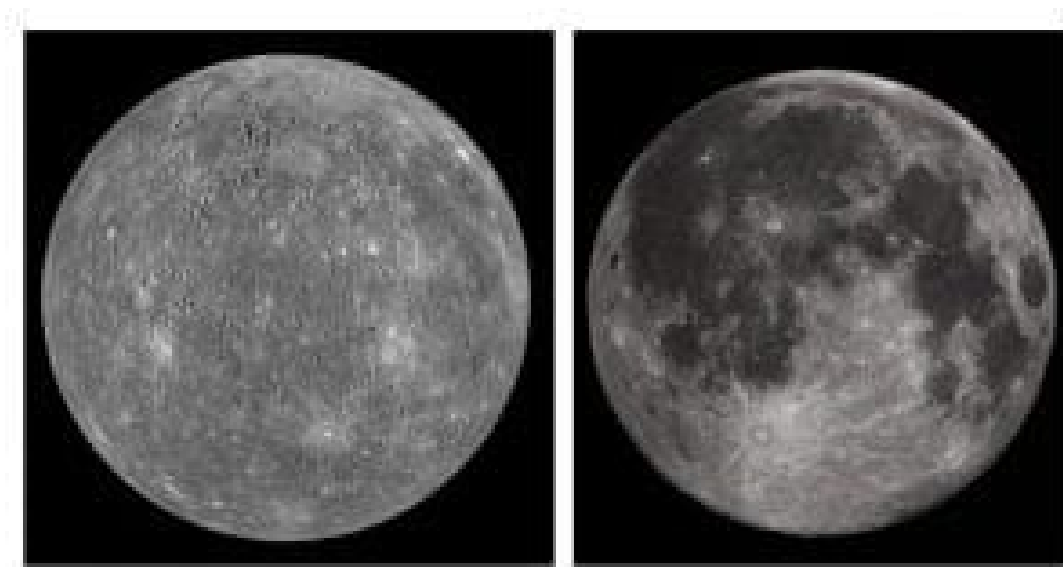


Figure 1 – Mercury (left) and the Moon (right). Composite image of Mercury taken by MESSENGER. **Compare the cratered surfaces of these worlds – what does this tell us about erosive processes and atmospheres on Mercury and the Moon?**

For most planets, their size and distance from the Sun can tell us most of what we need to estimate if they have the conditions to host life. Moons can also have the conditions for life but this depends on their distance from their host planet; we will see some examples of this with the jovian moons Europa and Enceladus. As noted in the previous chapter, the size of a world determines how quickly it will lose heat that is inside of it; the rate of heat loss depends inversely on the radius of the world. Mercury has a radius that's only 38% the size of the Earth's radius, and the Moon's radius is 27% of the Earth's, so both of these worlds lose their heat more quickly than larger terrestrial planets like the Earth and Venus.

For a world to be geologically active, there needs to be an internal heat source. Any heat that is left inside Mercury or the Moon is not enough to drive geological activity today, so we say that the Moon and Mercury are geologically dead. While it is not a strict requirement that a world be geologically active to harbor life (as far as we know), it has been extraordinarily beneficial for life on

Earth, as plate tectonics and volcanism help to keep the Earth's climate stable.

The Earth's atmosphere is crucial to life on the surface of our planet. The surface gravity of the Moon is just one-sixth that of Earth and it is too weak to hold on to an atmosphere. Without a protective atmosphere to both shield and store energy from the Sun, the surface of the Moon experiences huge temperature swings. Without a global magnetic field, the Moon's surface is bombarded by charged particles from the solar wind. The small traces of water that have been detected on the Moon are frozen as ice at the polar regions or in permanently shaded craters.

Like the Moon, Mercury does not have a significant atmosphere. However, Mercury a weak magnetic field has been detected on the innermost planet in our solar system, and this could help protect the surface from some of the pummeling by the solar wind. Because Mercury spins slowly, the days are very long on Mercury: one day on Mercury is the equivalent of 176 days on Earth. This leads to huge swings in Mercury's temperature — temperatures reach a sizzling 430 °C (800 °F) on the sunlit side and drop to -170 °C (-274 °F) on the shaded side; water ice has been found in craters at Mercury's north pole. Such a wide range in temperatures does not seem conducive for any type of life.

Want to know more: Liquid water

Since all life on Earth requires water, it makes sense to also assess if liquid water might exist on the surface of a world. A phase diagram shows the pressure (P) and temperature (T) ranges for the liquid-gas, solid-liquid, and solid-gas phase transitions (sometimes these diagrams are called P-T diagrams). Figure 2 below shows the phase diagram for water. The surface pressure for the Earth is shown for reference as the red line at ~ 100 kPa (or 1 atm), with the freezing and boiling points of water indicated at 0 °C and 100 °C.

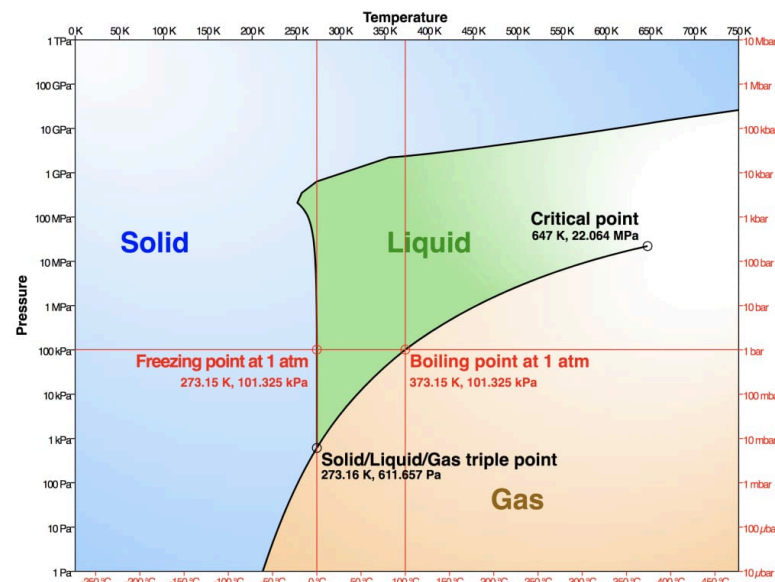


Figure 2 – The phase diagram for water. Note that the y-axis is a log scale. **Identify the range of temperature and pressure for the inhabited surface of Earth.**

The Moon and Mercury have no atmospheres and therefore they have extremely low surface pressures. Water ice can exist in the coldest regions on the Moon and Mercury but can not make the transition to liquid water — the solid ice will go directly to gas (that is, it sublimates) at temperatures higher than the solid-gas transition point.

Concept Check: Phases of water

The surface pressure on Mars today ranges from 30 Pa on the heights of Olympus Mons to 1,115 Pa in the deep depression of Hellas Planitia. The average surface pressure on Mars is about 650 Pa. Can liquid water exist on the surface of Mars? If so, what temperature is needed?

Venus

Venus is sometimes referred to as Earth's twin because the two planets have similar sizes and densities. From a great distance, Venus might be mistaken for a habitable world, and in the very distant past, it is possible that oceans of water and life did exist on Venus. Pictures of the Venusian surface show clouds that with a pinkish hue, and Venus shows evidence of recent volcanism^[1]. But the runaway greenhouse that occurred on Venus has made the planet inhospitable for life today. The loss of water and the build up of a thick carbon dioxide atmosphere has led to a crushing atmospheric pressure, 90 times stronger than experienced on Earth's surface.

This dense atmosphere leads to a bone-dry surface with a fairly uniform temperature of 850 °F on Venus. The Venusian surface is too hot for any (known) life to survive on. When the Venera 13 spacecraft touched down on Venus in 1982, it captured data for about 2 hours before the electronics were fried and the pressure crushed it to death (Figure 3).

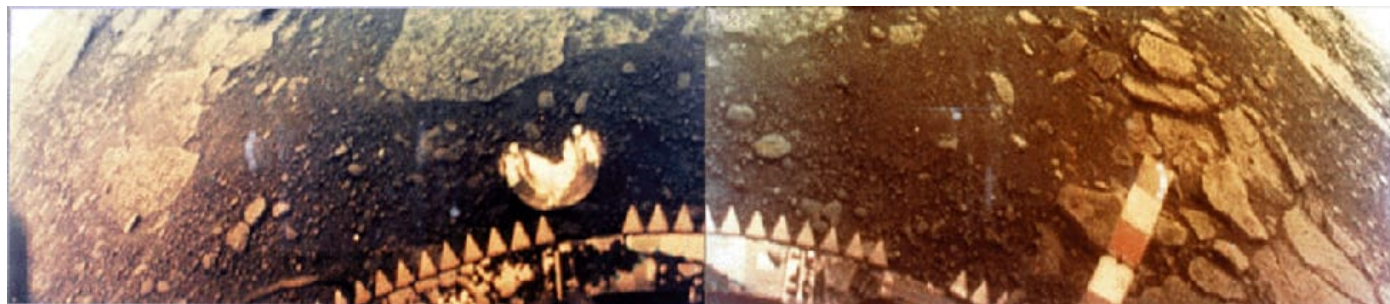


Figure 3 – Surface of Venus. These wide-angle views of the surface of Venus are from the Venera 13 spacecraft. Everything is orange because the thick atmosphere of Venus absorbs the bluer colors of light. The horizon is visible in the upper corner of each image.

Another big difference between Earth and Venus is the rotation rates – how fast the planets spin: a day on Venus lasts 243 Earth days. As evidenced by volcanic mountains Venus does have a molten layer to its core but the slow rotation rate impedes the formation of a global magnetic field. Without a magnetosphere for protection from the solar wind, the top layers of Venus' atmosphere are slowly being stripped away.

Life in the Clouds of Venus?

Before completely dismissing the possibility of life on Venus, let us look at its atmosphere in more detail. The temperature of the Venusian atmosphere varies with altitude, as seen in Figure 4. In the upper troposphere, between 30 and 60 kilometers above the surface, a thick cloud layer is composed primarily of sulfuric acid droplets. Sulfuric acid (H_2SO_4) is formed from the chemical combination of sulfur dioxide (SO_2) and water (H_2O). Sulfur dioxide is added to the atmosphere via volcanic eruptions on Venus' surface. In the atmosphere of Earth, SO_2 is one of the primary gases emitted by volcanoes, but it is quickly diluted and washed out by rainfall. The temperature in parts of this cloudy region are between 273–373 K, making this a layer with the conditions for habitability. Any bacteria that could thrive in this zone would take in sulfur as a nutrient. On Earth, bacteria that survive on sulfur are found in rocks that date back to around 3 billion years, prior to [the Great Oxidation Event](#).

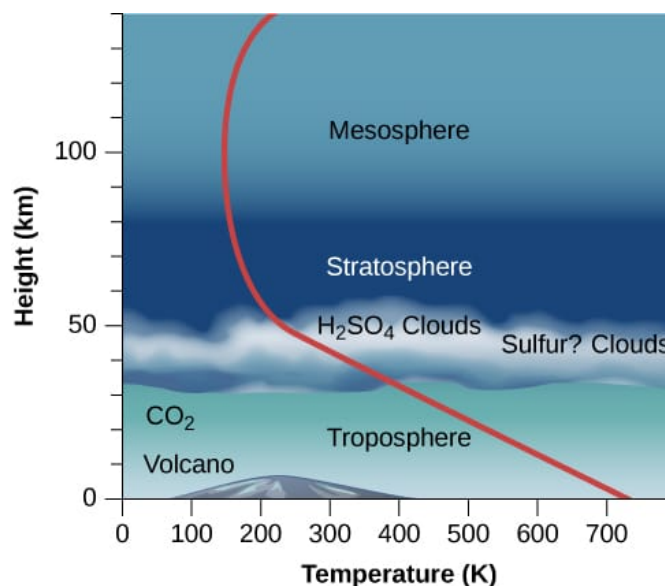


Figure 4 – Venus' Atmosphere. The layers of the massive atmosphere of Venus shown here are based on data from the Pioneer and Venera entry probes. Height is measured along the vertical axis, the horizontal axis shows temperature, and the red line allows you to read off the temperature at each height. Notice how steeply the temperature rises below the clouds, thanks to the planet's huge greenhouse effect. **What is the temperature on the surface of Venus and how does this compare to Earth?**

In an article published in the journal *Nature* in 1967, “Life in the Clouds of Venus?”, co-authors Harold Morowitz and Carl Sagan note that:

“While the surface conditions of Venus make the hypothesis of life there implausible, the clouds of Venus are a different story altogether. As was pointed out some years ago, water, carbon dioxide and sunlight—the prerequisites for photosynthesis—are plentiful in the vicinity of the clouds. Since then, good additional evidence has been provided that the clouds are composed of ice crystals at their tops, and it seems likely that there are water droplets toward their bottoms. Independent evidence for water vapour also exists.”

No life has yet been detected in Venus' clouds, although there was cautious excitement in 2020 when the detection of phosphine (PH₃) in the atmosphere of Venus was announced. The “Extraordinary Claims” box below describes the arc of this announcement and subsequent denouncement.

Extraordinary Claims: Life in the clouds of Venus

In September 2020, an article was published in the journal *Nature* titled “Phosphine gas in the cloud decks of Venus.” The article immediately gained widespread attention, as the molecule phosphine (PH₃) is produced on Earth by bacteria. In other words, phosphine could be a potential **biosignature**, or sign of life, on Venus. The authors note in the paper's abstract that “PH₃ could originate from unknown photochemistry or geochemistry, or, by analogy with biological production of PH₃ on Earth, from the presence of life.”

The claimed detection of PH₃ was based on spectroscopic data taken using two different telescopes. Very small amounts of phosphine are found in Earth's atmosphere, primarily from microbial life living in very low-oxygen environments (phosphine is toxic to humans). Phosphine can also be produced during the production of methamphetamine, which is also a man-made, or biological, source. Because of the potentially revolutionary impact of this discovery if true — this would be the first detection of any life elsewhere in the cosmos — other scientists were eager to verify the detection and rule out any ways that PH₃ could be produced naturally on Venus.

In February 2021, a study was published by astrobiologists at the University of Washington that used atmospheric models to reproduce the PH₃ detection but found instead that sulfur dioxide (SO₂) was a far more likely explanation than phosphine. NASA's SOFIA mission also searched for the PH₃ line but did not find any evidence for it. Others have speculated on ways that PH₃ could be produced abiotically (meaning *not* by life) on Venus, including by phosphorous from a volcanic eruption interacting with sulfuric acid in Venus' clouds.

While the case is not totally closed on the possibility of PH_3 -based life in the clouds of Venus, other studies would need to corroborate the evidence for this claim to be accepted.

References

Bandari, A. (2022, November 29). *No Phosphine on Venus, According to SOFIA – SOFIA: Stratospheric Observatory for Infrared Astronomy*. <https://blogs.nasa.gov/sofia/2022/11/29/no-phosphine-on-venus-according-to-sofia/>

Greaves, J. S. et al. (2021). Phosphine gas in the cloud decks of Venus. *Nature Astronomy*, 5(7), 655-664.

Lincowski, A. P. et al. (2021). Claimed detection of PH_3 in the clouds of Venus is consistent with mesospheric SO_2 . *The Astrophysical Journal Letters*, 908(2), L44.

Witze, A. (2021). Life on Venus claim faces strongest challenge yet. *Nature*, 590(7844), 19–20. <https://doi.org/10.1038/d41586-021-00249-y>

The Gas Giant Planets

The jovian planets are gaseous (Jupiter and Saturn) or a slushy ice mix (Uranus and Neptune) and do not have solid surfaces; thus, any life on these giant planets seems virtually impossible. These planets orbit far from the Sun and do not intercept enough solar energy for life as we know it to thrive (the upper atmosphere of Jupiter, for example, has an average temperature of -110°C ^[2], much colder than the lowest known temperature for any psychrophiles), especially without surface chemistry to spark life and sustain evolution. That said, we can still make a critical assessment similar to that made for Venus to see if there are layers in the atmospheres of the gas giants could potentially have conditions that could harbor life.

The Jovian Atmospheres and Clouds

Looking at the atmospheres of the jovian giants in Figure 5, we see that all of the crucial elements for life — CHNOPS — are all seen in various molecules. For Jupiter and Saturn, there is a temperate region which has water (H_2O) droplets. There are also organic compounds, such as methane, that may provide a source of nutrients, and there is still leftover heat from formation that keeps the atmosphere warm, at least close to the surface.

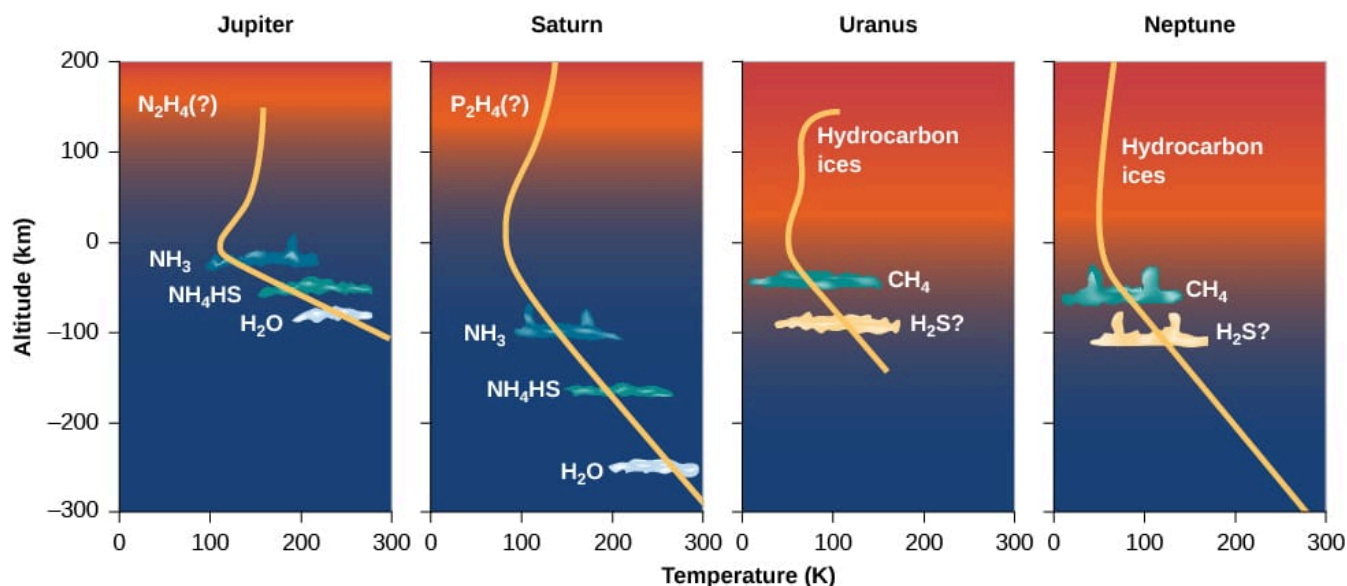


Figure 5 – Atmospheric Structure of the Jovian Planets. In each diagram, the yellow line shows how the temperature (see the scale on the bottom) changes with altitude (see the scale at the left). The temperature profiles show where different molecules might condense in the planet atmospheres. **Are there layers in any of these atmosphere that mimic habitable Earth conditions?**

The possible layer of water clouds in Jupiter's clouds present an intriguing environment for possible life. In the 1970s, Carl Sagan and E. E. Salpeter imagined an entire ecosystem existing within this layer, with organic life forming from methane and electricity.

Once formed, these organisms could exist on energy from the Sun and nutrients in the atmosphere (such organisms are called phototrophs). Evolution could act on these organisms and larger, buoyant life forms that can adapt to different pressure levels in the atmosphere could exist and compete for resources. The downside of this scenario is that the atmosphere of Jupiter is very turbulent, with storms raging for centuries with winds that may be too high to sustain such life.

Carl Sagan discusses the origin of life.

1. How do experiments to synthesize biomolecules in terrestrial labs inform our understanding of the origin of life elsewhere?
2. What might life on Jupiter look like? What are environmental limitations for putative organisms on Jupiter?



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=121>

Video source: @yeoldebrian

Panspermia

In our discussions of the possibility of life elsewhere in our solar system, we have not specifically addressed the question of whether life on Earth could have originated somewhere else and ended up on our planet. This possibility is not unfeasible: we have evidence that other worlds in our solar system, like Venus and Mars, were once habitable places long ago, and perhaps back then a microbe could have hitched a ride to Earth on a meteor. Could life have come from outside our solar system? This seems less likely, merely based on the extreme distances that the life would have to travel and survive before getting to Earth, but still worth considering. A more speculative idea is that an extraterrestrial civilization could have brought life here intentionally.

The video clip below reviews some of the ideas related to panspermia.

1. What are some adaptations of life on Earth that might allow organisms to survive in space?
2. What transport mechanisms might exist for tiny organisms to move to other planets?
3. How long can extremophile organisms remain dormant while traveling through space?



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/>

Interplanetary

If life developed on Earth so quickly, and other terrestrial planets (Venus and Mars) had similarly habitable conditions, it is easy to imagine that life may have evolved independently on those planets around the same time it did on Earth. We have strong evidence, for example from stromatolites and microfossils, that life formed on Earth fairly rapidly (3.5 to 4 billion years ago), perhaps even during the Hadean eon when liquid water oceans first appeared.

If life did form independently on another planet in our solar system, then it could have travelled to Earth in its early stages. There were far more collisions in the early days of the solar system and this could have provided a vehicle for a rock with life to be blasted from one rocky planet and land on another. Of course, life surviving the impact poses challenges but this is nonetheless a plausible scenario. We have learned from extremophiles on Earth that microscopic life can be remarkably resilient. Another possibility is that small dust grains could be released during the impact and these could make their way to another planet (rather than surviving on the rock itself).

Comets, Asteroids, and Meteors

Comets, asteroids and meteors are leftover debris from the formation of the solar system and therefore provide a way of sampling the conditions in the very young solar system, before life emerged on Earth. Could life have formed on these relics? With the exception of the asteroid Ceres (more on that in a bit), these objects are far too small to have any heat inside of them (remember that small worlds lose their heat more rapidly than larger ones). There is therefore no heat to drive geological activity and there are no protective magnetic fields. Adding to this the frigid temperatures that prohibit liquid water from existing, we can rule out any life on these objects.

However, the building blocks for life, amino acids, have been found on the surface of all of these types of objects. For example, the Rosetta mission detected glycine on the comet 67/P. Analyses of the Murchison meteorite, which landed on Earth in Australia in 1969, have shown that it contains more than 90 amino acids, many not found on Earth, including some of the 20 amino acids that are used by all life on Earth. Organic materials have also been found on the asteroid Ryugu, from which a sample was collected by the Japanese Aerospace Exploration Agency's Hayabusa2 mission and successfully brought back to Earth for analysis. (The OSIRIS-REX mission dropped off samples from asteroid Bennu on Earth in September of 2023.) While life does not exist on these bodies, they could have brought the ingredients for life to Earth or others places in the solar system via a collision. This is strictly not panspermia, since life itself was not delivered, and is sometimes called pseudo-panspermia to make this distinction clear.

Life on Ceres?

The asteroid Ceres, orbiting in the asteroid belt between Mars and Jupiter, is large enough to also be classified as a dwarf planet. Ceres was explored in detail by NASA's Dawn mission from 2015-2018. Dawn detected organic molecules on Ceres' surface, some of them the building blocks for amino acids used for life on Earth, although it is not clear if these organics formed on Ceres or were brought there from other bodies. Dawn studied bright white spots on Ceres (Figure 6) that are associated primarily with the central peaks of large craters, such as the Occator crater, and found that the light-colored mineral is primarily salt, released from the interior^[3]. After repeated close flybys, data from the Dawn spacecraft indicated that Ceres has (or has had) briny water beneath its surface that can be brought up through volcanoes. The most dramatic is the 4 kilometer tall ice volcano called Ahuna Mons. This raises the possibility that Ceres is still geologically active today, and that it was once a habitable world.

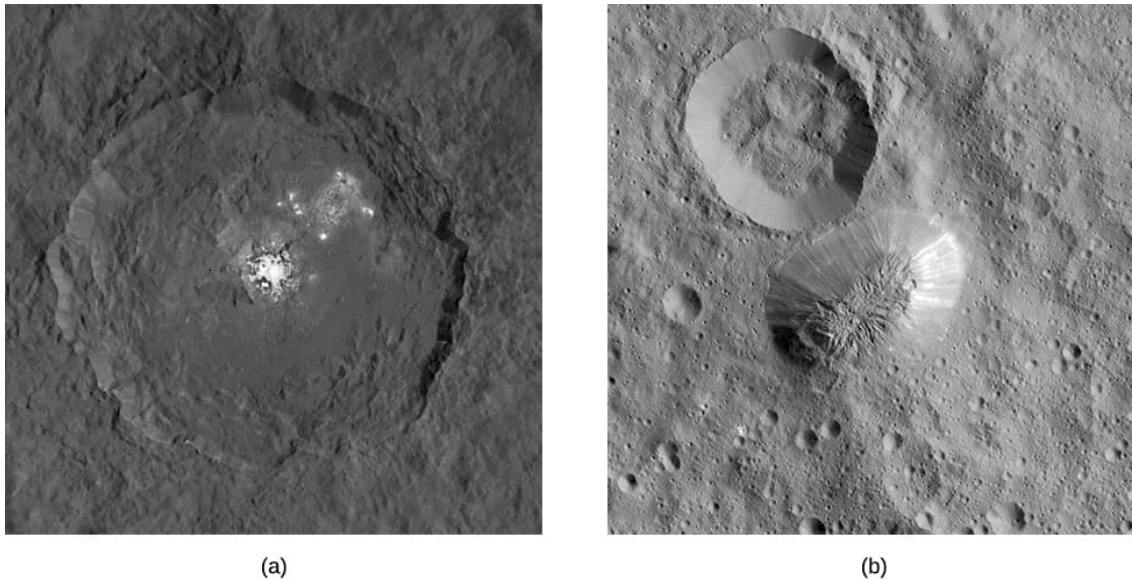


Figure 6 – White Spots in a Large Crater on Ceres. (a) These bright features appear to be salt deposits in a Ceres crater called Occator, which is 92 kilometers across. (b) Ahuna Mons is an isolated mountain on Ceres, 4 kilometers high. It is thought to be an intrusion of ice from the interior.

The TEDx Talk below discusses some potential avenues for life coming to Earth via panspermia. After watching the video, consider the following discussion questions:

1. In addition to tardigrades, what type of bacteria can survive harsh environments? What can these bacteria do that makes them particularly robust?
2. Discuss the possibility of life originating in the Jezero crater on Mars.
3. Which two missions are currently searching for signs of organic matter in asteroids and comets? How and where are they searching?



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=121>

Video Credit: @TEDx

Interstellar Panspermia

The principle of life being brought to Earth from another solar system shares the same basic premise as interplanetary panspermia, but with much longer distances for the life to travel. The challenges are the same but on a much larger scale, as the life must survive a longer trip once it is ejected from another solar system. Thus, there is a much lower probability of life arriving from an interstellar rock or comet than from within our solar system. However, this possibility seems more pertinent now that two interstellar objects, 1I/Oumuamua and 2I/Borisov, have been detected in our solar system, where the designation “I” stands for Interstellar. Oumuamua is believed to be an asteroid and Borisov is likely a comet, so we have confirmation that objects from other star systems can reach our own. In addition to interstellar asteroids or comets, life could potentially be transferred via much smaller dust grains (on the order of microns, or 10^{-6} meters) that still can potentially provide protection from dangerous radiation on a journey between stars.

Directed Panspermia

There is of course another possibility, that life was planted on Earth by a technologically advanced extraterrestrial civilization. This idea may seem far-fetched but it has received serious consideration, notably from Francis Crick (who co-discovered the double-helix structure of DNA) and Leslie Orgel. In 1973, Crick and Orgel published a paper describing their idea of so-called “directed panspermia,” in which life on Earth was brought here as microbes on a spacecraft from another planet that is occupied by advanced life^[4]. One of their arguments in favor of the idea that Earth was “infected” with alien life is that humans could conceivably do that same thing; that is, we could send microbes out to seed another world with life. They call this premise the idea of “cosmic reversibility” — if we could do it, so could another civilization. One way to find evidence that we were infected in this way would be to search for signs of it in the genetic code. More recent ideas of directed panspermia suggest that the extraterrestrial civilization would use a comet to spread life, rather than a spaceship.

Key Concepts and Summary

Small bodies like the Moon and Mercury do not have sufficient surface gravity to hold on to an atmosphere and have also lost most of their internal heat. Venus is similar in size and mass to the Earth and may have had habitable conditions at one time. However, a runaway greenhouse effect on Venus produced a massive, crushing atmosphere and raised the surface temperature so that complex biochemistry would only be possible in the upper atmosphere. The gas giant planets, Jupiter and Saturn have massive atmospheres with no real solid surfaces and relatively small high-density cores. The search for life in the solar system is really a search for life as we know it on Earth. If similar biochemistry is found on other solar system bodies, this might support the theory of panspermia – the idea that life originated beyond our planet (or solar system) and was transported here.

Review Questions

Summary Questions

1. Are Mercury and the Moon geologically active today?
2. Why are Mercury and the Moon not likely to host any life?
3. Why can't life exist on the surface of Venus today?
4. What are Venus cloud layers made of, and in which layer could life possibly exist?
5. Discuss the reported detection of phosphine in Venus' clouds and why this was of interest to Astrobiologists. As of today, has the detection been confirmed?
6. What makes the surface and interiors of the jovian gas giants inhospitable for any life?
7. How did Carl Sagan imagine life in the clouds of Jupiter? Do you think this scenario is likely? Explain why or why not.
8. What is the main idea of panspermia?
9. How could life travel between planets?
10. Where else in our solar system have amino acids been found?
11. Could life exist on the asteroid/dwarf planet Ceres? Explain.
12. What additional challenges are posed when considering interstellar panspermia?
13. What is the main idea of directed panspermia? Do you think directed panspermia is likely? Why or why not?

Exercises

1. <https://www.science.org/content/article/active-volcano-shows-venus-living-planet> 4.03: Searching for Life In Our Solar System#return-footnote-121-1
2. <https://science.nasa.gov/resource/solar-system-temperatures/> 4.03: Searching for Life In Our Solar System#return-footnote-121-2
3. Read more: <https://www.nasa.gov/centers-and-facilities/jpl/mystery-solved-bright-areas-on-ceres-come-from-salty-water-below/> 4.03: Searching for Life In Our Solar System#return-footnote-121-3
4. Crick, F. H., & Orgel, L. E. (1973). Directed panspermia. *Icarus*, 19(3), 341-346. [https://doi.org/10.1016/0019-1035\(73\)90110-3](https://doi.org/10.1016/0019-1035(73)90110-3) 4.03: Searching for Life In Our Solar System#return-footnote-121-4

This page titled 4.3: Searching for Life In Our Solar System is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by .

4.4: The Search for Life on Mars

Although it is dry and seemingly barren today, our next-door neighbor Mars may have once been a habitable planet. Finding any life on Mars, even a microbe, would represent a leap in our understanding of life in the universe, as it would mean that life existed in two separate locations in one solar system. Humans have been exploring the Red Planet with unmanned spacecraft since the 1960s and a great deal of evidence has emerged in support of a warmer and wetter Mars in the past. In this chapter, we review some of the key findings that point toward a once habitable Mars.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss findings from early missions in the 20th century to explore Mars
- Describe the terrain and geological features of Mars
- Discuss several lines of evidence that suggest Mars had liquid water on its surface in the past
- Describe the various rover space missions to Mars and what they have found so far

History and Early Missions

The possibility that Mars hosts, or has hosted, life has a rich history dating back to at least the middle of the nineteenth century. William Herschel observed Mars in 1783 and noted that Earth and Mars share some commonalities, such as having an atmosphere and going through seasonal cycles. By 1840, the length of a day on Mars was known to be very similar to that for the Earth, at just over 24 hours. These similarities between Earth and Mars fueled speculation that Mars, too, could harbor life. The Italian astronomer Giovanni Schiaparelli made detailed sketches of the Martian surface in 1877 and he labeled some straight line features as “canali,” an Italian word that can be translated to mean channels or canals, where channels are a natural geological feature and canals are manmade. That same year, 1877, American astronomer Asaph Hall discovered two small moons orbiting Mars; Hall named these moons Phobos and Deimos, meaning fear and panic, after the sons of Ares (the Greek counterpart of Mars).

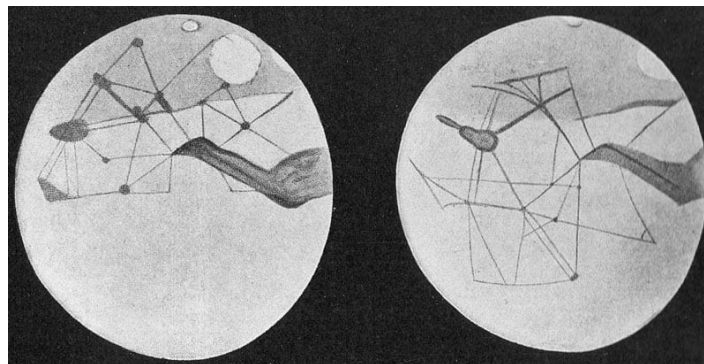


Figure 1 – Percival Lowell's sketches of Mars, circa 1890s.

The American astronomer Percival Lowell, the brother of the poet Amy Lowell, was intrigued by the possibility of life on Mars suggested by Schiaparelli's straight line “canali” features. Lowell, who came from a wealthy Boston family, built an observatory in Flagstaff, Arizona, called Lowell Observatory, in 1894 to undertake his observations of the surface of Mars. Lowell also published several books about the possibility of life on Mars, including *Mars As The Abode Of Life* and *Mars And Its Canals*. Not all scientists subscribed to Lowell's view that Mars was inhabited, and in 1894 Edward Barnard made his own observations and did not see the straight line canals observed by Schiaparelli and Lowell. To add to the public's imagination, H.G. Wells published his sci-fi novel *The War of the Worlds* in 1897.

The idea that Mars was inhabited by advanced life persisted throughout a good deal of the 20th century. The book *Intelligent Life in the Universe*, written by I.S. Shklovskii and Carl Sagan in 1966, includes a chapter dedicated to the possibility that the two moons of Mars are not natural but were artificially created by inhabitants of Mars. This was one year after the first images of Mars

came in from the Mariner 4 mission. English artist David Bowie chimed in in 1971 with his hit song “Life on Mars?”. In 1976, prior to the Viking missions landing on Mars, Carl Sagan and Joshua Lederberg published the article “The prospects for life on Mars: A pre-Viking assessment” in the journal *Icarus*. In the abstract, they note:

“Large organisms, possibly detectable by the Viking lander cameras, are not only possible on Mars; they may be favored.”

Mariner and Viking Missions

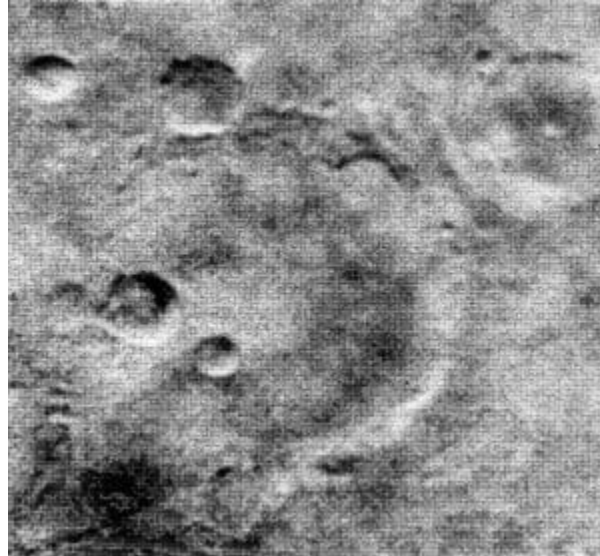


Figure 2. The first picture taken of the surface of Mars. This photo was taken by Mariner 4 and was transmitted back to Earth in 1965.

A global view of Mars emerged after images from the Mariner flyby and orbiter missions were transmitted back to Earth. The first visitor was the US Mariner 4, which flew past Mars in 1965 and transmitted 22 photos to Earth. These pictures showed an apparently bleak planet with abundant impact craters. In those days, craters were unexpected; some people who were romantically inclined still hoped to see canals or something like them. In any case, newspaper headlines sadly announced that Mars was a “dead planet.” In 1971, NASA’s Mariner 9 became the first spacecraft to orbit another planet, mapping the entire surface of Mars at a resolution of about 1 kilometer and discovering a great variety of geological features, including volcanoes, huge canyons, intricate layers on the polar caps, and channels that appeared to have been cut by running water. Mariner 9 also collected data about the Martian atmosphere using two spectrometers on board. Geologically, Mars didn’t look quite so dead after all. The data from Mariner 9 allowed the Viking 1 and 2 missions to launch in 1975 and successfully land on Mars in 1976.

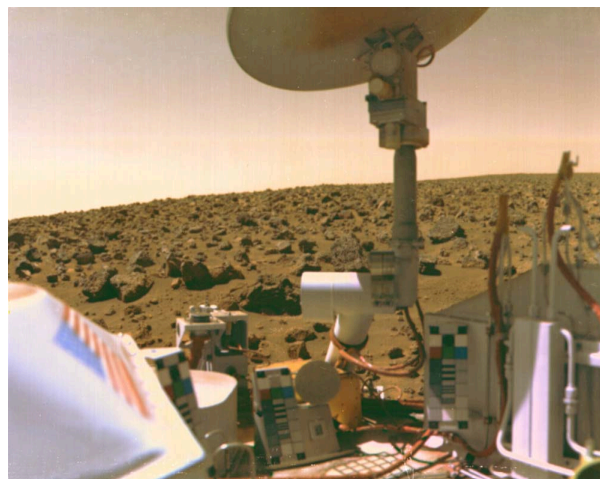


Figure 3 – The Viking 2 lander at Utopia Planitia on Mars.

The Viking camera’s did not reveal any large organisms on Mars (unless they hopped on the cameras to avoid being seen). But it did show a compelling view of the rock-strewn rusty surface of Mars. The Viking landers landed in different regions of Mars: Viking 1 at Chryse Planitia and Viking 2 at Utopia Planitia. Both of these regions are fairly smooth plains in the northern lowlands

of Mars and were chosen to help ensure a safe landing, but they also contain features that, from previous images, looked like they could be associated with water. The Viking missions contained the tools to perform soil experiments on Mars, with the goal of looking for evidence of life. These experiments are discussed in detail below.

Extraordinary Claims: The Face on Mars



Figure 4 – the “face” on Mars.

People like human faces. We humans have developed great skill in recognizing people and interpreting facial expressions. We also have a tendency to see faces in many natural formations, from clouds to the man in the Moon. One of the curiosities that emerged from the Viking orbiters’ global mapping of Mars was the discovery of a strangely shaped mesa in the Cydonia region that resembled a human face. Despite later rumors of a cover-up, the “Face on Mars” was, in fact, recognized by Viking scientists and included in one of the early mission press releases. At the low resolution and oblique lighting under which the Viking image was obtained, the mile-wide mesa had something of a Sphinx-like appearance.

Unfortunately, a small band of individuals decided that this formation was an artificial, carved sculpture of a human face placed on Mars by an ancient civilization that thrived there hundreds of thousands of years ago. A band of “true believers” grew around the face and tried to deduce the nature of the “sculptors” who made it. This group also linked the face to a variety of other pseudoscientific phenomena such as crop circles (patterns in fields of grain, mostly in Britain, now known to be the work of pranksters).

Members of this group accused NASA of covering up evidence of intelligent life on Mars, and they received a great deal of help in publicizing their perspective from tabloid media. Some of the believers picketed the Jet Propulsion Laboratory at the time of the failure of the *Mars Observer* spacecraft, circulating stories that the “failure” of the *Mars Observer* was itself a fake, and that its true (secret) mission was to photograph the face.

The high-resolution *Mars Observer* camera (MOC) was reflown on the *Mars Global Surveyor* mission, which arrived at Mars in 1997. On April 5, 1998, in Orbit 220, the MOC obtained an oblique image of the face at a resolution of 4 meters per pixel, a factor-of-10 improvement in resolution over the Viking image. Another image in 2001 had even higher resolution. Immediately released by NASA, the new images showed a low mesa-like hill cut crossways by several roughly linear ridges and depressions, which were misidentified in the 1976 photo as the eyes and mouth of a face. Only with an enormous dose of imagination can any resemblance to a face be seen in the new images, demonstrating how dramatically our interpretation of geology can change with large improvements in resolution. The original and the higher resolution images can be seen in Figure 5 below.

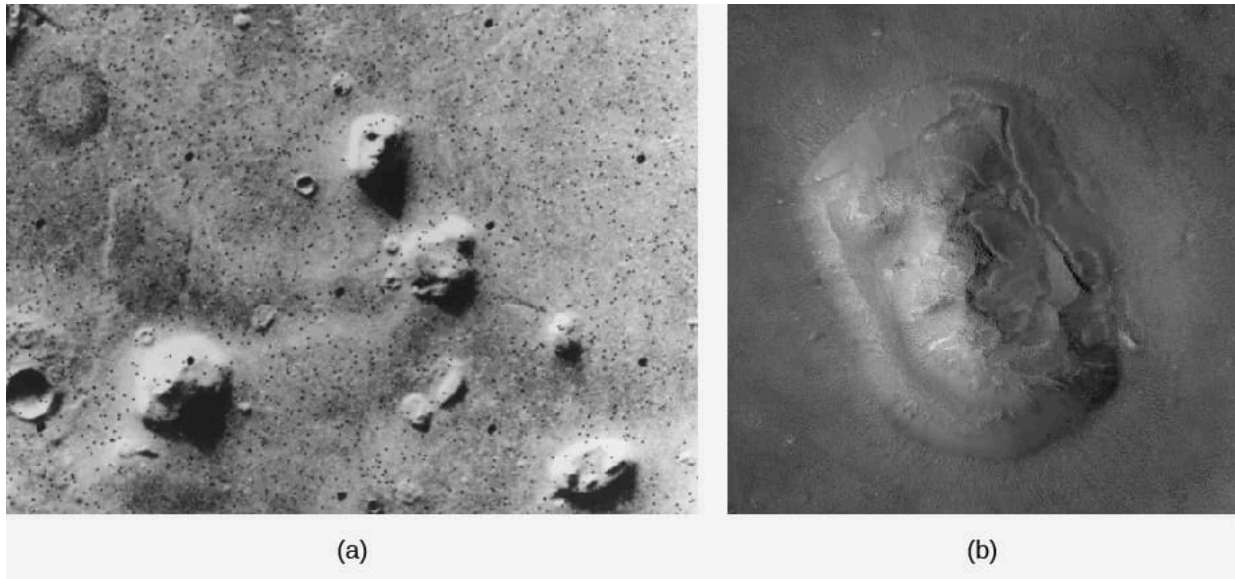


Figure 5 – Face on Mars. The so-called “Face on Mars” is seen (a) in low resolution from Viking (the “face” is in the upper part of the picture) and (b) with 20 times better resolution from the Mars Global Surveyor.

After 20 years of promoting pseudoscientific interpretations and various conspiracy theories, can the “Face on Mars” believers now accept reality? Unfortunately, it does not seem so. They have accused NASA of faking the new picture. They also suggest that the secret mission of the *Mars Observer* included a nuclear bomb used to destroy the face before it could be photographed in greater detail by the *Mars Global Surveyor*.

Space scientists find these suggestions incredible. NASA is spending increasing sums for research on life in the universe, and a major objective of current and upcoming Mars missions is to search for evidence of past microbial life on Mars. Conclusive evidence of extraterrestrial life would be one of the great discoveries of science and incidentally might well lead to increased funding for NASA. The idea that NASA or other government agencies would (or could) mount a conspiracy to suppress such welcome evidence is truly bizarre.

Alas, the “Face on Mars” story is only one example of a whole series of conspiracy theories that are kept before the public by dedicated believers, by people out to make a fast buck, and by irresponsible media attention. Others include the “urban legend” that the Air Force has the bodies of extraterrestrials at a secret base, the widely circulated report that UFOs crashed near Roswell, New Mexico (actually it was a balloon carrying scientific instruments to find evidence of Soviet nuclear tests), or the notion that alien astronauts helped build the Egyptian pyramids and many other ancient monuments because our ancestors were too stupid to do it alone.

Geology of Mars

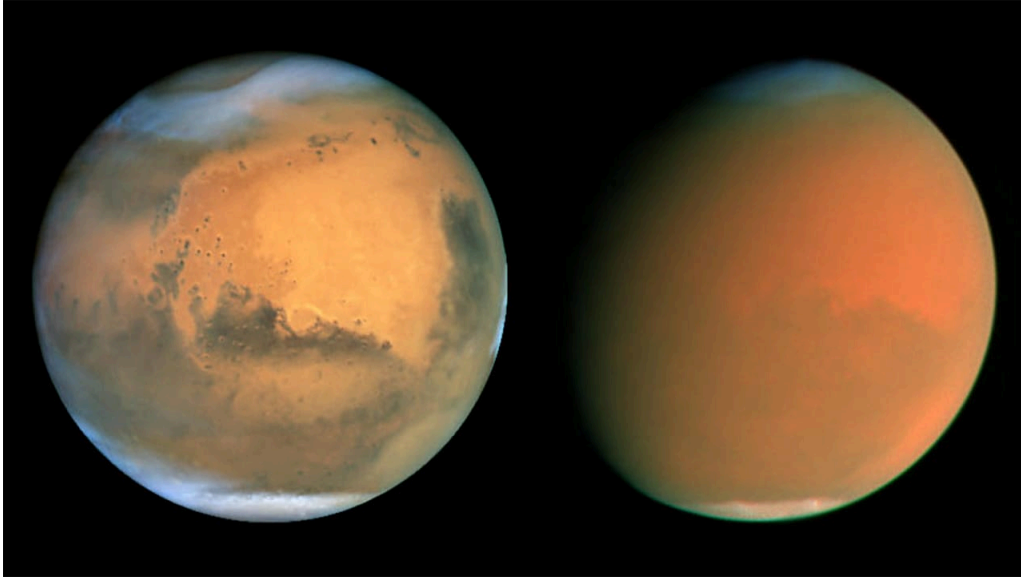


Figure 6 – Two views of Mars taken four months apart in 2001 by the Hubble Space Telescope. The left photo was taken in June 2001, when a dust storm was just beginning. By September (right photo), Mars was enshrouded in dust from the storm. You can still see Mars' polar ice caps in the right photo but just barely.

After the Viking missions concluded, Mars was not visited by spacecraft again until the 1990s. Beginning with the Mars Pathfinder and Mars Global Surveyor missions in 1997, there have been over a dozen successful missions to explore Mars in the past 25 years. Each mission has contributed in various ways to our global understanding of Mars as a planet. Several missions, which we highlight in the next section, have had a specific goal of assessing the habitability of Mars or looking for signs of life.

The map below (Figure 7) was created using an altimeter — an instrument that uses lasers to measure the altitude (or height) of different parts of the surface — onboard the Mars Global Surveyor. The map is color-coded by the height of the surface, where the yellow regions are “ground level” at 0 km, blue/purple regions are at depths of about 5-7 km, and the grey/white regions have the highest altitudes, at over 12 km high for the white circular regions (these white circles are the tops of volcanoes). The Martian volcano Olympus Mons stands at a height of over 22 km, which is more than twice the height of Earth's tallest mountain from base to peak (Mauna Kea in Hawaii, part of which is under water), making Olympus Mons the tallest known mountain in the solar system.

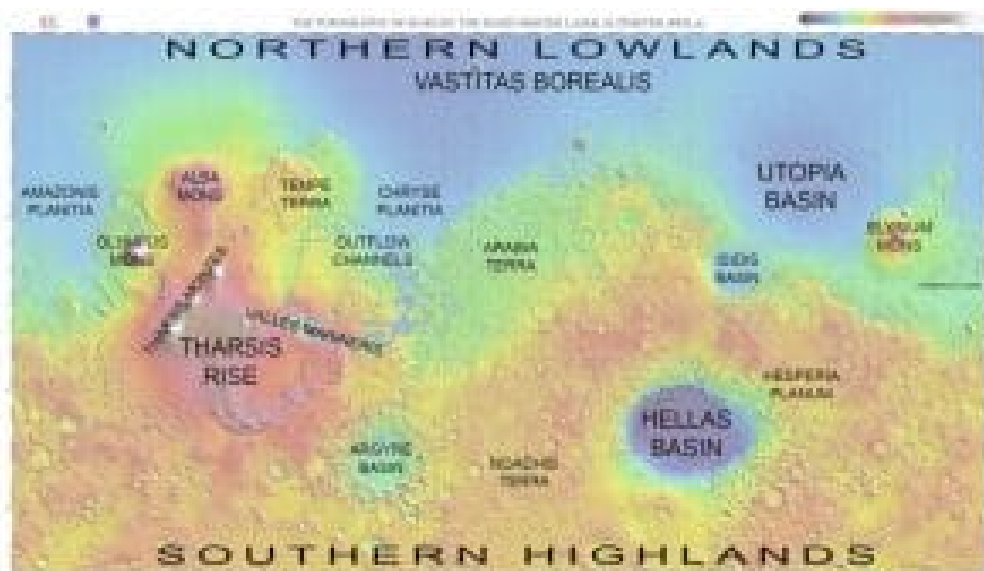


Figure 7 – Topographic map of Mars taken with the Mars Orbiter Laser Altimeter (MOLA). The different colors show elevation, with orange at 0 km, dark purple at about 7 km deep and white corresponding to terrain that is at or above 12 km.

Other notable features are seen on the Martian surface, such as the Valles Marineras (Mariner Valleys, named for the Mariner 9 mission which first revealed them). Valles Marineras is a canyon system that extend for about 5000 kilometers (nearly a quarter of the way around Mars) along the slopes of the Tharsis bulge. If it were on Earth, this canyon system would stretch all the way from Los Angeles to Washington, DC.

Concept Check: the Surface of Mars

Looking at the topographic map above of Mars (Figure 7), which regions are at the highest altitudes (elevation) and which regions are at the lowest altitudes? What do you think created Hellas Basin?

Can you find the Chryse and Utopia Planitia, the landing sites for Viking 1 and 2? Are these regions as deep as the Hellas Basin?

To further explore the terrain of Mars, you can load up Google Mars! Go to <https://www.google.com/mars/> and start exploring. For example, you can click on “Spacecraft” at the top and see the locations of different Martian landers and rovers.

Globally, Mars has two distinct regions: the heavily cratered highlands in the southern hemisphere and the smoother low-lying volcanic plains (the “lowlands”) in the northern hemisphere. The Moon also has similar surface geology, with the smoother plains (the lunar mare) representing a younger surface that was “re-paved” by lava around 3.5 billion years ago. Based on counting the numbers of craters in different regions and analysis of layered rocks, the geological history of Mars can be broken down into three eons: the Noachian, the Hesperian, and the Amazonian.

General Properties of Mars

We can gain some insight into the current conditions on Mars by considering its size and distance from the Sun. Mars is smaller than the Earth, with a radius about half that of the Earth and a mass of about 10% that of the Earth. The density of Mars is 3.9 g/cm^3 (Earth's density is 5.5 g/cm^3), which suggests a composition consisting primarily of silicates but with a small metal core. The smaller size of Mars compared to Earth means that Mars loses its internal heat more rapidly than the Earth, since the cooling rate depends on the radius. Mars has a surface gravity that is just 38% the value of Earth's surface gravity, meaning that material can escape more easily from Mars' gravitational pull. Finally, we note that Mars is about 50% further away from the Sun than the Earth is, meaning that Mars receives less energy from the Sun. Venus is closer to the Sun than Earth and consequently experienced a runaway greenhouse effect. Mars, on the other hand, experienced a runaway “icehouse” effect.

Today, Mars has lost almost all of its original atmosphere because of its low surface gravity and stripping by the solar wind (a stream of highly energetic particles emitted by the Sun). Mars also no longer has a protective magnetosphere so, if any life did currently exist on Mars, it would be challenging for it to survive in a high-radiation environment, which includes cosmic rays.

The Search for Life

In the quest to find life on Mars, either in the past or present, we can look evidence of liquid water and for the essential elements that make up all life on Earth — CHNOPS. Given the commonalities between Mars and Earth noted earlier, such as a similar cycle of seasons and surface geology, we might expect that life could have formed in the same manner around the same time on early Mars as on Earth if there was liquid water in the past.

The Viking Biology Experiments

The earliest missions to Mars provided some hints that liquid water—one of life's primary requirements—may once have flowed on the surface, and later missions have strengthened this conclusion. The NASA Viking landers, whose purpose was to search directly for evidence of life on Mars, arrived on Mars in 1976.

There were four different experiments that the Viking landers were equipped to conduct to look for byproducts of life in the Martian soil. The experiments were: gas-exchange (GEX), pyrolytic release (PR), labeled release (LR) and the gas chromatograph mass spectrometer (GCMS). The GEX and LR experiments both worked by adding nutrients to the soil and then attempting to detect gases released by organisms that consumed the nutrients. The PR experiment simulated the early Martian atmosphere and looked for signs of photosynthetic organisms. GCMS was slightly different in that its goal was to look for organic materials in the soil. Of the four experiments, only the LR produced positive results. However, the LR results have generally been deemed

insufficient evidence for life since the GCMS experiments did not find any organics in the soil. However, some of the scientists who designed the LR experiment still interpret the results as supporting life. Extraordinary claims...

Viking's onboard instruments found no organic molecules (the stuff of which life is made), and no evidence of biological activity in the martian soils it analyzed. The lack of any biological activity in the Martian soil is not particularly surprising because, despite the evidence of flowing liquid water in the past, liquid water on the surface of Mars is **generally not stable today**. Over much of the Martian surface, temperatures and pressures are so low that pure water would either freeze or boil away (under very low pressures, water will boil at a much lower temperature than usual). To make matters worse, unlike Earth, Mars does not have a magnetic field and ozone layer to protect the surface from harmful solar ultraviolet radiation and energetic particles. However, Viking's analyses of the soil said nothing about whether life may have existed in Mars' distant past, when liquid water was more abundant. Water in the form of ice exists in abundance on Mars, not too deep beneath its surface (for example, at the ice caps). Water vapor is also a constituent of the atmosphere of Mars.

Surface Features

Early evidence for surface features shaped by the presence of water came from the Mariner 9 mission, including the Valles Marineris, ancient river channels, tear-drop shaped islands and the presence of ice caps. Mariner 9 observed valley networks and outflow channels. Figure 8 below shows valleys and channels taken by the Viking Orbiter and the Mars Global Surveyor.

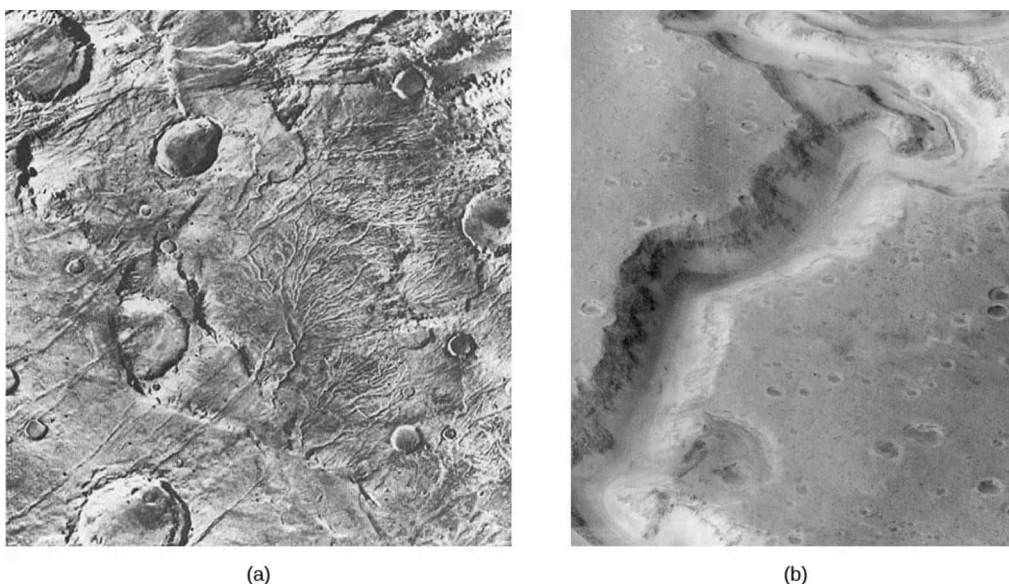


Figure 8 – Runoff and Outflow Channels on Mars. The photo on the left was taken by the Viking Orbiter in 1976 and shows a branched valley network in the old Martian highlands. The photo on the right was taken in 1998 with the Mars Orbiter Camera on board the Mars Global Surveyor mission and shows a channel called Nanedi Valles.

The camera aboard the Mars Global Surveyor provided high-resolution (down to a few meters, which is good enough to see a small truck or bus) images of gullies on Mars. The gullies were found on the steep walls of valleys and craters at high latitudes and appear to be carved by flowing water. These gullies are very young: not only are there no superimposed impact craters, but in some instances, the gullies seem to cut across recent wind-deposited dunes. Perhaps there is liquid water underground that can occasionally break out to produce short-lived surface flows before the water can freeze or evaporate. The Mars Reconnaissance Orbiter, which arrived at Mars in 2006 and is still active today, also took high resolution images of Martian gullies.

Aqueous Minerals

One way to find signs of past water on Mars is to look for minerals that can only form in the presence of water. These are known as aqueous minerals (the word aqueous comes from the Latin word *aqua*, meaning water) and there are several different types.

Clay minerals are found in sediments on the sea floor on Earth, and clays have been suggested as the original site for the **first life on Earth**. They have also been found in a few different locations on Mars, such as the Nili Fossae region.



Figure 9 – Hematite spherules (“blueberries”) seen around the Opportunity rover’s tracks on Mars. The photo is color-enhanced to show the spherules.

In 2004, the Opportunity rover found jarosite and hematite at Meridiani Planum on Mars; these are both minerals that form in a watery environment. Jarosite is a sulfate that can indicate interactions with acidic water. The hematite was found in the form of small spheres that resemble blueberries and as such have taken on this descriptive name. Figure 9 to the right shows blueberries on Mars seen next to the Opportunity rover’s tracks.

The Curiosity rover landed on Mars in 2012 and has now been there for over 4,000 sols (a sol is a Martian day, which is slightly longer than an Earth day). Curiosity landed in Gale Crater, which was selected as the landing site because of evidence that there were clay minerals and sulfates in the sediments on Mt. Sharp. The video below provides some background on Gale Crater and shows Mt. Sharp at its center. Curiosity has been climbing up Mt. Sharp and studying the composition of its layers. As noted in the video clip, the geological layers starting from the bottom provide a way of looking back at the history of the crater and are especially important in understanding when water existed inside the crater in the past.

Guide to Gale Crater



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=131>

Video Credit: @NASAJPL

The strongest evidence of an ancient habitable environment came from analyzing a very fine-grained rock called a mudstone—a rock type that is widespread on Earth but was unknown on Mars until Curiosity found it. The mudstone can tell us a great deal about the wet environments in which they formed. Curiosity has an instrument called CheMin (short for Chemistry and Mineralogy) on board the rover that can analyze soil and rocks and identify what minerals, and in what quantity, are present. The rover first

collects a sample and then deposits it into CheMin for analysis. Figure 10 shows the results for ten sites where samples were collected by Curiosity, starting at the bottom of Gale Crater (Yellowknife Bay) and subsequently as the rover moved up Mt. Sharp. For each site, a pie chart shows what minerals were found. The legend shows the colors used for the pie charts — notice that clay minerals, jarosite and hematite are all found, and their amounts vary as the rover climbs up the mountain. This provides not only confirmation of past liquid water on Mars but also a detailed look back at how the water varied (for example, the clay minerals appear at Yellowknife Bay and then reappear as the rover reaches Murray Buttes).

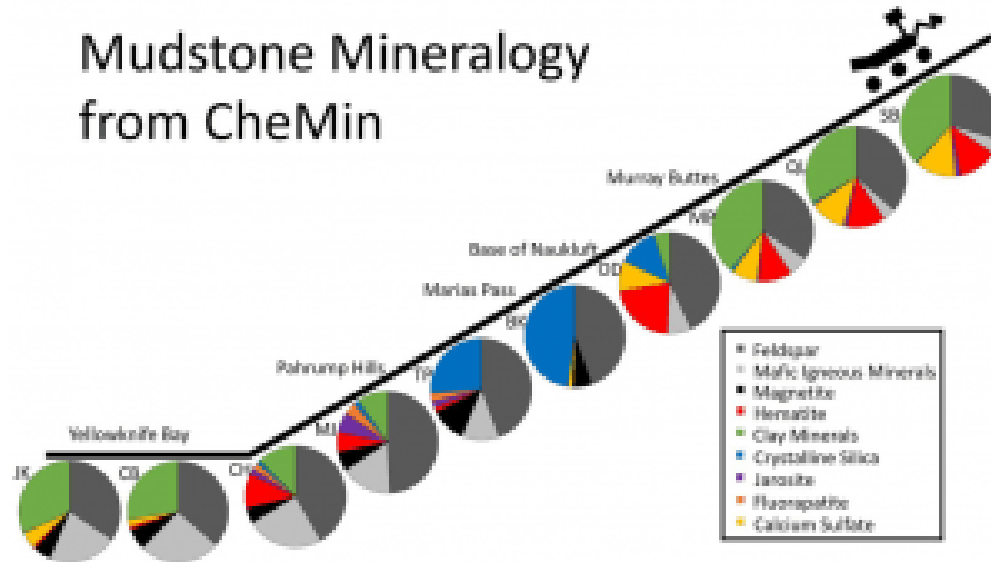


Figure 10 – Mudstones found by the Curiosity rover as it has climbed up Mt. Sharp in the Gale Crater.

Ice

There is water on Mars today but it is in the form of ice. This alone is significant, as it is easy to imagine that some of this ice existed as liquid water when Mars had a warmer environment. Today, ice is found at the polar ice caps, inside craters and below the surface. The Phoenix Mission directly detected water ice on Mars in 2008 (Figure 11). The *Phoenix* spacecraft landed near the north polar cap in summer. Controllers knew that it would not be able to survive a polar winter, but directly measuring the characteristics of the polar region was deemed important enough to send a dedicated mission. The most exciting discovery came when the spacecraft tried to dig a shallow trench under the spacecraft. When the overlying dust was stripped off, they saw bright white material, apparently some kind of ice. From the way this ice sublimated over the next few days, it was clear that it was frozen water. (Sublimation is the process of a solid turning directly into gas when it is heated)

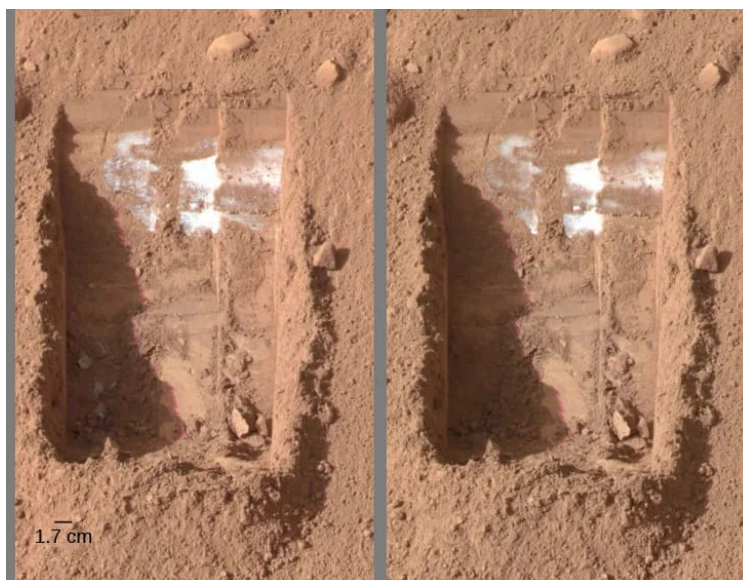


Figure 11 – Evaporating Ice on Mars. We see a trench dug by the Phoenix lander in the north polar region four martian days apart in June 2008. If you look at the shadowed region in the bottom left of the trench, you can see three spots of ice in the left image which have sublimated away in the right image.

The northern polar ice cap is made primarily of water ice, with some frozen CO_2 (“dry ice” we use on Earth). Measurements from the Mars Global Surveyor have established the exact elevations in the north polar region of Mars, showing that it is a large basin about the size of our own Arctic Ocean basin. The ice cap itself is about 3 kilometers thick, with a total volume of about 10 million km^3 (similar to that of Earth’s Mediterranean Sea). If Mars ever had extensive liquid water, this north polar basin would have contained a shallow sea. There is some indication of ancient shorelines visible, but better images will be required to verify this suggestion.

Organic Compounds

Evidence of liquid water flowing in the past clearly points to Mars once being a habitable planet, albeit more than 3 billion years ago. But life as we know it also needs certain chemical elements that are essential to building amino acids, proteins and DNA. The basic building blocks for life are carbon, hydrogen, oxygen, nitrogen, phosphorous and sulfur (CHNOPS). The Perseverance rover, which landed on the surface of Mars in July of 2021 (two other Mars missions, the UAE’s Hope orbiter and China’s Tianwen 1 mission, also reached Mars around the same time) is taking the search for life on Mars one step further than the Curiosity rover. Curiosity’s main goal was to establish if Mars once was habitable, and much evidence from the Curiosity rover described above supports this claim. Perseverance is looking for evidence of potential biosignatures on Mars. If any biosignatures — meaning signs of life — are found by Perseverance, they will still need to be analyzed in great detail to rule out the possibility that they formed naturally rather than by life.

Concept Check

Go to the Mars rock samples site at <https://mars.nasa.gov/mars-rock-samples> and pick one of samples to study. Describe what was collected and how it may contribute to our understanding of life on Mars.

The Perseverance mission is unique — it is the first Mars mission to collect samples that will eventually be sent back to Earth for analysis and it also has a helicopter, named Ingenuity! Perseverance is collecting rock and soil samples as it moves around the crater and visits areas such as an ancient river delta. The samples are sealed into tubes and these will eventually will be picked up by a future mission to Mars; some estimates put the sample return date to Earth in the 2030s.

Perseverance has an instrument appropriately named SHERLOC (the H in SHERLOC stands for Habitable — Scanning Habitable Environments with Raman & Luminescence for Organics & Chemicals) that can detect the presence of organic molecules, as well as aqueous minerals. SHERLOC detected organics in Jezero crater, where the rover is exploring. The Curiosity rover also found organics in Gale crater, but the Perseverance detections have added value as they can be correlated with minerals inside the rocks.

The video clip below gives a tour of Jezero Crater, where Perseverance landed and has been collecting rock and soil samples. In the video, notice that there are mudstones in Jezero crater, just as in the Gale crater that the Curiosity rover is exploring.

NASA's Perseverance Rover explores the surface of Mars.



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=131>

Video Credit: @NASAJPL

Prospects for Life

Five decades of robotic exploration have allowed us to develop a picture of how Mars evolved through time. Early Mars had epochs of warmer and wetter conditions that would have been conducive to life at the surface. However, Mars eventually lost much of its early atmosphere and the surface water began to dry up. As that happened, the ever-shrinking reservoirs of liquid water on the martian surface became saltier and more acidic, until the surface finally had no significant liquid water and was bathed in harsh solar radiation. The surface thus became uninhabitable, but this might not be the case for the planet overall.

Reservoirs of ice and liquid water could still exist underground, where pressure and temperature conditions make it stable. There is recent evidence to suggest that liquid water (probably very salty water) can occasionally (and briefly) flow on the surface even today. Thus, Mars might even have habitable conditions in the present day, but of a much different sort than we normally think of on Earth.

Our study of Mars reveals a planet with a fascinating history—one that saw its ability to host surface life dwindle billions of years ago, but perhaps allowing life to adapt and survive in favorable environmental niches. Even if life did not survive, we expect that we might find evidence of life if it ever took hold on Mars. If it is there, it is hidden in the crust, and we are still learning how best to decipher that evidence.

Want to know more: Rare Earth?

- With the discovery of planets orbiting other stars in the 1990s, enthusiasm was mounting for the search for life on other worlds. In 2000, a popular book, “Rare Earth,” by Ward and Brownlee, suggested that primitive life is very common but technological life is rare. Without a single other example of life beyond Earth, the book spurred a lot of discussion among scientists and the interested public.

The main premise of the book was that too many random, chance conditions were required for homosapiens to evolve. In addition, once our technological species appeared, there were new threats to the survival of life (nuclear war, climate change,

over-population). Some of the essential conditions and circumstances identified by Brownlee and Ward are listed in the Table below.

Essential Characteristic	Reasoning
Right mass planet	Too low mass and the planet could not hang on to its atmosphere; too high mass and the high surface pressure would impede evolution of animals.
Right mass star	Too high mass and the star would have a very short lifetime and emit most of its energy at damaging ultraviolet wavelengths; too low mass and strong magnetic activity from the star would strip away the planet atmosphere.
Habitable zone orbit	Too close and the planet would experience runaway greenhouse warming, like Venus; too far and liquid water would be frozen and biochemical reaction rates would be sluggish.
Atmosphere	An atmosphere is needed to protect life from ultraviolet radiation.
Oxygen	The invention of photosynthesis was required to produce atmospheric oxygen for efficient aerobic metabolisms. Microbes could be successful organisms in anaerobic conditions, but not complex eukaryotes.
Liquid water	Some surface water, but not so much as to cover all of the land (hard for technology to develop under water).
Plate tectonics	To participate in a negative feedback loop and stabilize the climate.
Global magnetic field	Without a global magnetic field, the wind of charged particles from the surface of the host star will strip away the atmosphere.
Large moon	To stabilize the tilt of the planet.
Mars-like world	So prebiotic chemistry and life could get a head start.
Gas Giant outer planet	To act as a sink for incoming comets.
Right location in the galaxy	Too close to the galactic center and high energy radiation from massive stars would be a threat; too far out in the galactic disk and there would be less chemical enrichment.
Mass extinctions	Some mass extinctions to allow for biodiversity, but not too many.

The counter point made by many scientists was that the argument was too anthropomorphic. Some characteristics make sense (right mass planet, atmosphere, water) but others seem less relevant, for a few different reasons:

Not a bottleneck: The circumstances that allowed life on our planet might not be a requirement for life elsewhere. Some characteristics do not present a bottleneck. There are hundreds of billions of “right mass stars,” and very few OB stars with short lives and peak energy emission of damaging ultraviolet radiation. It now appears that there are also billions of planets similar in mass to the Earth and orbiting in the habitable zone. Many of these stars and planets reside in intermediate regions in the galactic disk.

Redundant: Some of the characteristics are interrelated. In the same way that there is redundancy in the requirement: “you have to have a tree and it has to have leaves and it has to have a trunk and branches,” some of the characteristics follow “for free” from others. A large subset of Earth-like planets in the habitable zone will have liquid water, plate tectonics, and global magnetic fields. They will naturally outgas or accrete atmospheres.

Uncertain impact: Other factors in Ward & Brownlee’s list have an uncertain impact. Did we really need Mars to jump-start life on Earth? If we didn’t have Jupiter, would there be fewer objects trapped in the asteroid belt? Would a few more impacts have been a good thing for life on Earth? Does it matter if the obliquity of the Earth precesses at a faster or slower rate?

Mass extinctions fall on the spectrum of natural selection, which must be fundamental for life. It does seem likely that simple microbial life will be far more common than complex organisms. There is likely a pyramid scheme for life, with single cell organisms at the bottom, and complex organisms like whales, velociraptors, and humans near the top. This is exactly what

astrobiologists are trying to learn.

Key Concepts and Summary

Craters on the surface of small bodies like Mars show that erosive processes no longer exist on these objects. The small mass and size of Mars means that this planet has lost most of its atmosphere. However, features on the surface of Mars suggest that this planet once had bodies of surface water, making one of our closest neighbors a key target for exploration with NASA missions. NASA rovers dig beneath the surface and analyze Martian soil for signs of biochemical processes or even simple cellular life.

Review Questions

Summary Questions

1. When did the first dedicated experiments to search for life on Mars begin? Discuss some of the ideas about life on Mars that predated these searches.
2. What are the main properties of the planet Mars (for example, its density and surface gravity)? How do these properties relate to the possible habitability of Mars?
3. How is the geology of Mars similar to that on Earth? How does it differ?
4. What molecules make up the atmosphere of Mars?
5. Does Mars have a magnetic field today, and did it have one in the past? Explain your answer.
6. What are aqueous minerals? How can their presence point toward the conditions for life on Mars?

Exercises

4.5: The Jovian Moons

Although the massive gas and ice giant planets of the outer solar system almost certainly do not harbor any life, some of their large moons are intriguing possibilities. Once presumed to be frigid geologically dead worlds, missions to the outer solar system have shown some of these moons to be dynamic worlds, with active volcanism. Scientists now think that a number of the outer solar system's moons may harbor liquid oceans. Among these, Europa, Titan and Enceladus have thus far been of greatest interest to astrobiologists.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain how tidal heating provides a source of energy for some of the jovian moons
- Discuss “ocean worlds” and why they may harbor life
- Discuss the environment of Saturn’s moon Titan and what types of life could exist
- Describe how lakes beneath ice sheets on Earth could help scientists learn how to explore icy moons

Tidal Heating



Figure 1 -The four Galilean moons of Jupiter are shown to relative scale and arranged in order of their orbital distance from Jupiter: Io (closest; far left), Europa, Ganymede, and Callisto (furthest; far right).

A quick glance at Figure 1 shows that the moons of Jupiter (at least the four largest) look quite different from the Earth’s moon. The moon on the far left is Io and it is the most volcanically active world in our solar system. Io is almost the same size as our Moon: the Moon has a radius of 3476 km, and Io has a radius of 3640 km and a mass that is 1.2 times greater than the Moon’s mass. Additionally, Io is nearly the same distance from Jupiter that the Moon is from the Earth. With just these comparisons in mind, we’d expect Io to have been long geologically dead. So, why is Io so different? Where is the energy coming from to drive the active volcanism on Io?

The difference is the size of Earth compared with the size of Jupiter. Jupiter is a behemoth compared to Earth, with 318 times the mass of the Earth and a radius about 11 times larger than that of Earth. This means that the gravitational pull between Io and Jupiter is much stronger than the pull between the Moon and the Earth. But there’s more to the story with gravity. At any given time, one side of Io is closer to Jupiter. This means that the pull of gravity is stronger on this closer side and the net result is that Io is literally stretched out. On its own, this is not enough to explain Io’s energy source, as we know the Moon is geologically dead and also experienced this kind of stretching. This type of force due to stronger gravity on one side of an object is called a **tidal force**. The difference is that Io is also experiencing gravitational tugs from Europa, Ganymede and Callisto, as shown by the small arrows in part (A) of Figure 2.

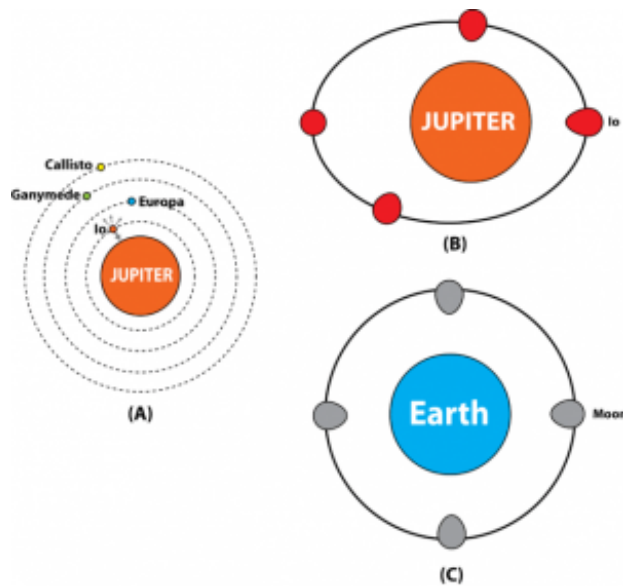


Figure 2 – Tidal flexing. Panel (A) shows the orbits of the four Galilean moons around Jupiter. The arrows surrounding Io indicate the immense gravitational pull that Jupiter exerts on Io. Panel (B) shows a zoom in of Io's orbit around Jupiter. The strong gravity from Jupiter, along with Io's elliptical orbit, create heat from tidal flexing. In panel (C), the Moon's orbit around Earth is shown for comparison. Note that the Moon's orbit is actually more elliptical than Io's; Io's elliptical orbit is exaggerated in B.

The orbits of Io, Europa and Ganymede are locked into an **orbital resonance**. For every 1 orbit around Jupiter that Ganymede makes, Europa makes 2 orbits and Io makes 4 orbits. As a side effect, the same sides of these moons always face Jupiter (we say they are **tidally locked** with Jupiter). When Io receives these periodic tugs from Europa and Ganymede, its orbit becomes more elliptical, meaning it deviates from a perfect circle. This is shown in part B of Figure 2 and the animation below, where Io is stretched and squeezed by different amounts as it orbits Jupiter. This means that Io experiences stronger or weaker tidal forces as it orbits Jupiter, and this continued change in its shape leads to friction and heating. **Tidal heating** is thus the mechanism that supplies Io with a source of heat.

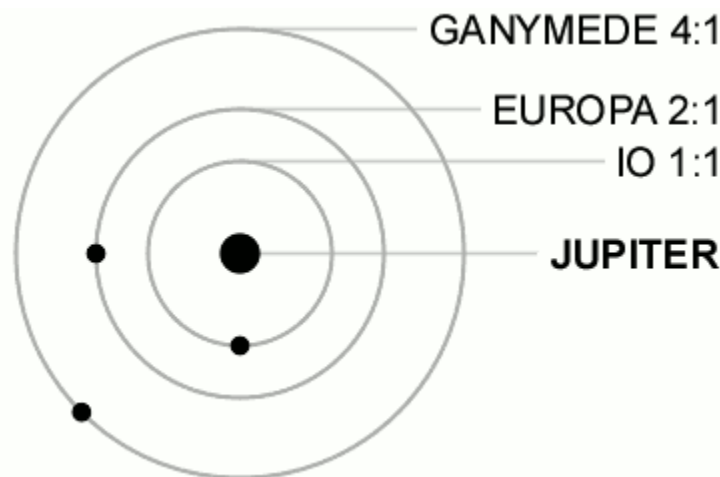


Figure 3 – Animation of Io, Europa and Ganymede's 1:2:4 orbital resonance. The moons light up green or magenta when two are lined up.

The Moon is also tidally locked to the Earth — we always see the same side of the Moon from Earth. But the Earth's orbit is not continually changed like Io's so the Moon's shape is not changing as it orbits Earth (See part C of Figure 2).

Jupiter's Potentially Habitable Moons

Io orbits closest to Jupiter (Figure 1) and so experiences the strongest tidal heating from Jupiter's massive gravity. This effect is so strong that Io is thought to be the most volcanically active body in our solar system. At far right, Callisto shows a surface scarred by billions of years' worth of craters—an indication that the moon's surface is old and that Callisto may be far less active

than its sibling moons. Between these hot and cold extremes, Europa, second from left, orbits at a distance where Jupiter's tidal heating may be “just right” to sustain a liquid water ocean beneath its icy crust.

Io, Ganymede, and Callisto

Because of Io's extreme levels of volcanism, it is very difficult to imagine any kind of life that could survive in this hostile environment. Ganymede is the largest moon in the solar system — it is larger than the planet Mercury — and also experiences some tidal heating from its elliptical orbit around Jupiter. However, an ocean beneath the icy-rock surface of Ganymede would be under a fairly thick crust (150 km) and would be more challenging to explore. Callisto does not receive tidal heating from Jupiter but it is sufficiently large that it may still retain some internal heat and could host a subsurface ocean.

We will not explore Ganymede and Callisto's possibilities for life further here. However, ESA's [JUICE](#) (Jupiter Icy Moons Explore) mission was launched in April 2023 and is on its way to do exactly that. The JUICE spacecraft will reach Jupiter in 2031...stay tuned!

Europa

Jupiter's moon Europa revealed itself to the Voyager and Galileo missions as an active world whose icy surface apparently conceals an ocean with a depth of tens to perhaps a hundred kilometers. As the moon orbits Jupiter, the planet's massive gravity creates tidal heating on Europa, just to a lesser extent than on Io (since Io is closer to Jupiter).

Evidence for a salty water ocean

The first hint that Europa was different from our own Moon came from spectroscopic observations of its surface in the 1950s and 1960s. Recall that we can use spectroscopy — the splitting apart of light from an object — to learn about the chemical makeup of an object. In the case of Europa, the light is reflected sunlight that bounces off the surface of Europa. Water and ice have strong spectral signatures in infrared light, and infrared spectroscopy of Europa's surface showed that it was covered in ice. Images of Europa's surface were taken with the Voyager 1 and 2 spacecraft in 1979 and these images showed a cracked icy surface with few craters. The lack of craters suggested that perhaps liquid water exists beneath the icy crust that can upwell and “pave” over any craters, similar to the way lava wells up from Io's surface and smooths over regions next to the volcanic eruption. Since tidal heating causes volcanic activity on Io, it is consistent that tidal heating could keep any liquid water below the surface of Europa heated.

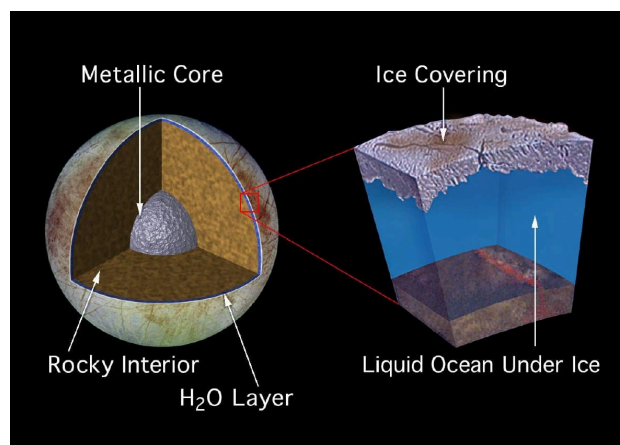


Figure 4 – Image showing the interior structure of Europa (left), with the zoom on the right showing the possible subsurface ocean.

More evidence for Europa's subsurface ocean came from the Galileo mission in the 1990s. Galileo took more close-up images of Europa's surface, adding to the collection taken by the Voyagers, and made measurements of Europa's magnetic field. Just like the Earth, Jupiter has strong magnetic fields and consequently a large magnetosphere surrounding the planet. It turns out that Jupiter's changing magnetic field actually creates a secondary (or “induced”) magnetic field around Europa. Magnetometers aboard the Galileo spacecraft mapped out Europa's magnetic field and showed that it is consistent with a salty, liquid water layer beneath Europa's icy surface.

Other evidence supporting the existence of the liquid water ocean come from the detection by the Hubble Space Telescope of water vapor plumes erupting from Europa's surface (this same idea of cryovolcanism is relevant for Saturn's moon Enceladus below); measurements of Europa's moment of inertia, which depends on its internal structure, also support the liquid water hypothesis. The video clip below summarizes the evidence for Europa's global liquid water ocean.

Europa



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=138>

Video Credit: @NASAJPL

Source of Energy

Europa has probably had an ocean for most or all of its history, but habitability requires more than just liquid water. Life also requires energy, and because sunlight does not penetrate below the kilometers-thick ice crust of Europa, this would have to be chemical energy. On Earth, life near the bottom of the ocean, far removed from any sunlight above, gets its energy from a process called **chemosynthesis**. Plants on Earth take in energy from the Sun to create their food — this is photosynthesis. Chemosynthesis works similarly to photosynthesis, except that the energy comes not from the Sun but from chemical reactions. The environment at the bottom of Europa's ocean may be similar to hydrothermal vent communities on Earth's ocean floor, where life may have gotten its first start.

One of Europa's key attributes from an astrobiology perspective is that its ocean is most likely in direct contact with an underlying rocky mantle, and the interaction of water and rocks—especially at high temperatures, as within Earth's hydrothermal vent systems—yields a *reducing chemistry* (where molecules tend to give up electrons readily) that is like one half of a chemical battery. To complete the battery and provide energy that could be used by life requires that an *oxidizing chemistry* (where molecules tend to accept electrons readily) also be available. On Earth, when chemically reducing vent fluids (for example, H₂S or methane) meet oxygen-containing seawater, the energy that becomes available often supports thriving communities of microorganisms and animals on the sea floor, far from the light of the Sun.

The Galileo mission found that Europa's icy surface does contain an abundance of oxidizing chemicals. This means that availability of energy to support life depends very much on whether the chemistry of the surface and the ocean can mix, despite the kilometers of ice in between. That Europa's ice crust appears geologically "young" (only tens of millions of years old, on average) based on the low number of craters and that it is active makes it tantalizing to think that such mixing might indeed occur. Understanding whether and how much exchange occurs between the surface and ocean of Europa will be a key science objective of future missions to Europa, and a major step forward in understanding whether this moon could be a cradle of life.

Future Exploration

Europa is of such great interest to astrobiologists that a NASA mission, called the [Europa Clipper](#), is set to launch in October of 2024. While the Europa Clipper will not land on the surface of Europa and do any drilling (this is still a challenging technological feat, as it would need to drill beneath about 20 km of ice), it will conduct extensive experiments while in close orbit of Europa.



Figure 5 – Jupiter's Moon Europa, as Imaged by NASA's Galileo Mission. The relative scarcity of craters on Europa suggests a surface that is “geologically young,” and the network of colored ridges and cracks suggests constant activity and motion. Galileo's instruments also strongly suggested the presence of a massive ocean of salty liquid water beneath the icy crust.

Saturn's Potentially Habitable Moons

Enceladus

Another tantalizing ocean world is the small (500 km diameter) moon of Saturn, Enceladus. The story of how Enceladus was found to be a world with an icy surface, with a salty water ocean beneath, follows similar lines as the evidence for Europa. Spectroscopy of Enceladus revealed the presence of ice, and Voyager 1 and 2 took the first up-close pictures of its surface. The Voyager photos showed some heavily cratered regions but a very smooth region near Enceladus' south pole. The Cassini mission began studying Saturn in 2004 and visited Enceladus multiple times before it ended in 2017.

Enceladus



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=138>

Video Credit: @NASA

In 2005, the Cassini mission performed a close flyby of Enceladus, and made the remarkable discovery that plumes of gas and icy material were venting from the moon's south polar region. The Cassini spacecraft eventually flew directly through the plumes in 2015 and detected the presence of water, salts, organic compounds, silicates and hydrogen gas (H_2). All of this chemistry is suggestive of interactions between liquid water and rocky material beneath Enceladus' icy surface. As with Europa, this could point to hydrothermal vents within Enceladus' subsurface ocean, which is an environment with all of the ingredients for life to form. The plumes are associated with the blue features called "tiger stripes" in the southern part of Enceladus (see the left side of Figure 6). The tiger stripes are fissures on the surface that are a pathway between the surface and the liquid water beneath. The plumes represent a type of geological activity called **cryovolcanism**. The prefix *cryo* means cold and this type of volcanism involves icy materials being spewed out, rather than hot lava.

Concept Check

Enceladus is a small moon, only about one-seventh the size of Earth's Moon, yet it is geologically active. What causes the active volcanic activity on Enceladus? In other words, what is the source of energy?

Strong evidence for a global liquid water ocean beneath the surface, rather than a body of water only below the south polar region, came from Cassini measurements of a tiny wobble (called **libration**) of Enceladus as it rotates on its axis and revolves around Saturn. The size of the libration can only be explained by a global ocean; if the libration was smaller, this would mean that the subsurface layer was at least partially solid rather than liquid.

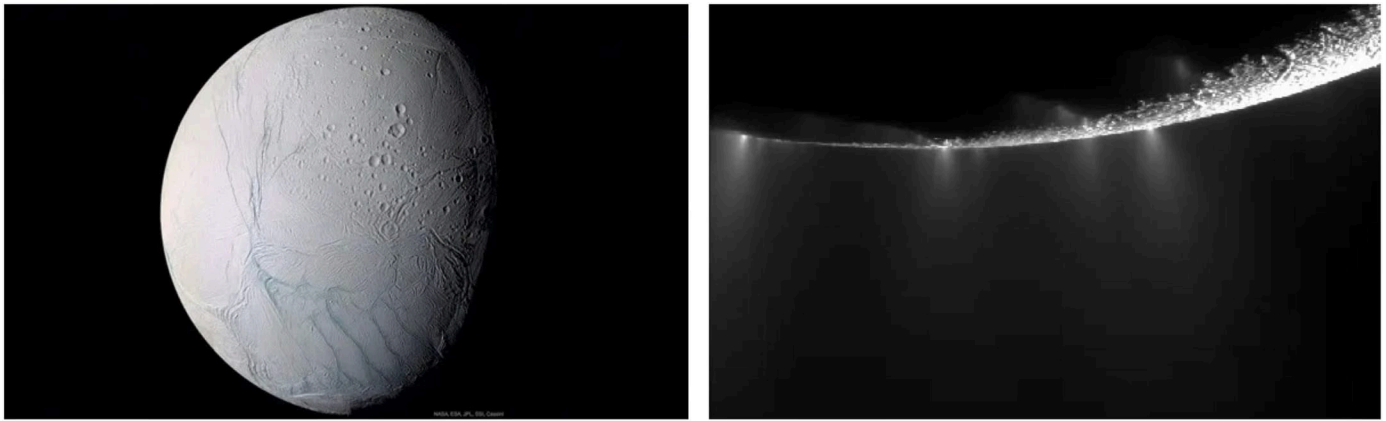


Figure 6 – Saturn’s moon Enceladus. Both images were taken with the Cassini spacecraft. On the left, the giant rifts (called “tiger stripes”) on the surface of Enceladus are seen near its south polar region. On the right is a zoom into the south polar region, showing geysers and plumes of ice and gas.

Titan

Saturn’s largest moon Titan was discovered in 1655 by the Dutch astronomer Christiaan Huygens. In the 1940s, Gerard Kuiper used spectroscopy and found that Titan has an atmosphere containing methane. Images of Titan taken by the Voyagers confirmed the atmosphere but were unable to see through the clouds to the surface. The Cassini mission, which was so valuable for learning about Enceladus, also did a (literal) deep dive into Titan: Cassini’s Huygens probe descended through Titan’s atmosphere and landed on its surface in 2005.

Saturn’s moon Titan



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=138>

Video Credit: @NASA

Titan’s Atmosphere and Surface

Titan is very different from both Enceladus and Europa. Although it may host a liquid water layer deep within its interior, it is the surface of Titan and its unusual chemistry that makes this moon such an interesting place. Titan’s thick atmosphere—the only

one among moons in the solar system—is composed mostly of nitrogen (N_2) but also of about 5% methane. Voyager detected hydrocarbons in Titan's atmosphere, indicating an active chemistry in which sunlight interacts with the atmospheric nitrogen and methane to create a rich mix of organic compounds. In the upper atmosphere, the Sun's ultraviolet light breaks apart the methane and nitrogen molecules and recombines them into more complex organic compounds that are collectively known as "tholins" (Carl Sagan coined this term). The tholins shroud Titan in an orange haze, as seen in the right panel of Figure 7.

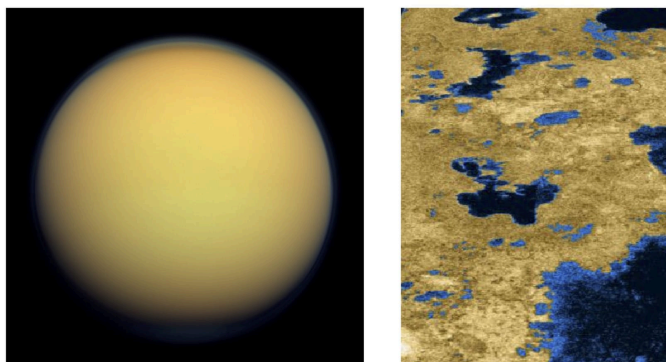


Figure 7 – Views of Titan from the Cassini mission. On the left, Titan and its hazy orange atmosphere. On the right, a radar image from Cassini showing liquid lakes on Titan's surface.

In the atmosphere, the conditions are right for some of the hydrocarbons to condense and fall as rain to Titan's surface. This methane or ethane rain creates lakes and rivers on Titan's surface. When combined with an atmosphere and a cycle analogous to Earth's water cycle, this sounds like a habitable environment! We do need to keep in mind that the lakes are made of liquid methane and ethane and not liquid water; it is far too cold for liquid water to exist on Titan's surface. On Earth, we follow the liquid water and find life. On Titan, perhaps we will follow the liquid methane or ethane and find life with a different biochemical basis. A radar image of Titan's lakes, taken by Cassini in 2006, is shown in the right in Figure 7.

The Huygens Probe

In January 2005, the Huygens probe descended to the surface of Titan and relayed data, including imagery of the landing site, for about 90 minutes. You can watch a video about the descent of Huygens to Titan's surface below (the footage is real and not CGI!):

Huygen's probe



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=138>

Future Exploration

NASA has selected a new mission to Titan for launch in 2027. Called *Dragonfly*, this mission is a drone that will fly in Titan's atmosphere, with emphasis on study of prebiotic chemistry (see details at <https://dragonfly.jhuapl.edu/>). Other future proposed missions include a balloon operating in the atmosphere and a “boat” floating in one of the Titan lakes.

Life Under the Ice on Earth

What can we expect to find living in an ocean underneath a thick, icy shell? We can gain some insight by seeing what has been found by exploring subglacial lakes — meaning lakes that are buried beneath glacial ice — on Earth. Subglacial lakes exist in Antarctica and many international research stations have been established there, including the Russian Vostok station and the U.S.'s McMurdo Station. Two subglacial lakes in Antarctica, Lake Vostok and Lake Whillans, have been particularly well-studied for the possibility of any life within them. Both are freshwater lakes, with Lake Whillans about 800 m (or 0.5 miles) beneath the ice and Lake Vostok at a deeper 4 km (2.5 miles) below the ice. While these are not an exact analogs to Europa or Enceladus' oceans, it is still informative to assess the habitability of these types of environments as well as trying out new techniques to drill beneath thick ice. Contamination is also a very big challenge to overcome when drilling through ice.

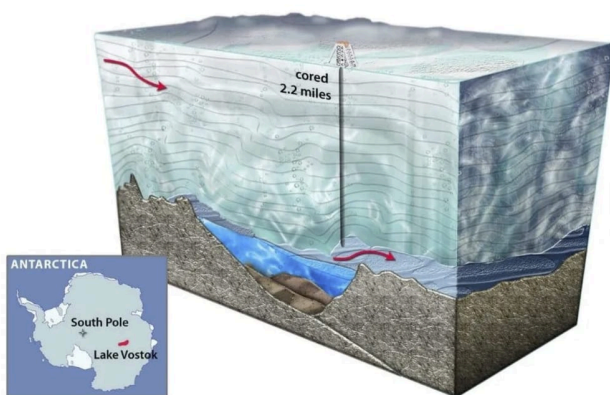


Figure 8 – An artist's sketch of a cross section of Antarctic ice, showing Lake Vostok underneath layers of ice.

Lake Vostok was confirmed to exist via radar mapping in the 1990s and is about the size of Lake Ontario. Scientists were able to drill close to the lake's surface in 2013. Microbes, some previously unknown, were found living inside the ice. While the layer of ice studied is very close to the lake's surface, this is still not direct evidence of life in Lake Vostok.

Lake Whillans was also drilled into in 2013. Samples of sediments and water were collected and analyzed.

Key Concepts and Summary

The conditions for the chemistry of life might exist on moons as well as planets. The larger moons orbiting Jupiter and Saturn are intriguing because tidal heating could provide an extra source of energy. There is evidence for an ocean of water under the ice mantle surface of Jupiter's moon, Europa. Since life as we know it requires water, Europa is a key target in our search for life in the solar system and NASA is launching an exploratory mission to Europa. Saturn is twice the distance from the Sun as Jupiter, but plumes of water have been observed coming from the surface of Enceladus (one of the moons orbiting Saturn). Titan is Saturn's largest moon and the Cassini mission has returned images of a thick atmosphere and pools of liquid methane and ethane on the surface, making Titan a key location for the study of biochemistry in cold environments that makes use of solvents other than water.

Review Questions

Summary Questions

1. Explain the process of tidal heating and how it provides energy to some of the jovian moons.

Exercises

1. **Europa's Ocean.** Describe the three pieces of evidence for a salty liquid water ocean beneath Europa's icy surface.
2. Why is it easier to study what is inside Enceladus' ocean than what is in Europa (or Ganymede or Callisto's) ocean?
3. Research the moons Ganymede, Callisto and Triton (a moon of Neptune) and discuss the evidence for liquid water and organic molecules on these worlds. Based on what you found, do you think these worlds are habitable? Explain why or why not.

5: THE SEARCH FOR LIFE BEYOND OUR SOLAR SYSTEM

5.1: The Habitable Zone

5.2: Exoplanets- Early History and Direct Imaging and Doppler Detection Methods

5.3: Transit, Astrometry and Gravitational Microlensing Techniques

5.4: Exoplanets- Statistics and Modern Discoveries

This page titled [5: The Search for Life Beyond Our Solar System](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

5.1: The Habitable Zone

In assessing the possibility that life could exist on another world, one of the first questions to ask is: “Is it in the habitable zone?”. The habitable zone around a star is the region in which the temperature on the surface of the planet would be in the proper range to allow liquid water to exist. If a planet does fall into the habitable zone around its host star, we cannot immediately conclude that life exists but it is a useful starting point in looking for other worlds that contain life.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss the concept of the habitable zone around a star
- Describe the habitable zone around the Sun
- Explain the difference between the equilibrium and surface temperature of a planet
- Compare the surfaces and atmospheres of Earth, Venus and Mars
- Explain the concept of a continuous habitable zone

Defining the Habitable Zone Around a Star

One of the mantras in Astrobiology is “Follow the water!”. This makes sense from our Earthly perspective, since absolutely all life on Earth requires liquid water. On Earth, where there is water, there is life. Of course, other liquids could serve this same purpose in other environments, but we stay focused on liquid water for defining the habitable zone around a star.

The habitable zone around a star is the region where the temperature is appropriate for liquid water to exist on the surface of a world. Sometimes this region is called the “Goldilocks Zone,” an analogy to the fairy tale Goldilocks and The Three Bears and the “just right” temperature for porridge (not too hot or too cold – Figure 1).

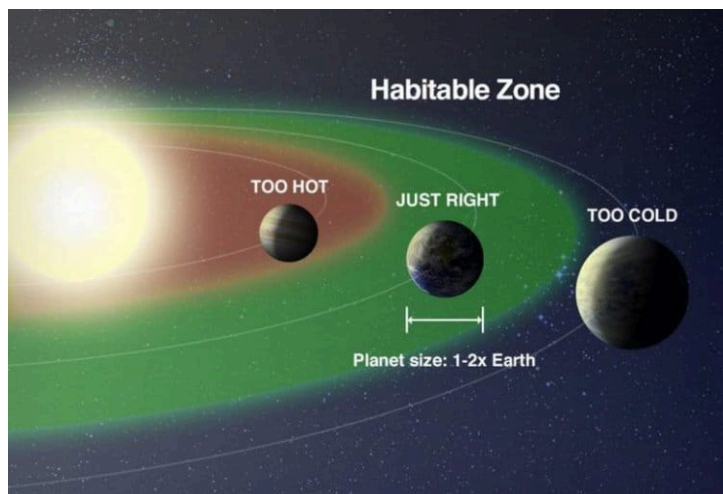


Figure 1 – The habitable zone is sometimes referred to as the “Goldilocks” zone, as it is the region where the temperature is not too hot and not too cold for liquid water to exist.

Strictly speaking, this means that a planet in the habitable zone has a temperature in the range between the freezing point and the melting point of water: 273-373 K (which is 0-100°C, or 32-212°F). While this may seem simplistic, as the presence of liquid water depends not only on temperature but on the surface pressure of a world and whether there is an atmosphere, it is a good way to begin the process of identifying potentially habitable exoplanets around another star. Other factors can then be considered to rule out the world as being habitable. For example, the Moon is in the habitable zone around the Sun but it is not a habitable world; the Moon is much smaller than the Earth and has lost all of its internal heat. While we tend to mainly think of rocky worlds as places that can harbor life — as life on Earth likely began at the liquid water-land boundary — many exoplanets that are gas giants fall into the habitable zone around their star. The first exoplanet detected around a Sun-like star, 51 Pegasi b, is actually a

gas giant.

Let's start by looking at the habitable zone in our own solar system, as shown in Figure 2. Currently, the Sun is about halfway through its life, with an age of about 4.6 billion years. Over time, the Sun will get hotter and the habitable zone around it will shift further away from the Sun — we discuss this point further in the Continuous Habitable Zone section below. Today, the Earth and Mars fall into the Sun's habitable zone; Venus is too close to the Sun for liquid water oceans to exist on its surface — any previous liquid water has completely evaporated. The Earth is at a distance of 1 AU from the Sun. What are the inner and outer boundaries of the Sun's habitable zone?

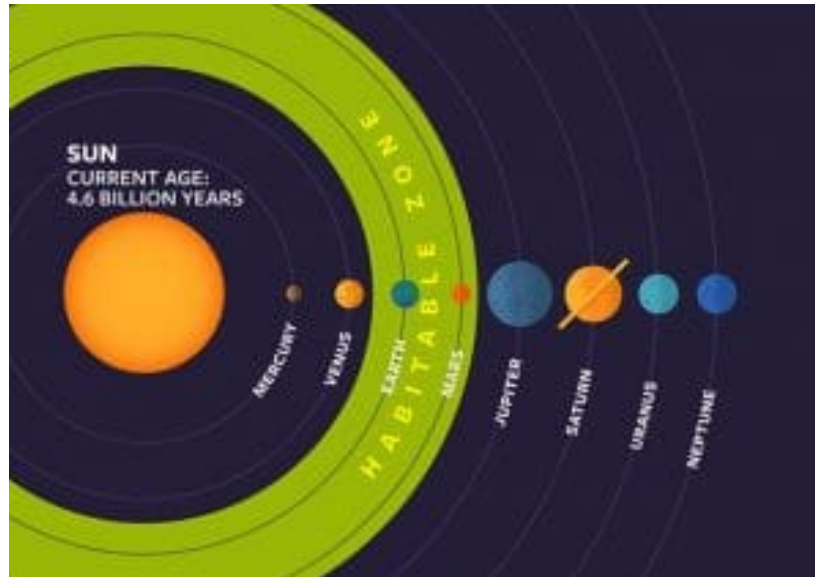


Figure 2 – The Habitable Zone around the Sun is a distance where planets might have the right conditions for liquid water to exist on their surfaces. **Which solar system planets are currently in the Sun's habitable zone?**

Two ranges for the habitable zone are often quoted: one is a conservative estimate, with a more narrow range of distances, and the second is the optimistic estimate, which is wider. The conservative estimate accounts for changes due to water in a planet's upper atmosphere that can lead to a "moist" greenhouse effect, which sets in earlier than the "runaway" greenhouse effect. For the Sun, the conservative range is 0.95 – 1.37 AU, while the optimistic range extends the borders to 0.85 – 1.7 AU. Figure 2 shows the optimistic habitable zone around the Sun; the conservative estimate would not include Mars, which is 1.5 AU from the Sun.

The box below shows an example of how to estimate the boundaries for the habitable zone around any star.

Worked Example: The Boundaries of the Habitable Zone Around a Star

To approximate the inner boundary of the habitable zone around a star, we can calculate the distance from the star where the temperature is equal to the boiling point of water. The outer boundary of the habitable zone is calculated as the distance from the star where the temperature would equal the freezing point of water.

Recall from the Stefan-Boltzmann law that the luminosity (total power output) of a star is:

$$L = \sigma T^4 4\pi d^2 \quad (5.1.1)$$

where d is the distance from the star. By taking the ratio of these values and using the conservative inner and outer habitable zone boundaries for our solar system ($d_{\text{inner,Earth}} = 0.95$ AU and $d_{\text{outer,Earth}} = 1.37$ AU), we obtain the following relationships:

$$d_{\text{inner,star}} = 0.95 \sqrt{L_{\text{star}}/L_{\text{Sun}}} \quad (5.1.2)$$

$$d_{\text{outer,star}} = 1.37 \sqrt{L_{\text{star}}/L_{\text{Sun}}} \quad (5.1.3)$$

For example, let's calculate the habitable zone boundaries around the star TRAPPIST-1. A quick search shows that the luminosity of TRAPPIST-1 is less than just one-thousandth of the luminosity of the Sun: $L_{\text{T-1}} = 0.00052 L_{\text{Sun}}$. In one step, we can put this

luminosity into the relationships above and find that the inner and outer boundaries for the habitable zone around TRAPPIST-1 are $d_{\text{inner}} = 0.022 \text{ AU}$ and $d_{\text{outer}} = 0.032 \text{ AU}$. In our solar system, Mercury is 0.39 AU , so the habitable zone for TRAPPIST-1 is *extremely* close to the star compared with our habitable zone.

Questions

1. Calculate the inner and outer boundaries of the habitable zone around the star Pegasi 51 (this is the star that 51 Pegasi b orbits).
2. Calculate the optimistic habitable zone boundaries around Pegasi 51.

Show Answers

1. $d_{\text{inner}} = 1.08 \text{ AU}$, $d_{\text{outer}} = 1.56 \text{ AU}$
2. $0.97 - 1.94 \text{ AU}$. Note that a luminosity of $1.3 L_{\text{Sun}}$ was used for these calculations.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Comparing Venus, Earth & Mars

Although they are very different places today, the planets Venus, Earth and Mars all started out with very similar conditions. They all formed at the same time from materials in the solar nebula, meaning that they all started off with the same compositions mixed together in a hot liquid. As these molten planets cooled, their interiors differentiated and their early atmospheres and oceans formed. The materials in the atmosphere came primarily from volcanic outgassing with some contributions possibly from impacts from comets (all three planets experienced many impacts during the Late Heavy Bombardment from 4.1 – 3.8 Gyr ago). Some materials that were outgassed include a mix of CHONS: sulfur dioxide (SO_2), hydrogen sulfide (H_2S), carbon dioxide (CO_2), and ammonia (NH_3). Note that there was no free O_2 in these early atmospheres, as oxygen will quickly bond with other elements once it enters the atmosphere. Oxygen in Earth's atmosphere is a byproduct of life.

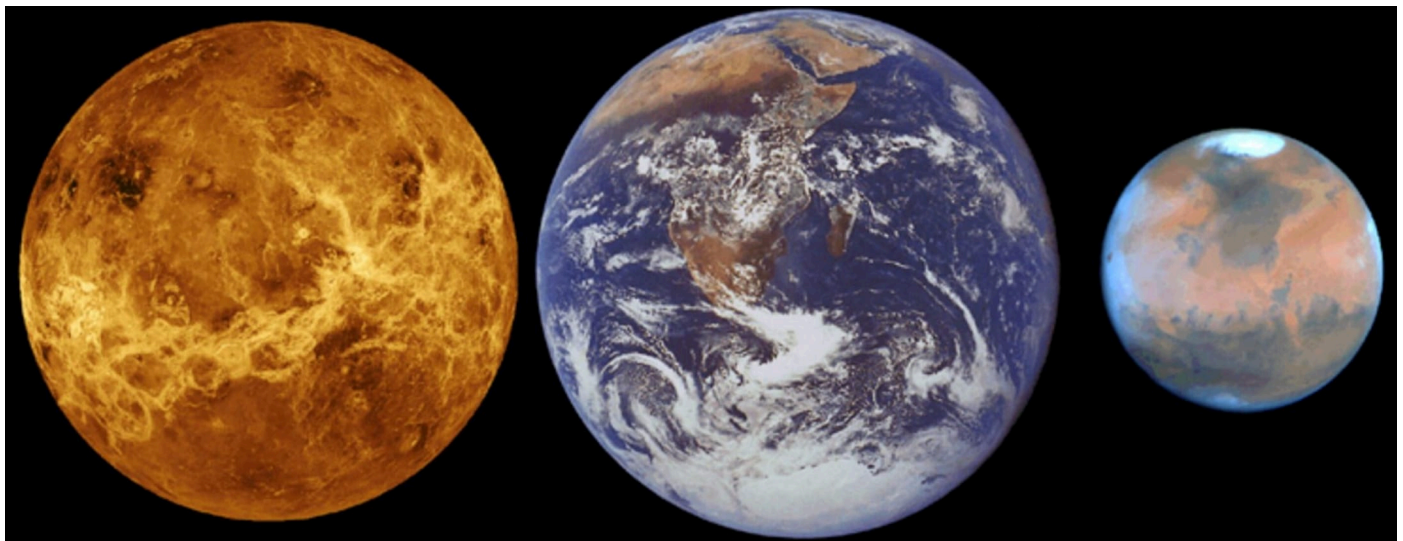


Figure 3 – Venus, Earth, and Mars. The planets are shown to scale, with Mars having about half the radius of Earth.

Today, the Earth's atmosphere is dominated by molecular nitrogen (N_2) and oxygen while both Mars and Venus have atmospheres almost entirely dominated by CO_2 . These differences can be explained by considering both the size and distance from the Sun of Venus and Mars.

Table 1: Comparison of Venus, Earth, and Mars. Values in the first five rows are relative to the Earth's value.

	Venus	Earth	Mars
Mass	0.82	1	0.11

Radius	0.95	1	0.53
Distance from the Sun (AU)	0.72	1.00	1.52
Surface gravity	0.91	1	0.38
Atmospheric pressure	90	1	0.007
Main atmospheric gas(es)	CO ₂	N ₂ , O ₂	CO ₂

Venus and the Earth are very close in size while Mars is significantly smaller. Table 1 summarizes the data. In terms of radius, Venus has nearly the same radius (95% Earth's size) while Mars has about half — 53% — the radius of Earth. This means that Mars likely has already lost most or all of its internal heat while Venus may still retain internal heat to this day that can drive geological activity. The surface gravity of Venus is very close to the Earth's (at 91%) while Mars has only 38% the surface gravity of Earth. This means that it is much easier for any atmosphere Mars had in the past to be lost to space, and this is consistent with the extremely thin atmosphere that Mars holds onto today. We devote an entire chapter to exploring the possibility of water (and life) on Mars.

It is quite possible that for the first billion years after it formed, Venus had a moderate temperature atmosphere, surface water, and perhaps even life. However, because Venus is closer to the Sun, it intercepts more energy than the Earth. Liquid water would have gradually evaporated, forming a potent greenhouse gas that trapped more solar energy, leading to a positive feedback loop that further warmed the planet, driving water out of the crust of Venus.

Ultraviolet radiation dissociates water molecules into its component atoms of hydrogen and oxygen. The Venus Express spacecraft, launched in 2005 by the European Space Agency (ESA), has measured the escape of hydrogen and oxygen (roughly in a ratio of 2:1) from the upper atmosphere of Venus. Over the 4.56 billion year lifetime of the planet, Venus has lost its water. Carbon dioxide is chemically bound in the crust of Earth rocks. However, in the presence of a hot, dry Venusian climate, carbon dioxide would evaporate out of the mantle of the planet, producing the heavy CO₂ atmosphere observed today. The current condition of Venus is the fate of an Earth-like planet that is too close to its star to retain liquid surface water.

Equilibrium Temperature

Considering that the amount of power from the Sun received by a planet should be equal to the amount of power emitted by the planet, we can calculate the equilibrium temperature of a planet (see Box 2 below for details). The equilibrium temperature of Venus and Earth are $T_{\text{eq,Venus}} = 260 \text{ K}$ and $T_{\text{eq,Earth}} = 255 \text{ K}$. How does this compare with the actual measured values at the surface of Venus and Earth? The actual temperature at the surface of Venus is 730 K (457°C)! Venus experienced a runaway greenhouse effect, with all of its CO₂ now baked out of rocks and in the atmosphere.

The Earth's actual surface temperature is 288 K (15°C), putting it right into the habitable zone range for liquid water. The Greenhouse effect is often portrayed in a negative way, as the increase of greenhouse gases in the Earth's atmosphere causes the global temperature to rise and exacerbates climate change— and this is a real concern if Earth is to avoid the fate of Venus prematurely. However, keep in mind that without a Greenhouse effect the Earth would be an ice ball!

Worked Example: Calculating the Equilibrium Temperature of a Planet

To find the equilibrium temperature of a planet, we can consider the power that reaches the planet from its star and is absorbed by the planet, and the power that is emitted by the planet. This planetary energy budget is in balance so we can set them equal to each other: $P_{\text{in}} = P_{\text{out}}$

The power reaching the surface of the planet is the product of the amount of energy reaching the planet, the cross-section of the planet that receives this energy, and the fraction of this light that is absorbed by the surface. This can be written:

$$P_{\text{in}} = (L_{\text{star}}/4\pi d_p^2) \times (\pi R_p^2) \times (1 - a) \quad (5.1.4)$$

where L_{star} is the luminosity of the star, d_p is the distance from the star to the planet, R_p is the radius of the planet, and a is a variable called the **planetary albedo**. The planetary albedo is a measure of how much light is reflected back by the planet, meaning it doesn't reach the surface. This can be due to ice on the surface or in clouds. In our solar system, $a_{\text{Earth}} = 0.32$ and

$a_{Venus} = 0.75$; Mercury has the lowest albedo at 0.12. The luminosity of the star can be written using the Stefan-Boltzmann law as $L_{star} = \sigma T_{star}^4 4\pi R_{star}^2$.

The power emitted by the planet is the produce of the energy flux radiated by the planet (in W/m^2) multiplied by the total surface area of the planet:

$$P_{out} = (\sigma T_p^4) \times (4\pi R_p^2) \quad (5.1.5)$$

where T_p is the temperature of the planet.

Setting $P_{in} = P_{out}$ and solving for T_p , we obtain the following equation for the planetary temperature:

$$T_p = T_{star} \sqrt{R_{star}/(2d_p)} (1-a)^{(1/4)} \quad (5.1.6)$$

We can rescale this relationship so that the input units are AU for d_p and we use the Stefan-Boltzmann law to rewrite T_{star} and R_{star} in terms of L_{star} :

$$T_p = 278.1 (L_{star}/L_{Sun}) \sqrt{1/d_p} (1-a)^{(1/4)} \quad (5.1.7)$$

Example 1: Equilibrium temperature of Venus

For Venus, we just need the distance from the Sun in AU and Venus' planetary albedo: $d_V = 0.72$ AU and $a = 0.75$. For the luminosity of the star, since it is the Sun, $L_{star} = L_{Sun}$

$$T_{Venus} = 278.1 (L_{Sun}/L_{Sun}) \sqrt{1/0.72} (1-0.75)^{(1/4)} \quad (5.1.8)$$

This is much cooler than the actual surface temperature of Venus, which has undergone a runaway greenhouse effect. To calculate the adjusted temperature that factors in the greenhouse effect, an extra term ge must be factored into the equation, which appears as $(ge)^{-1/4}$. This term varies from 0 to 1, where 1 means no greenhouse effect. Try recalculating the temperature of Venus with different values for the ge term. Which gives a good match to Venus' actual surface temperature of 730 K? [Ans: about 0.15]

Question: Calculate the equilibrium temperature of Mars.

Show Answer

The equilibrium temperature of Mars can be found in the exact same way as for Venus: now we need the average distance of Mars from the Sun and its planetary albedo. Looking these up, we find: $d_M = 1.52$ AU and has an albedo of $a = 0.16$ (reference: <https://astronomy.swin.edu.au/cosmos/a/Albedo>). For the luminosity of the star, since it is the Sun, $L_{star} = L_{Sun}$

$$T_{Mars} = 278.1 (L_{Sun}/L_{Sun}) \sqrt{1/1.52} (1-0.16)^{(1/4)} \quad (5.1.9)$$

Exoplanetary Habitable Zones

The discussion of the habitable so far has focused on the Sun. However, we can intuitively imagine how the habitable zone would differ for stars that are hotter and cooler than the Sun.

Thought Question: How do you expect the habitable zone to change for a star that is hotter than the Sun? Will it be closer, further, or the same distance away from the star as the Sun's habitable zone? Will it be narrower or wider, or the same width as the Sun's habitable zone? Now answer the same question for a star that is cooler than the Sun.

Show Answer

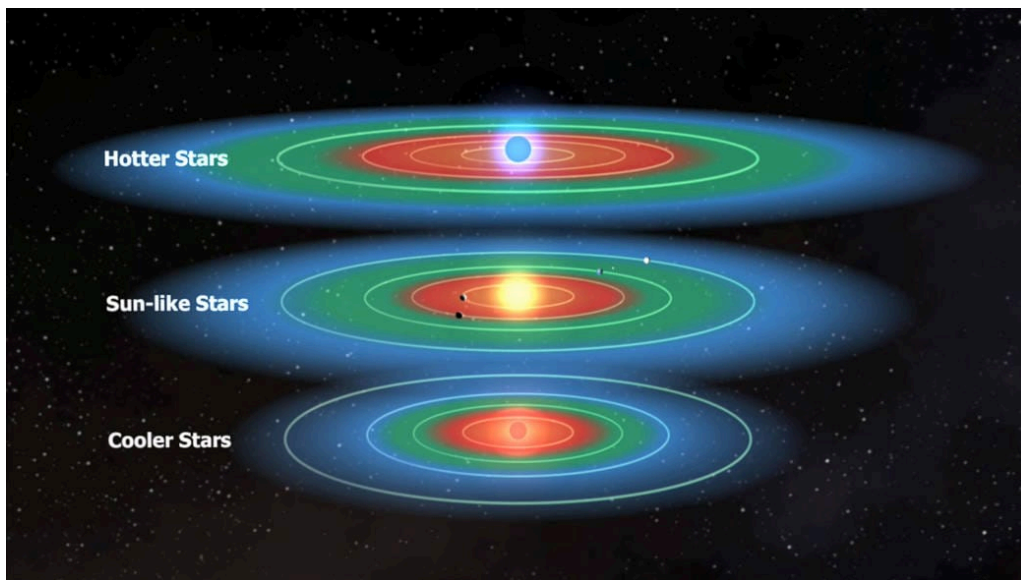


Figure 4 – The habitable zone around a hot, average (Sun-like) and cool star. **What types of stars have habitable zones that are wider and extend to greater distances than the HZ for the Sun?**

As seen in Figure 4, a star hotter than the Sun (such as those with spectral types OBAF) has a wider habitable zone that is further away from the Sun. This makes sense, as a hotter star on the Main Sequence has a higher luminosity and thus more energy reaches the planet than from a cooler star. Because there is more of it, this energy can spread further before the temperature dips below the freezing point of water, so the habitable zone is wider. The same logic can be applied to the cooler star, which has a narrower habitable zone that is closer to the star.

Let's look at the habitable zone around the star TRAPPIST-1, which we considered above. You can see a nice animation of the habitable zone for TRAPPIST-1 [here](#). The inner and outer boundaries of TRAPPIST-1's habitable zone were estimated to be from 0.022-0.032 AU (see Box 1 above). The closeness of TRAPPIST-1's habitable zone to the star is exemplified by seeing that the TRAPPIST-1 system needs to be zoomed in 25-times for it to be compared on the same image as our solar system.

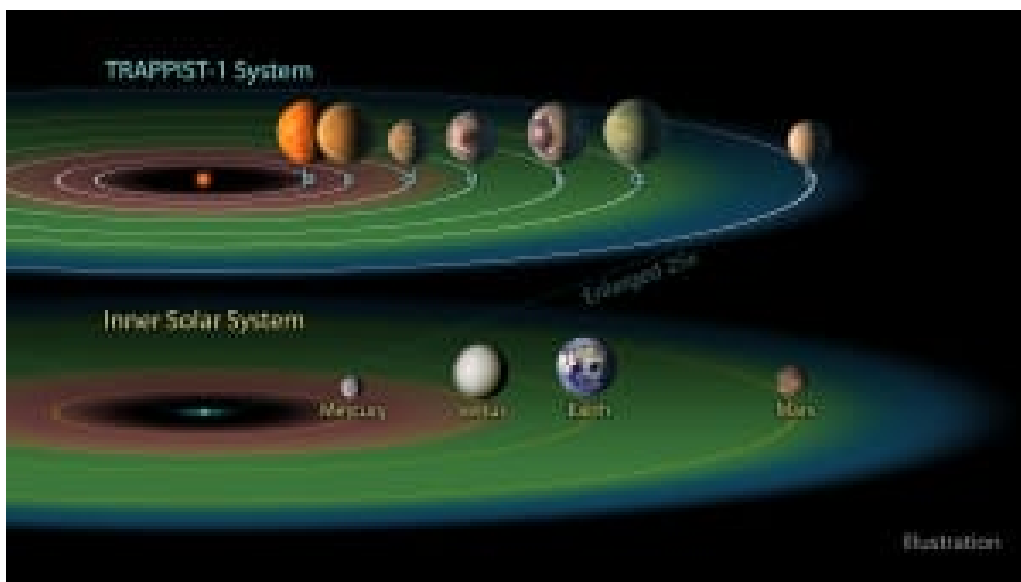


Figure 5 – The habitable zone around the star TRAPPIST-1. Note that we have magnified the TRAPPIST-1 system, which would entirely fit inside the orbit of Mercury in our solar system.

Another way to visualize the habitable zone around stars of different types (temperatures) is to calculate how much light is received at a planet. Figure 6 below shows how the intercepted starlight varies with the temperature of the host star. The brightness percentages on the x-axis are shown relative to the Earth (by definition, the Earth is 100%, so a planet with a percentage greater than 100% receives more light from its host star than the Earth does). Some known exoplanets that are around

the same size as the Earth are plotted for comparison. The boundaries for the habitable zones are shown; why does the width of the HZ increase as the temperature of the star increases?

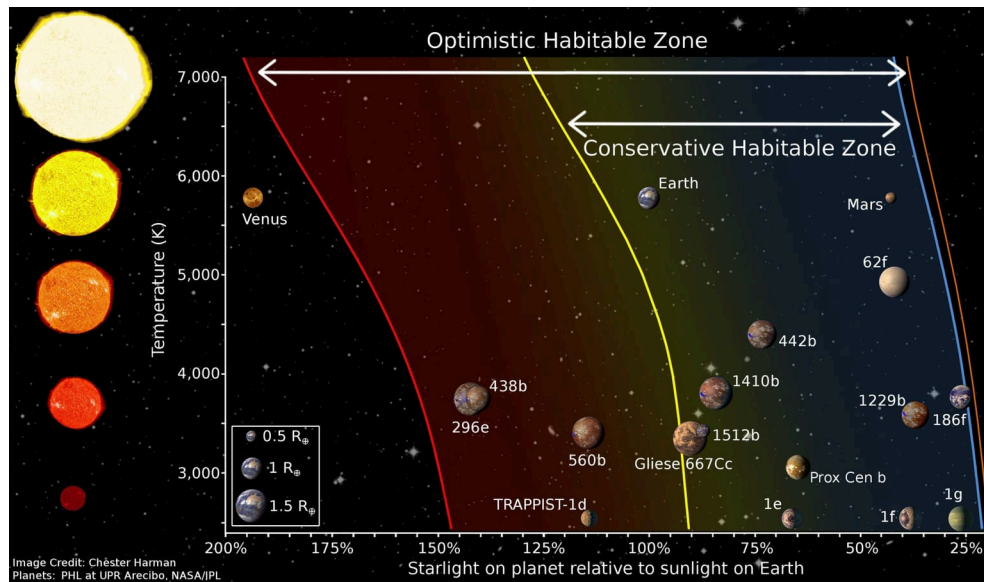


Figure 6 – The amount of intercepted starlight is shown relative to that of Earth for different mass (temperature) stars. The horizontal x-axis could also be shown as distance from the star. Earth-sized exoplanets are shown for comparison.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The Continuous Habitable Zone

The Sun is getting more luminous as it ages. This means that over time the habitable zone around the Sun will be pushed further away, and eventually Earth will be in the “too hot” zone. This is true of all stars on the Main Sequence. A star’s brightness increases by about 0.7% every 100 million years. Since forming 4.5 billion years ago, the Sun is now 30% brighter today than it was when it started shining continuously. In about 2 billion years from now, the Sun’s optimistic habitable zone’s inner edge will be out past 1 AU, so Earth will no longer fall into this zone. Figure 7 shows how far the Sun’s habitable zone will extend once the Sun becomes a red giant star, in about 8 billion years from now. Notice that Mercury and Venus have been swallowed up by the Sun at this point.

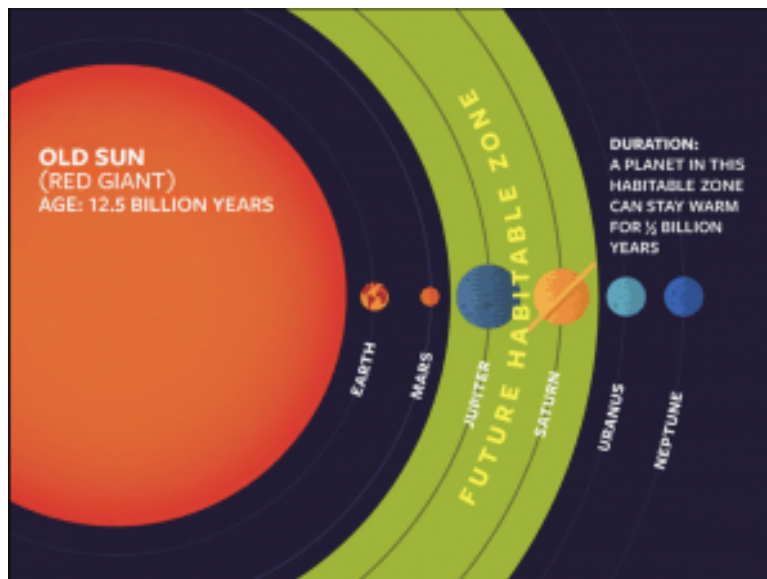


Figure 7 – The Sun’s HZ in 8 billion years. The HZ will no longer contain Earth and the Sun will completely engulf Mercury and Venus.

Key Concepts and Summary

The Habitable Zone is the region around a star where it is “not too hot” and “not too cold” for liquid water to exist. The boundaries of the habitable zone are set by the luminosity of the host star. For billions of years, this luminosity will be constant as hydrogen is being fused into helium in the core of the star. However, after the star burns all of the hydrogen in its core, it will evolve and begin burning an outer layer of helium. When this occurs, the luminosity of the star increases and the boundaries of the habitable zone move out. In addition to the luminosity of the star, the atmosphere of a planet can trap radiation, further warming the planet. In our solar system, the Earth, our Moon and Mars reside in the habitable zone – since we have not found life on the Moon or Mars, this shows that the concept of a habitable zone is an idealization, perhaps necessary but not really sufficient to infer the existence of life as we know it.

Review Questions

Summary Questions

1. What is the habitable zone around a star?
2. What property (or properties) of a star determine the location of its habitable zone?
3. What are the optimistic and conservative boundaries for the habitable zone around the Sun?
4. Which planets are in the Sun's habitable zone?
5. What is the difference between the equilibrium temperature of a planet and its actual (measured) surface temperature? What causes these two temperatures to vary, especially in the case of Venus?
6. How do the atmospheres of Earth, Venus and Mars vary? What led to these differences?
7. How does the habitable zone around a star change as the star evolves?

Exercises

1. Generate the habitable zone around another star. Choose an exoplanetary system that you think could have some habitable planets. [Try <http://astro.twam.info/hz/> or <https://ccnmtl.github.io/astro-simulations/circumstellar-habitable-zone-simulator/>]
2. Greenhouse effect PHeT sim: <https://phet.colorado.edu/en/simulations/greenhouse-effect>

5.2: Exoplanets- Early History and Direct Imaging and Doppler Detection Methods

Perhaps nothing captures the imagination more fully than envisioning life on another planet. What was still in the realm of science fiction just a few decades ago, especially through television shows like *Star Trek*, is now commonplace: as of 2024, more than 5,000 exoplanets — planets orbiting a star other than our Sun — have been discovered. Their detection and study has brought the search for life in the universe to a new level. Exoplanets can be categorized as rocky or gaseous, and their atmospheres can be searched for biosignatures. Unlocking more clues from exoplanets will narrow down the possibilities for habitable worlds in our galaxy.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe the challenges and progress in taking direct pictures of exoplanets
- Explain the center-of-mass of a solar system and star “wobbling” that enables detection
- Describe how spectra can be used to detect exoplanets (the Doppler Technique) and how the minimum mass of an exoplanet can be found
- Summarize the types of planets (sizes, masses, orbital periods) that can be detected with imaging and radial velocity (“Doppler”) data.

Early History

There are several techniques for discovering exoplanets. The first exoplanets were detected by the gravitational effect that they exerted on their host stars. More recent techniques have been able to image the planets. Each method is sensitive to specific types of exoplanets, and when we piece the information together, we can begin to understand the diversity of exoplanets. Humans have long wondered whether other solar systems with planets like our own Earth might exist among the billions of stars in our galaxy, and this moment will go down as the time when we figured this out.

The first failures and successes

The discovery of worlds around other stars has a long history with many false starts. In the 1960s, Peter van de Kamp interpreted a small wobble in the position of Barnard’s star as an exoplanet. Observations by other astronomers contradicted that result, although van de Kamp never admitted that his claim was in error. In 1991, Lyne and Bailes reported in the prestigious *Nature* journal the discovery of a planet orbiting the pulsar star PSR 1829-10. They had measured frequency of pulse arrival times and used the Doppler effect to infer the presence of a planet, but later realized that they had not properly accounted for the velocity of the Earth around the Sun. When Lyne retracted the result at a meeting of the American Astronomical Society in January 1992, he received a standing ovation for his scientific integrity and courage.

There were also a few signals that were initially published with alternative interpretations that later turned out to be exoplanets. In 1988, Campbell, Walker and Yang observed a periodic radial velocity signal in the red giant star, Gamma Cephei. They tentatively interpreted this as photospheric variability in the star, but additional data by Hatzes and colleagues in 2002 confirmed that this was indeed an orbiting exoplanet. Another example occurred in 1989 when Latham and colleagues published the discovery of a companion to the star HD 114762; the team cautiously interpreted this as a possible brown dwarf. However, by 2012, this object was reclassified as a massive exoplanet.

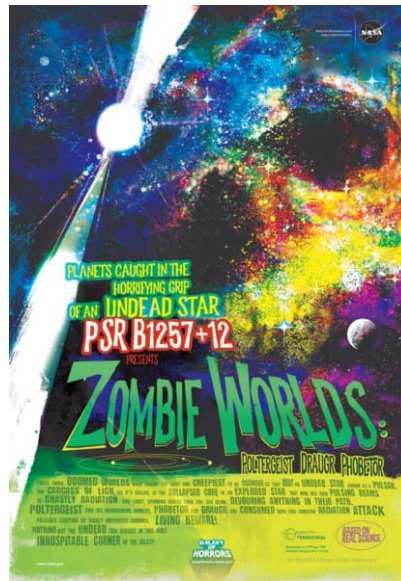


Figure 1 – Pulsar Planets. NASA Exoplanets Exploration Program's Exoplanet Travel Bureau poster for the planets around the pulsar PSR B1257+12. One of these exoplanets, PSR B1257+12 b, has been named Poltergeist, a nod to its unlikely existence after the supernova explosion of its host star

The first confirmed exoplanets were very peculiar. Aleksander Wolszczan and Dale Frail measured periodic variations in the frequency of pulse arrival times to detect two small planets orbiting the pulsar neutron star, PSR 1257+12 in 1992. In 1994, they found one more planet in this system. These discoveries were puzzling because this planetary system should not have survived the supernova explosion of the host star before it evolved to become a spinning neutron star. The planets likely formed in a debris disk around the pulsar. In retrospect, perhaps that discovery should have told us that planet formation was a ubiquitous process. If planets can form around an exploding supernova, then we should have expected that exoplanets were common.

In 1952, astronomer Otto Struve made the remarkable assertion that if a Jupiter-like planet resided very close to its host star, that the gravitational tug of the planet on the star would produce radial velocity variations that might be detected in the stellar spectra with high enough precision measurements.

Twentieth-century astronomers worked to improve the precision of their techniques, and in 1995, the first exoplanet was finally discovered around a sun-like star. Most astronomers consider the dawn of exoplanets to be November 1995 when Michel Mayor and Didier Queloz discovered a gas giant planet orbiting the sun-like star 51 Pegasi at the Observatoire de Haute-Provence using the Doppler technique. The interpretation of this radial velocity signal as an exoplanet remained controversial for a few years but is no longer questioned. 51 Peg b was the first confirmed detection of an exoplanet around a main sequence star, and the 2019 Nobel Prize was awarded to Mayor and Queloz.

Direct Imaging

Just take a picture!

Seeing is believing, so it would be ideal if we could simply point a telescope at a star and take a picture of the orbiting planets. This method is called direct imaging, and the biggest challenge is separating the reflected light of the planet from the light of the star. The problem is that the planet is typically a billion times fainter and lost in the glare of the star.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Coronagraphs

One way to block out the light from a star so that some of the faint planets orbiting the star can come into view is to use an instrument called a coronagraph. Conceptually, this is similar to trying to see a faint object (like a tennis ball, coming over the net) that is lost in the glare of the Sun. If you use your hand to cover the Sun, you can now see the tennis ball. If you were lucky enough to be in the path of totality under clear skies during the solar eclipse of 2017, then you had the perfect demonstration of

how a coronagraph works. In this case, the moon acted as a coronagraph, allowing people to see stars close to the Sun in the “day” sky.

The video below is an animation showing how a coronagraph works. Before the light from the star can reach the detector, the coronagraph is used to artificially block out the starlight so that the image captured has most of this glare removed, as seen at the end of the animation when the exoplanet is revealed.

Using a Coronagraph to see the planets

Direct Imaging. What happens to the brightness of the orbiting planet when a coronagraph is used to block light from the host star?

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=162>

Animation source: <https://exoplanets.nasa.gov/>

Coronagraphs were developed in the 1930s by French astronomer Bernard Lyot, who was interested in studying the Sun's corona. Light from the Sun's photosphere completely blocks out the light from the corona, except during a total solar eclipse, so this was a technique that made studying the outer layers of the Sun possible. After exoplanet detections began ramping up in the early 2000s, coronagraphs began being used to take pictures of exoplanets. Coronagraphs still pose extreme challenges and only planets that are sufficiently far from their host stars can be revealed this way.

At the Keck telescope in Hawaii, astronomers used the adaptive optics (AO) system with a coronagraph to detect four planets orbiting the star HR 8799. This star was selected as a good target because it is relatively close (about 40 parsecs from the Sun) and because the star is young (a mere 30 million years) so that any planets would also be young and luminous. Exoplanets are named with lower case letters beginning with “b” in the order that they are discovered. In the image below, planets b, c, and d were found in 2008 and planet e was discovered a few years later. It is challenging to find planets close to the star where the coronagraphic suppression of light is more difficult.

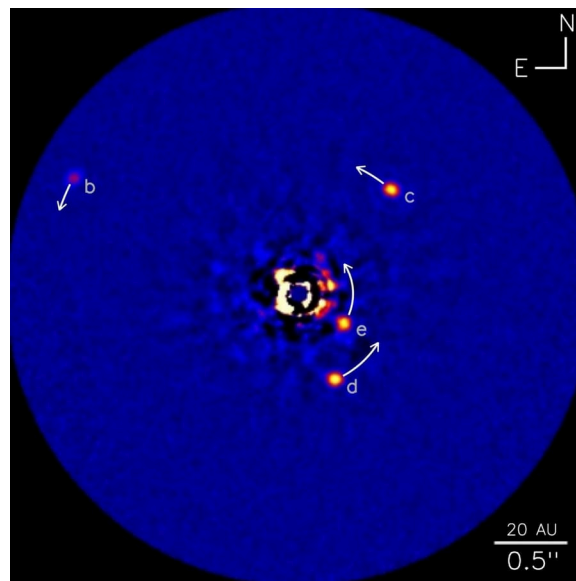


Figure 2 – HR8799. Four exoplanets orbiting HR 8799. The exoplanets, labeled b, c, d, and e (in the order they were discovered) were revealed using a coronagraph to mask out the light from this bright star. **What can we learn about observational biases for direct imaging by considering the order that planets were detected around HR 8799?**

Once the light from the planet's host star has been suppressed with a coronagraph, more information about the properties can be gleaned using other methods. Spectroscopy of the four planets around HR 8799 has been carried out using the Project 1640 instrument on the Palomar telescope and shows variable amounts of absorption from water, carbon monoxide, carbon dioxide, and methane.

The search for exoplanets using coronagraphs is an exciting technique, and impressive technological advances in this field are

permitting the detection of smaller planets located at smaller inner working angles (i.e., closer to the star). The James Webb Space Telescope has coronagraphs built into several of its infrared cameras, such as NIRcam and MIRI. In 2023, JWST took its first direct images of the exoplanet HIP 65426 b. The goal for the next 20 years is to develop space-based coronagraphs on large telescopes to image Earth-like planets around nearby stars and to obtain spectra to search for signatures of life.

Star Shades

Another way to block out starlight to enhance planets is to use a star shade. While coronagraphs are built into the telescope's optics, star shades are completely separate from the spacecraft carrying the telescope. The figure below shows the basic layout of a starshade system: the starshade would fly in formation with the spacecraft at a nearly fixed distance in front. And this distance is not tiny — the starshade needs to be located somewhere between 20,000 to 40,000 km ahead of the spacecraft, depending on the size of the spacecraft.

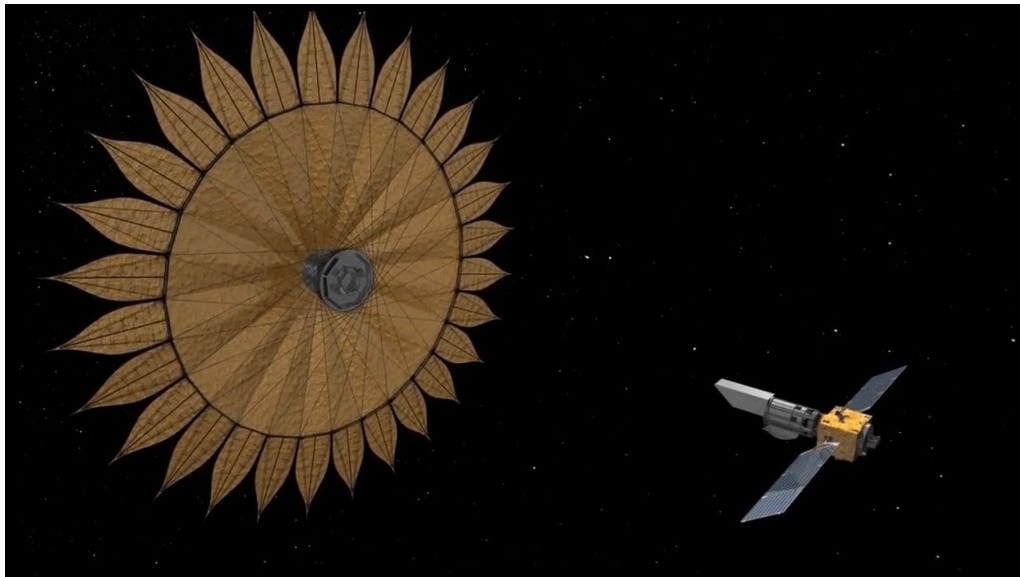


Figure 3 – Starshade. This artist's concept shows the geometry of a space telescope aligned with a starshade, a technology used to block starlight in order to reveal the presence of planets orbiting that star.

The distances need to be extremely precise, down to 1 meter (3.2 feet), to avoid having unnecessary light from the star leak through and contaminate the image. This is a technological challenge but engineers at NASA's Jet Propulsion Lab have carried out computer simulations that show this is feasible to do. The upcoming Nancy Grace Roman Space Telescope infrared mission will have a coronagraph on board and there is discussion of adding a starshade as well. The next generation of telescopes to image planets, such as NASA's [Habitable Worlds Observatory](#) (HWO), will include a coronagraph and a starshade.

Adaptive Optics

To obtain direct images of exoplanets around nearby stars with ground based telescopes, a technique called [adaptive optics](#) (AO) has been developed to correct for the atmospheric twinkling of starlight. AO combines optics, electrical engineering, and computers to correct for atmospheric distortion. In cases where there is more than one star in the field of view of the telescope, light from one of the nearby bright non-science stars is picked off with a mirror and sent to the AO system. If a bright star is not nearby, then a laser can be used to simulate a star on the sky. Hundreds of actuators on a deformable mirror in the AO system are moved around at high speed (thousands of Hertz) until the image of the star is concentrated into the smallest possible point of light. This counteracts the atmospheric distortion for all stars in the field of view, including the science star of interest.

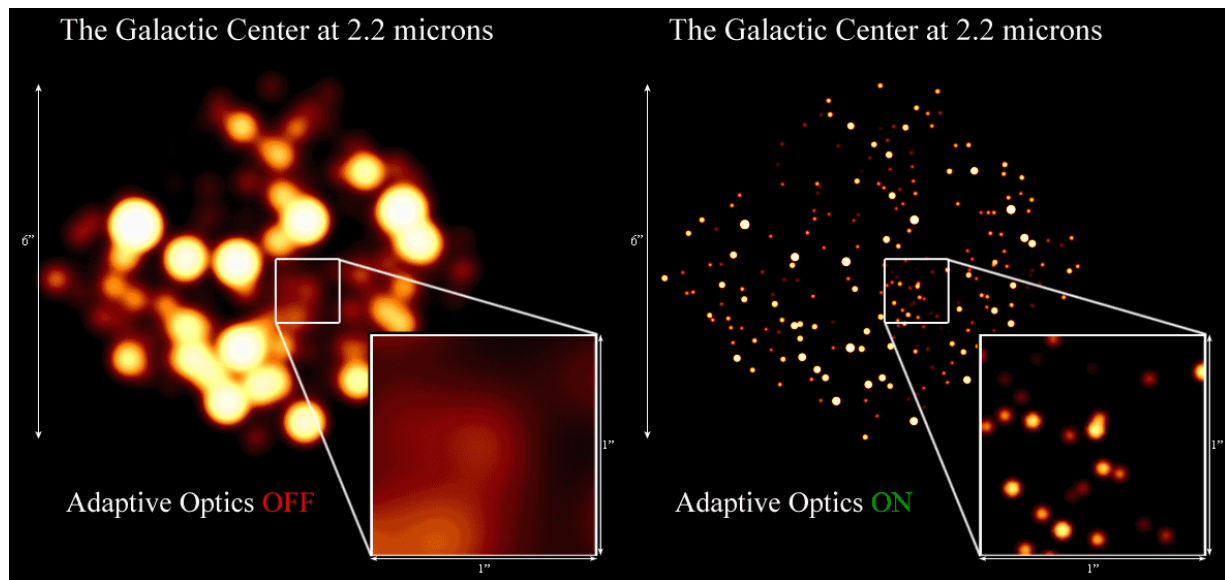


Figure 4 – AO Image of the Galactic Center. The blurred image on the left was taken with the Keck telescope with the Adaptive Optics system off. The image on the right shows the dramatic improvement in the resolution and image quality with the Adaptive Optics Center is on.

Direct imaging is the only method for finding exoplanets where the planet is directly seen. In the other methods, we must instead infer the presence of exoplanets by how they affect the star they are orbiting, either the star's motion or its brightness. These methods rely upon **indirect detection**.

The Doppler Technique

The Doppler technique (also called the radial velocity technique) was the first method to successfully detect exoplanets orbiting Sun-like stars. This technique measures the velocity of stars over time. All stars are traveling in orbits around the center of our Milky Way galaxy. For example, the Sun takes about 220 Myr to complete one loop around the galaxy. As the Sun travels around the galaxy, some stars appear to be moving toward us while others are moving away.

After subtracting the constant galactic velocity for a given star, a small remaining periodic motion in the velocity of a star (a “residual” velocity) can reveal that the star is being tugged around a common center-of-mass by another body, as shown in Figure 5 below. These residual velocities can be modeled to determine whether the orbiting object is a planet. With the Doppler technique, the planet is never observed (making this an “indirect” detection method). Instead, the time-varying velocity of the host star is modeled to infer the presence of an unseen planet. All stars have some nearly constant radial velocity; stars that exhibit a residual periodicity in their radial velocities have a gravitationally bound companion.

The Center-of-Mass of Solar Systems

The Sun is generally assumed to be the center of our solar system, with the planets, asteroids and comets all orbiting around it. This is almost true but there is a gravitational pull on the Sun from the planets, especially the massive planets. This tug on the Sun causes it to “wobble” around, moving in small loops. This looping pattern repeats but the center stays fixed and is called the **center-of-mass**. The center-of-mass is the point around which both the star and all planets revolve. This is shown in Figure 5a and 5b (left and right), where for simplicity, there is only one planet orbiting the star.

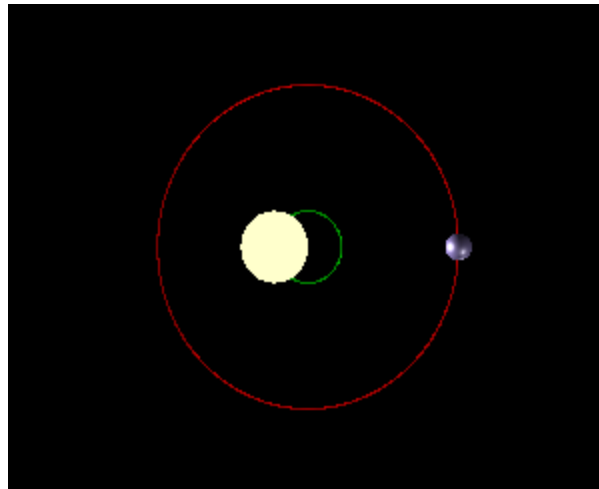


Figure 5a – Face-on Center of Mass. A star and planet orbit their common center of mass. The star “wobbles” (moves in a small circle) due to the gravitational tug from the planet. **When the star is on the left side of its orbit in this image, where is the planet?**

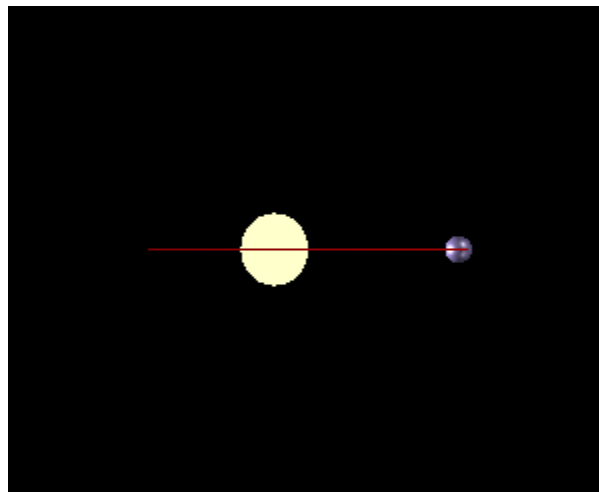


Figure 5b – Edge-on Center of Mass. A star and planet orbit their common center of mass. The star “wobbles” (moves in a small circle) due to the gravitational tug from the planet. **When the star is on the right side of its orbit in this image, where is the planet?**

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The Doppler Effect and Radial Velocity

How can this wobble of a star be used to detect exoplanets? The answer brings us back to spectroscopy. Specifically, the **Doppler effect**, which gives this technique its name. By taking a spectrum of a star, we can determine its temperature, chemical composition and motion. The motion we can find is called the **radial velocity** and is specifically the motion either directly toward or directly away from us. Think of drawing a radius between the instrument (a spectrograph) and the star — it is motion along this radial line we can determine. When a star (or any object emitting waves) is moving away from us, its spectral lines are shifted toward smaller wavelengths; this is called a **blueshift**. When the star is moving away from us, the spectrum shifts toward longer wavelengths, and this is called **redshift**. Figure 6 shows how this looks for the absorption lines in the spectrum of a star: the middle panel shows the lines if the star has no radial motion; in the top panel, the star is moving away from us and its lines are moved toward the red end of the spectrum; and in the bottom panel the star is moving away from us and its lines are moved toward the blue end of the spectrum.

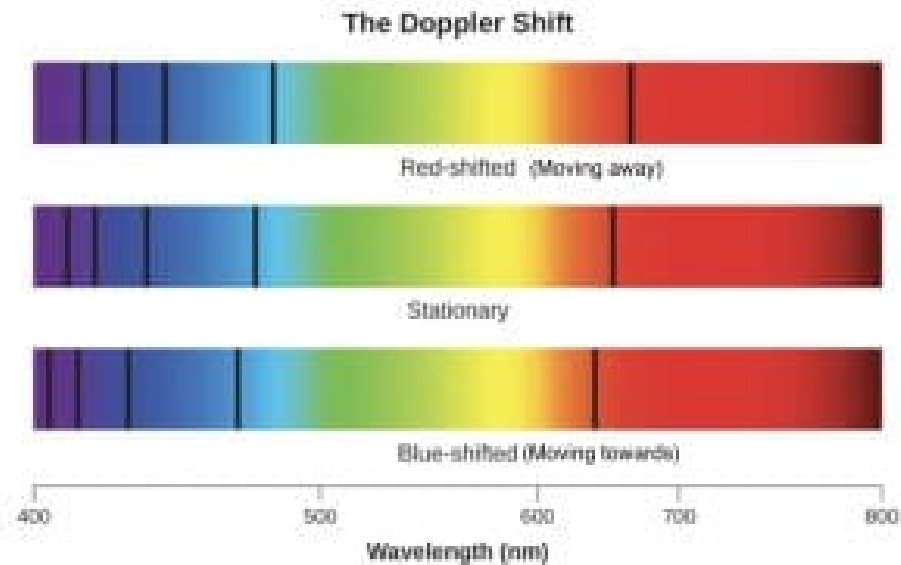


Figure 6 – Doppler Shifted Absorption Lines. **When the star is moving away from us, which direction do the absorption lines shift?**

If we observe a star long enough and find that its spectral lines are shifting back and forth, then the presence of another object (in this case, a planet) can be inferred. Keep in mind that the spectral lines we observe are for the star and not the planet; the planet's reflected light is far too faint to have its spectrum recorded. This can be seen in the animation below (Figure 7), where the spectral lines are for the star.

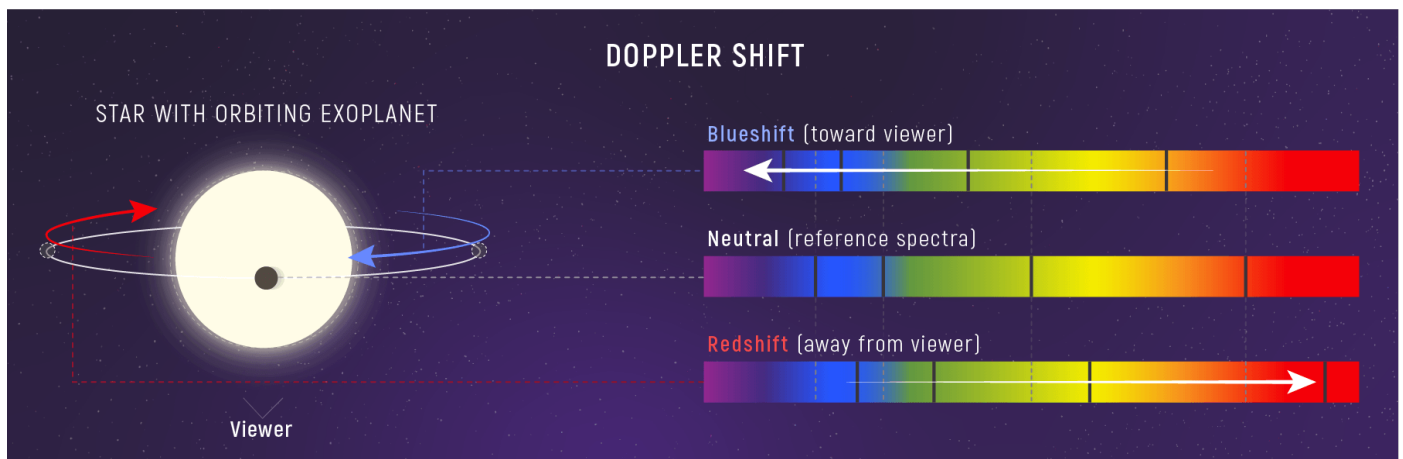


Figure 7 – Doppler Shift of Stellar Spectrum. As a planet orbits a star, the star wobbles. This causes a change in the appearance of the star's spectrum called Doppler shift. While the star is moving toward us, its spectrum is blueshifted (top spectrum): All of the wavelengths look compressed and absorption lines appear at shorter (bluer) wavelengths than when the star is not moving toward or away from us (middle spectrum). While the star is moving away from us, the light is redshifted: All of the wavelengths look stretched out and the absorption lines appear at longer (redder) wavelengths (bottom spectrum). Because the change in wavelength is directly related to relative speed, astronomers can use Doppler shift to calculate exactly how fast an object is moving toward or away from us. Astronomers can also track the Doppler shift of a star over time to estimate the mass of the planet orbiting it. **As the star is moving away from us, what direction is the planet moving? Do the spectral lines move redward or blueward at this time?**

The original version of this chapter contained H5P content. You may want to remove or replace this element.

In reality, high-resolution spectra are needed to find the wavelength shifts to the precision needed to determine properties of an exoplanet. A segment from an extracted high-resolution spectrum around the deep pair of sodium absorption lines is shown in the animation (Figure 8) below, where each shift of the spectrum simulates a different velocity shift. By measuring the periodic shift of the wavelengths for these lines ($\Delta\lambda$) relative to the rest wavelength (λ_0 , indicated by the red vertical lines for sodium), the velocity of the star over time can be calculated.

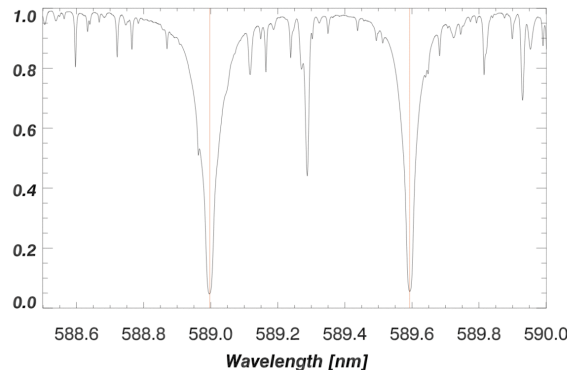


Figure 8 – Doppler Shift of High Resolution Spectrum. The Doppler shift is measured from each spectrum of the star and the time series radial velocities are calculated. The red lines indicate the rest wavelengths for the two sodium lines at 589.0 nm and 589.6 nm. **How is the velocity of the star determined from the shifting of the atomic absorption lines?**

The animation above is an extreme exaggeration (for the purpose of illustration) of the reflex Doppler shift that would occur from orbiting exoplanets. The spectral absorption lines in the animation above are between 5-20 pixels in width. A spectral line shift of just one pixel on a detector corresponds to a radial velocity change of about 1000 m/s. The amplitude of Doppler shifts caused by exoplanets would be invisible to the eye on the scale shown above.

To determine the radial velocity of a star from its spectrum, you only need to measure how far the wavelengths of the absorption lines have shifted due to the motion toward or away from us. This shift in wavelength, $\Delta\lambda$, is related to the radial velocity, v_{rad} , as follows:

$$\frac{\Delta\lambda}{\lambda_0} = \frac{v_{rad}}{c} \quad (5.2.1)$$

Here, λ_0 is the wavelength of a line if there was no radial motion, meaning the middle panel of Figure 6. We sometimes refer to λ_0 as the rest wavelength or lab wavelength. The constant c is the speed of light.

This same formula will give the reflex radial velocity for a star that is wobbling due to the presence of a planet, as shown in the example below.

Worked Example: Finding the Reflex Radial Velocity

You observe a star and notice that the spectral lines are moving back and forth. You know that the rest wavelength, λ_0 , of one of the lines is 656.28 nm and observe over time that it shifts back and forth by 0.00012 nm (a very small shift). What is the reflex radial velocity of this star?

We are given the shift in wavelength: $\Delta\lambda = 0.00012$ nm and that the rest wavelength is 656.28 nm. In one step, we can find the reflex radial velocity:

$$\frac{v_{rad}}{c} = \frac{\Delta\lambda}{\lambda_0} \quad (5.2.2)$$

$$v_{rad} = c \frac{\Delta\lambda}{\lambda_0} \quad (5.2.3)$$

Notice that the units of $\Delta\lambda$ and λ_0 are the same — they are both in nanometers. They can be in any units as long as they are the same and thus cancel each other out.

What does this value mean? This is the reflex (or residual) radial velocity — as the star moves in its small loop, the radial velocity varies and 54.8 m/s is the maximum value it reaches. The value will reach a maximum speed of 54.8 m/s and a minimum speed of -54.8 m/s. [Figure 10](#) below shows this idea, where in that case E is the maximum and A is the minimum value for the radial velocity.

Our ability to detect smaller planets is limited by the precision with which we can measure the shifts of spectral lines. Jupiter induces a radial velocity in the Sun of about 12 m/s with a periodicity of about 12 years. To detect Jupiter, we would need to measure a 12-year back-and-forth shift in the spectral lines of just 0.01 pixel. (The raw data collected by a spectrograph records data as pixels; the pixels are converted to wavelengths during processing.) This requires that we control the stability of our instrument to much better than this precision and maintain the control over at least 12 years. Impressively, this is the state of the art: astronomers can measure wavelength shifts to the level of 0.001 pixel (1 m/s) with decades-long stability. However, detecting the stellar reflex velocity from a planet like the Earth is a much greater challenge. The Earth induces a velocity of about 10 cm/s in the Sun. This corresponds to a one-year back-and-forth shift of 0.0001 pixel. New [instruments](#) are now being developed to try to find other examples of [Earth-like planets](#) that might harbor life.

Mass and Inclination of a Planetary System

Kepler's Third Law relates the orbital period of a planet around a star to the planet's average distance from the star: $P^2 = a^3$, where P is how long it takes the planet to make one orbit around the star and a is the semi-major axis of the orbit (or the average distance of the planet from its star). This handy version requires that P be given in years and a given in astronomical units (AU). After formulating the Law of Gravitation, Newton generalized this law to include the masses of the two orbiting objects (for a review, see [Newton's version of Kepler's Third Law](#)):

$$(M_1 + M_2)P^2 = a^3 \quad (5.2.4)$$

This means that if we know P and a , we can get the total mass of the planetary system. The mass of the star can be estimated based on its spectral type, so this means we can find the mass of the exoplanet! But the full mass of the exoplanet cannot always be determined — it depends on the orientation of the star system.

Looking carefully at Figure 5 again, notice that we are essentially viewing the star system **edge-on**, meaning that all of the star's movement is toward/away from us and along our line-of-sight. But what if the planetary system was tilted along our line of sight? Let's imagine the extreme case where we are seeing viewing the system **face-on**, as in the right panel in Figure 9.

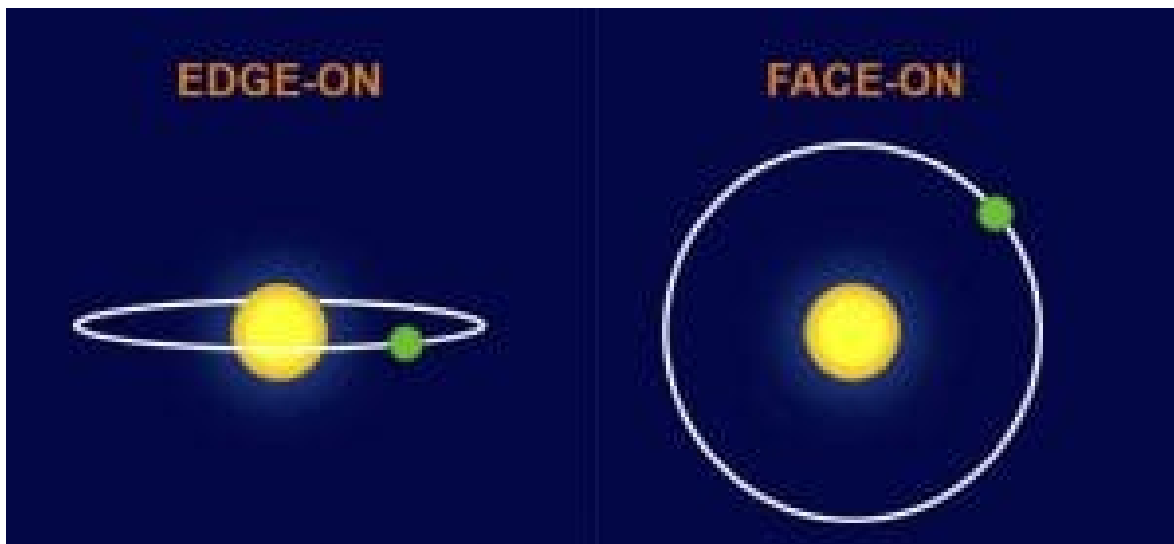


Figure 9 – Edge-on and face-on views of a planetary system. In the edge-on case (left), all of the star's motion is along our line of sight and the full radial velocity signal can be measured. **When the system is face-on in the plane of the sky, is there any motion toward or away from us?**

In the case of a face-on orientation of the planetary system, as shown on the right in Figure 9, none of the stars motion is along our line of sight so the Doppler technique will not yield any information. In reality, planetary systems can have *any* orientation along our line of sight. The orientation of a planetary system is characterized by a parameter i , which stands for inclination.

For the edge-on case, $i = 90^\circ$ and for the face-on case, $i = 0^\circ$. For a system that has an inclination that is less than 90° , the exact mass cannot be determined but the *minimum* mass that the planet could have is found. The minimum mass depends on $\sin i$, and a full quantitative description of how to find this minimum mass is given in the box below.

Detecting planets from velocities

To detect a planet with the Doppler technique, many spectra must be obtained that span the complete orbital period of the star-planet system. Of course, astronomers don't know the orbital period in advance, so in practice, they take as many spectra as possible and then use mathematical and statistical methods to search for repeating changes in the radial velocity.

Figure 10 depicts the orbit of a star as a dashed circular path that is viewed edge-on by an observer to the right. The velocity of the star is constant in this figure, however the projected radial velocity along our line of sight is changing. At point A, the star moves toward the observer and the spectrum of the star is blue-shifted (the measured wavelengths are shorter). At point E, the star moves away from the observer and the spectrum of the star is red-shifted (the measured wavelengths are shifted to longer, redder wavelengths). At point C, there is no radial velocity – all of the motion is tangential to our line of sight. At points B and D, only the projected radial component of the velocity (velocity times the sine of the orbital inclination) can be measured in the stellar spectrum.

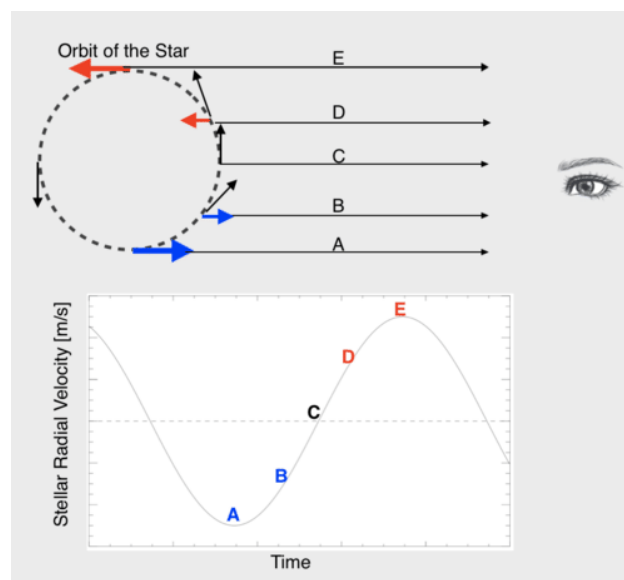


Figure 10. The Doppler technique is only sensitive to velocity along our line of sight (radial velocity). As the star orbits the center of mass, the projected component of the radial velocity is calculated by measuring the change in the wavelength of absorption lines in the stellar spectrum.

What is the radial velocity that we measure when the star is traveling tangential to our line of sight (point C)?

The radial velocity data for the star can be used to find the orbital parameters — and determine the minimum mass — of the planet. Details are given in the box below.

Want to know more: Finding exoplanet masses from velocity data

From the radial velocity curve of the star, the orbital period of the planet can be found (it is the same as the observed period for the star's wobble) along with the amplitude of the radial velocity for the star. The semi-amplitude, which is half of the full amplitude, is denoted as K and is shown in Figure 11 below.

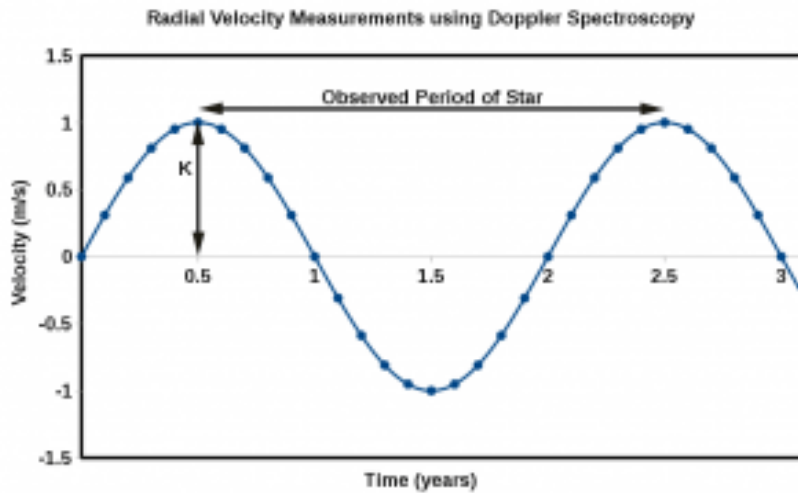


Figure 11 – The radial velocity for a star with a companion as a function of time.

The orbital parameters can then be used to calculate the semi-major axis and the mass of the planet times the orbital inclination using Newtonian physics.

Missing or unrecognized delimiter for \left

where

$$m_p \quad (5.2.5)$$

$$M_* \quad (5.2.6)$$

$$P \quad (5.2.7)$$

$$e \quad (5.2.8)$$

$$i \quad (5.2.9)$$

This can be rearranged to give the formula for the minimum mass of the exoplanet:

$$m_p \sin i = K \sqrt{1 - e^2} \left(\frac{P}{2\pi G} \right)^{\frac{1}{3}} (M_* + m_p)^{\frac{2}{3}} \quad (5.2.10)$$

Statistical probability for $\sin(i)$

For randomly oriented inclinations, the probability of drawing a particular $\sin(i)$ is not uniform. Instead, the probability of observing an orbit with an inclination Δi between i_1 and i_2 scales as the difference between the cosine of those angles:

$$\mathcal{P} = |\cos i_1 - \cos i_2| \quad (5.2.11)$$

So, there is an 86% probability that the inclination of a random orbit is between 90° and 30° . For these typical orbits, the true mass of the planet would be within a factor of two of the measured $m_p \sin i$. Conversely, there is only a 14% probability that an orbital inclination is between 30° and 0° (yielding a true mass more than twice $m_p \sin i$).

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Real-world examples

In 1995, Michel Mayor and Didier Queloz discovered the first exoplanet around a main sequence star using the Doppler technique. The planet that they discovered has a mass similar to Jupiter's and orbits the star 51 Pegasi once every 4.23 days. The data from their 1995 discovery paper are shown in Figure 12 below. Because the signal repeats, the velocities can be folded so

that the orbital periods repeat. This “phase-folded” data is a helpful way to display relatively short periodic signals and clearly demonstrates that the signal is repeating in a predictable way.

In the data from the Mayor & Queloz 1995 paper, you can see that from September to November in 1994, the radial velocities for 51 Pegasi were changing. This perplexed Didier Queloz, who was a graduate student working with Professor Mayor. To better understand the velocity variations, he began taking more data. Mayor and Queloz developed a mathematical model (shown as the solid line) that fits the velocity changes of the star and also describes Keplerian orbital motion. It is quite an empowering feeling to develop a model like this and to then watch the future velocities of the star follow your predictions.

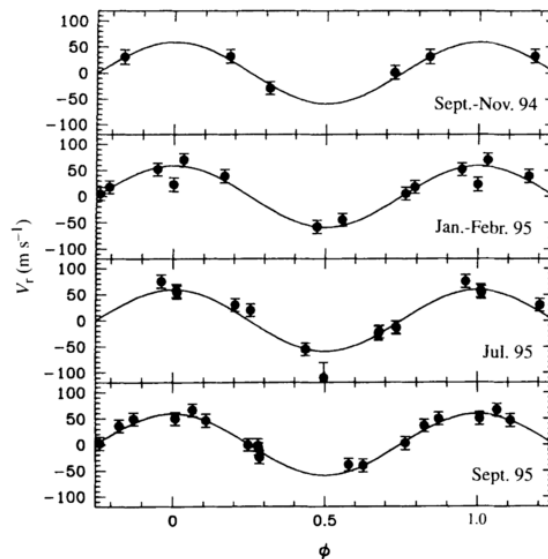


Figure 12 – Radial velocity of the star 51 Peg. **The radial velocity data in the images above were taken at different times between 1994 and 1995, but are plotted against an angle; what does this mean?**

Queloz and Mayor could not quite believe what they were seeing. A Jupiter-mass planet in a 4-day orbit was very different from any of the planets in our solar system. In their paper, they discuss the alternative interpretations such as stellar pulsations and effectively rule them out. They conclude that the most likely interpretation is an orbiting planetary mass companion. Massive exoplanets that are orbiting very close to their host star are called **hot Jupiters**. 51 Peg b is considered the poster child for hot Jupiters.

Because they derive the minimum mass, they note that the true mass may be larger than one Jupiter mass. Some skeptics believed that Mayor and Queloz had simply detected a stellar binary system that was oriented nearly face-on, which is a totally reasonable scenario to consider. However, an inclination of less than 1° is required to pump the minimum mass up to the stellar mass regime of 70 Jupiter masses. It can be rigorously shown that probabilistically there is only 1 chance in 10,000 that they would have observed such a low inclination orbit.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Want to know more: the first few exoplanet discoveries

Mayor and Queloz’s discovery of an exoplanet in such a short orbital period was a game-changer. A team of astronomers at San Francisco State University were also trying to find exoplanets using the Doppler technique. They had a backlog of data because they were collecting about 50 spectra per night, but it took 24 hours to analyze a single spectrum on the slower computers of the day. Spurred on by the discovery of 51 Peg b, Sun Microsystems made a donation to the SFSU team and they quickly crunched through their observations, discovering Jupiter-like planets around 70 Virginis and 47 Ursae Majoris in their sample of 120 stars. The second of these, 47 UMa b, was very much like 51 Peg b – a Jupiter-mass planet in an orbit of just a few days.

However, the data for 70 Vir b (shown below) told a different story. Instead of the sinusoidal velocity variation seen with 51 Peg b, this planet has a more peaked velocity variation revealing an orbit with eccentricity of 0.4. The planet mass (6.6 times the mass of Jupiter) is also much larger, producing greater velocity variations in 70 Virginis. To the skeptics, the discovery of yet another massive planet was further support for the idea that nearly face-on binary star systems were being discovered rather than

exoplanets. The fact that this object was in an eccentric orbit added to the doubts – planets in our solar system are in nearly circular orbits. Surely the rest of the cosmos would be like us?

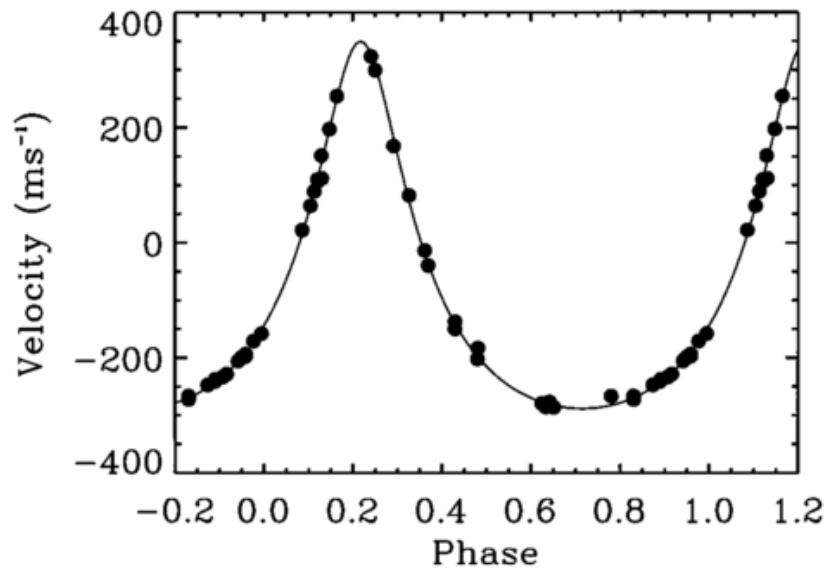


Figure 13 – Phase-folded radial velocities for the star 70 Virginis reveals a planet with a mass of 6.6 times the mass of Jupiter in a 116 day orbit. **Why does the shape of the radial velocity curve look different from the sinusoidal variation seen for 51 Peg b?**

Practice with the radial velocity simulator

There are a lot of concepts to understand with the Doppler technique. First, we use the spectrum of the star to measure wavelength shifts that correspond to changes in the velocity of the star. Second, we see only the projected radial component of the stellar velocity (this leads to a sinusoidal variation in the signal). Third, we don't know the inclination of the orbit – this tilt of the orbit means that the sinusoidal signal will be largest when viewed edge-on (90° inclination) and smaller as the inclination tilts toward face-on. This means that we only measure the minimum mass ($m_p \sin i$) with the Doppler method.

Let's get some practice with the Nebraska Astronomy Applet Project (NAAP) Radial Velocity [Simulator](#) to help develop your intuition and understanding of the Doppler method. Simulated radial velocities are “phase-folded” in the upper right hand plot. The radial velocities are obtained as a function of time, but if you repeat observations over more than one orbital period, the data can be folded so that all of the peaks and troughs of the curve line up. This is routinely done by astronomers to build evidence for repeating signals.

The default shows a star with the mass of the Sun being orbited by one Jupiter-mass planet with a perfectly circular (eccentricity is zero) orbit; the system is completely edge-on, with an inclination of 90° . To change values, you can either enter them in the box (be sure to hit the return key) or use the sliders. What happens to the radial velocity as you lower the inclination? (Recall that a face-on system has an inclination of 0° .)

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Key Concepts and Summary

In 1995, a planet was detected orbiting the sunlike star 51 Pegasi, launching the new field of “exoplanets” in astronomy. There are several different ways to detect exoplanets. Here we reviewed two methods for detecting exoplanets: direct imaging, where the light from the host star is suppressed so that the planet can be observed and the “Doppler” method, which detects the gravitational tug from an unseen planet that changes the radial velocity of the host star. Each of these observational methods is sensitive to different masses and orbital distances of planets.

Review Questions

Summary Questions

1. What is the main challenge faced in taking pictures (direct imaging) of exoplanets?
2. What is meant by the center of mass of a solar system? Make a sketch to help convey this idea.
3. How can the absorption spectrum of a star be used to show that the star has exoplanets orbiting around it?
4. How can the shift in spectral lines for a star with a planet be used to find the radial velocity of the star?
5. How can Newton's version of Kepler's Third Law be used to find the mass of a planet?
6. Why can only the minimum mass be found? In other words, what is needed to find the full mass?

Exercises

1. Consider the two methods for detecting exoplanets discussed in this chapter and make a chart listing the advantages of each method and what type of exoplanet it is best at finding.
2. RV simulator activities

5.3: Transit, Astrometry and Gravitational Microlensing Techniques

Learning Objectives

- Describe the transit method for detecting exoplanets and how the radius of an exoplanets can be found
- Explain how astrometry can be used to detect exoplanets
- Describe gravitational lensing and how microlensing can be used to detect exoplanets

Transit Method

By 1999, about a dozen planet candidates had been detected with the Doppler technique, and most were similar to the hot Jupiter 51 Peg b, with minimum masses of about one Jupiter mass and orbits of just a few days. Even the astronomers working with the Doppler method were deferential in calling these objects planet “candidates” in those early days because the true mass of the orbiting companion could not be determined with this detection technique.

However, astronomers knew that if these really were planets (and not face-on stellar binary systems), about one in ten of the short-period 51 Peg b type objects would have an orbit that was oriented so that the planet would “transit” or cross our line of sight to the star. When that happens, the planet blocks some of the starlight for the few hours that it transits.

Measuring the brightness of stars is relatively straightforward, and each of the stars with 51 Peg b type exoplanets were being monitored for a possible transit. While the Doppler technique cannot determine the inclination, the data do predict when a planet will transit if the inclination is viewed nearly edge-on. If a planet transits, we know that i must be close to 90° and, therefore, $\sin i \sim 1$. In this case, the Doppler measured mass is no longer just the minimum mass; it is the total, true mass of the planet.

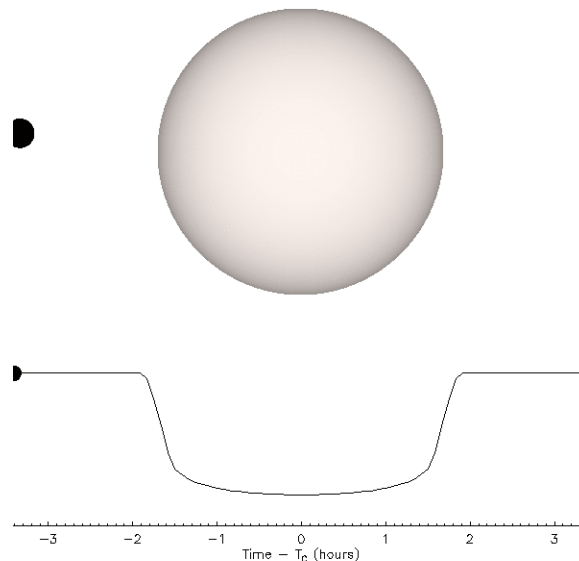


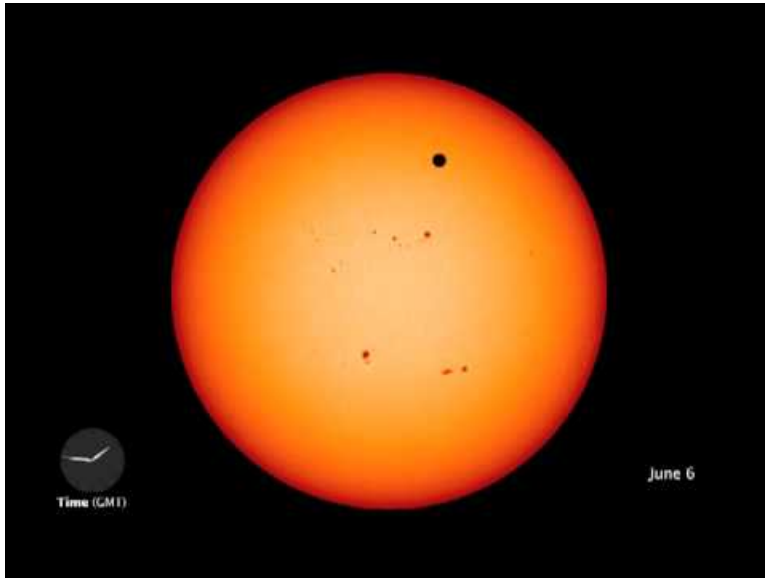
Figure 14 – A transit curve. When orbits are viewed edge-on, the planet transits, or crosses our line of sight to the host star. **What is the relationship between the size of the planet and the amount of starlight that is blocked (i.e., the depth of the transit)?**

Figure 14 above illustrates what is happening. During a transit the planet passes in front of the star (top animation), and as this happens the brightness of the star decreases (bottom animation). We do not actually have pictures like the top animation because stars are point sources of light, not resolved objects. The data that we have are the time series photometry (brightness measurements) depicted in the lower animation that show a regular and periodic decrease in the brightness of the star at precisely the putative transit time (predicted by the Doppler velocity measurements). Because we cannot resolve the image of the planet (or the star), this is still an indirect detection technique. However, even astronomers sometimes take some poetic license

here and say that we are seeing the shadow (or the silhouette) of the planet.

There is only one case where we can resolve the physical extent of star, and that is for our Sun. If the Earth is in the right position on our orbit around the Sun, we can see the inner planets, Mercury and Venus, transiting the Sun. This fortuitous alignment happens about 13 times a century for Mercury and fewer than 2 times per century for Venus. The video below shows a simulation of the 2012 the transit of Venus. You will have to wait until Dec 10, 2117, to see the next transit of Venus. **Can you estimate the transit depth for Venus crossing the Sun?**

The Transit of Venus



A YouTube element has been excluded from this version of the text. You can view it online here: <https://pb.libretexts.org/astrobio/?p=167>

Video Credit: @NASAgovVideo

The first transiting exoplanet

HD 209458 b was the first star where astronomers observed a transiting exoplanet. The planet had a mass similar to Jupiter's with a short orbital period of just a few days, and was independently detected by two teams using the Doppler method in the summer of 1999. HD 209458 b was the twelfth example of a hot Jupiter (a 51 Peg b analog) — recall that about one in ten close-in planets was expected to transit — and the planet hunters were anxiously awaiting the validation of “planet” status that a transit would bring.

In August and September 1999, the team led by Harvard graduate student David Charbonneau and Dr. Tim Brown, observed a transit at the time predicted by the Doppler velocity data. Their published data are shown in Figure 15. The ground based photometry (brightness measurements) is good enough that you can unambiguously see the transit event, but “blurring” caused by the Earth's atmosphere caused some of the starlight to be scattered. This blurring effect is what causes stars to appear to twinkle on the night sky. The team was able to combine data from multiple transits to reduce their errors to 0.015%. This precision is very impressive, but not high enough to detect a transiting planet like the Earth with a radius that is one tenth of Jupiter. A second team, led by Greg Henry, measured a partial transit in November of that same year, and the two discovery papers were published back-to-back in *The Astrophysical Journal*.

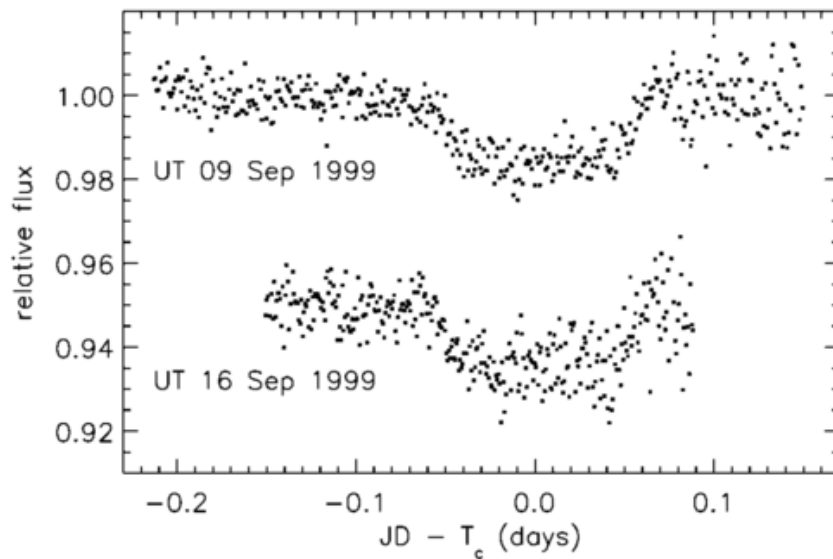


Figure 15 – Photometry (measurement of brightness) of the star HD 209458 showed a decrease in brightness for a few hours in August 1999 and again in September 1999 at the time when the Doppler observations suggested that the planet should be passing in front of the star. **What is the duration and fractional depth of the transit of HD 209458b? What causes scatter in the photometric measurements?**

Transit Depth

The depth of the transit tells us about the relative size of the planet and star. Imagine a planet that is large enough to completely cover (eclipse) the star during transit. In this extreme case, the brightness B of the star would drop to zero. However, the typical case is that planet is *much* smaller than the star. The planet's smaller circular disk blocks the light of the star's larger circular disk (stars and planets are both spheroidal and their cross-sections — a 2-D slice through — are very well-approximated by circles). The amount of light the planet blocks, called the **transit depth**, is then given by:

$$(\text{area of planet})/(\text{area of star}) = \frac{\Delta B}{B} = \frac{\pi r_p^2}{\pi R_*^2} = \left(\frac{r_p}{R_*}\right)^2$$

The transit depth equation above says that the fractional change in brightness ($\Delta B/B$) of the star is proportional to the radius of the planet (r_p) to the star (R_*). We measure the brightness of the star without a transit and the decreased brightness during transit. So, if we know the radius of the star, we can solve this equation to derive the radius of a planet that we cannot even see. Science is amazing!

In our everyday lives, we have a lot of experience with what artists call perspective. Objects that are closer to us look larger than objects that are farther away. This sometimes lead to an incorrect intuition that transit depth is related to the star-planet separation since the more distant planet are “closer” to us. However, the incredible distances to the stars that we observe mean that the difference in star-planet distances within a given planetary system are negligible.

Worked Example: Transit Depth

Calculate the transit depth for a star the size of the Sun with a gas giant planet the size of Jupiter.

The radius of Jupiter is 71,400 km, while the radius of the Sun is 695,700 km. Substituting into the equation, we get

$$\text{transit depth} = (R_{\text{planet}}/R_{\text{star}})^2 = (71,400 \text{ km}/695,700 \text{ km})^2 = 0.01 \text{ or } 1\%$$

which could easily be detected with the instruments on board the Kepler spacecraft.

Check Your Learning

What is the transit depth for a star half the size of the Sun with a much smaller planet that is the size of Earth?

Show Solution

We can solve this problem the same way, this time dividing the radius of the Sun in half and with Earth's radius (6,371 km) instead of Jupiter's:

$$\text{transit depth} = (R_{\text{planet}}/R_{\text{star}})^2 = (6,371 \text{ km}/347,850 \text{ km})^2 = 0.0003 \text{ or } 0.03\%$$

We could also use the fact that Jupiter's radius is about 11 times larger than the Earth's radius, and scale the answer for the Jupiter-sized planet:

$$0.01 \times (R_{\text{planet}}/R_{\text{star}})^2 = ((1/11)/(1/2))^2 = 0.01 \times (2/11)^2 = 0.01 \times 0.033 = 0.0003$$

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The Transit Simulator

Let's use the [NAAP transit](#) simulator from UNL to understand how transiting exoplanets are detected with time series photometric data. The default shows a gas giant planet (see the Planet Properties box and note that the mass and radius are relative to Jupiter) orbiting a star slightly more massive than the Sun; the inclination of the system is 86.9°. Jupiter has a radius that is 11.2 times larger than the Earth's, meaning that the Earth's radius is 0.089 times the radius of Jupiter. What happens to the light curve (upper right) when you change the radius of the planet to Earth's size? You can quantify the difference by looking at the value for the eclipse (transit) depth reported to the lower right of the light curve. How do you expect the light curve to change, if at all, if you lower the mass of the exoplanet?

Now let's look at the data for the first exoplanet detected with the transit method, HD 209458 b. In the Presets box, choose this planet from the drop-down menu and click "set". What is the eclipse depth? What is the inclination of this system? At what inclination would you no longer be able to measure the transit? The eclipse depth will show (no eclipse)" when there is no transit observed. (add show answer here, it's 82.7°) [will also add longer activities at end of chapter]

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The Kepler Mission

A decade before the first exoplanet was discovered, William Borucki, a scientist at NASA Ames Research Center in Mountain View CA, wrote a paper "The Photometric Method of Detecting Other Planetary Systems" (Borucki & Summers 1984) where he made the case for a space mission to detect planets orbiting other stars using the transit technique. It took more than twenty years of advocacy, but Borucki's persistence paid off. The first exoplanets were discovered in the mid-1990s, heating up interest, and after the first ground-based transit discovery in 2000, the exoplanet community rallied behind Bill Borucki's plan.

The first question in deciding whether to go forward with a space mission is: can this be done from the ground? If so, then the mission is a no-go. The justification for the space-based *Kepler* transit mission was the need to reach higher precision. Ground-based transit surveys could detect only 1% (0.01) changes in the brightness. The important limitations to the precision of ground-based photometry (brightness measurement) were blurring from the Earth's atmosphere and the limited number of dark nighttime hours when the telescopes point away from the Sun. From space, the atmospheric blurring is limited and the telescope can always point away from the Sun. The higher photometric precision was demonstrated by the Hubble Space Telescope, which measures 0.1% (0.001) changes in brightness. However, a special purpose telescope was really needed that could reach even higher precision with continuous monitoring of stars.

In 2009, NASA launched the *Kepler* Mission, which measured 0.01% or (0.0001) changes in the brightness of the host star. Figure 16 below shows that *Kepler* was remarkably successful in detecting a range of exoplanets with radii down to a couple times the radius of Earth.

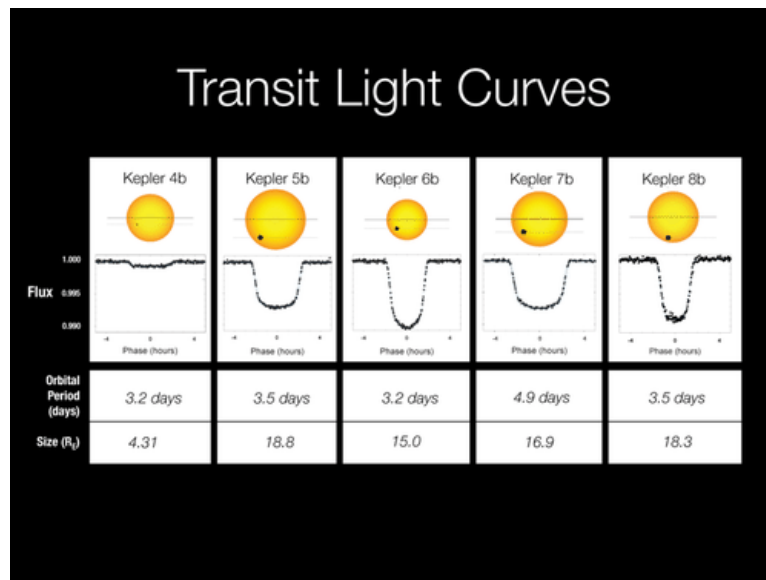


Figure 16 – Transit light curves from the NASA Kepler mission illustrate the difference in transit depths for different size planets. Compare the size of the planets for the deepest transits with the size of the planets for the shallowest transits. **The radius of Kepler 5b is larger than the radius of Kepler 6b — why might Kepler 6b have a deeper transit depth?**

The specific science goal for *Kepler* was to determine the occurrence rate of Earth-sized planets. The HAT-P-7 light curve in Figure 17 below compares a light curve from a ground-based transit survey for the planet HAT-P-7b with data from the *Kepler* space mission. What a difference the higher precision made!

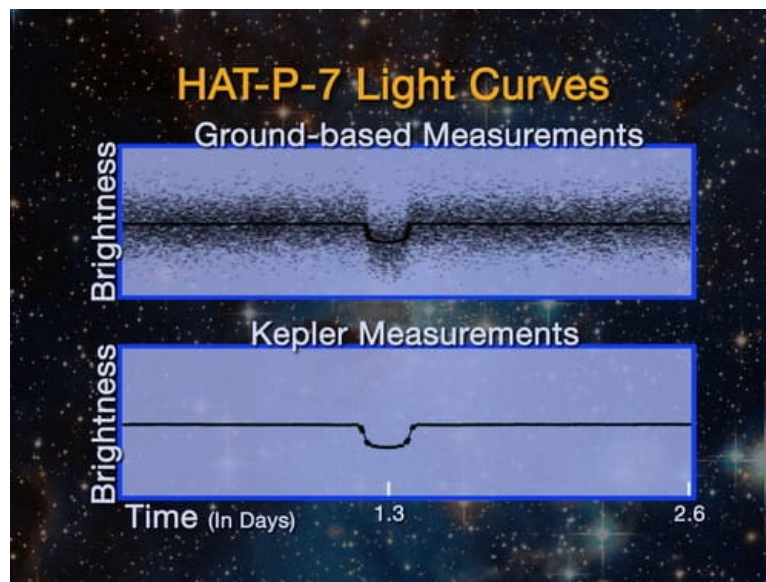


Figure 17 – This transiting planet was discovered with the ground-based Hungarian Automated Telescopes (HAT, top light curve). The host star was later observed with the *Kepler* spacecraft (bottom light curve). **Why is there more scatter in the data in the top light curve, compared to the *Kepler* light curve?**

In 2011, the first rocky exoplanet was discovered with the *Kepler* mission, known as Kepler-10 b. The lead author on the discovery paper, Natalie Batalha, was also the Deputy Scientist for the space mission, and her collaborative leadership style helped to make the NASA mission an incredible success. You can participate in the hunt for transiting planets in the *Kepler* data on the Zooniverse [Planet Hunters](#) website.

Multi-planet Transiting Systems

So far, we've only considered cases where there is just one transiting planet. In our solar system, the planets are nearly co-planar.

What if there is more than one transiting planet in the time-series light-curve data? Would we be able to disentangle the different signals? If the planets are different sizes, then the transit depths will also differ. A second important parameter is the orbital period. Kepler's third law says that the orbital period is related to the semi-major axis of the orbit so that more distant planets have longer period orbits. With longer period orbits, the time interval between transits is also longer. So there are two important pieces of information that go into a model of a transiting system: the spacing between transits (i.e., the orbital period) and the depth of the transits.

The simulation below shows a system with three planets with different sizes and distances from the host star. Before playing the simulation, can you predict how the light curve will look when two planets of different sizes cross in front of the star?

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=167>

Animation source: <https://exoplanets.nasa.gov/>

A good example of what actual data looks like for a system with more than one planet is shown in Figure 18 below for the Kepler 11 system. The six planets in this system all orbit closer than the distance to Mercury in our solar system.

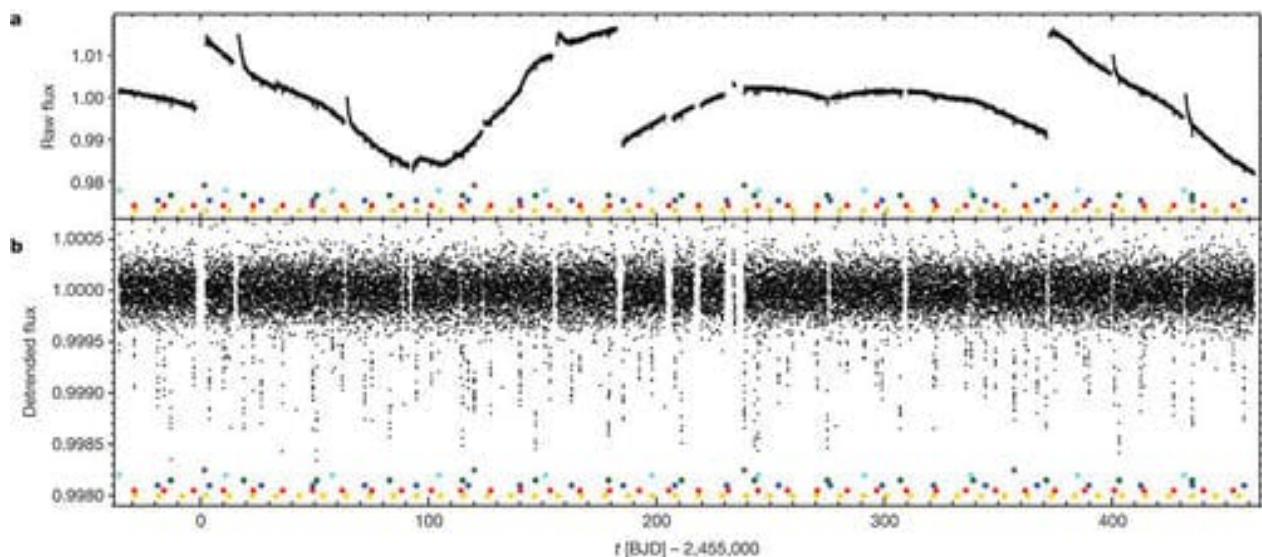


Figure 18 – The time series light curve of Kepler-11. The top panel shows the original photometry collected over about 500 days. After cleaning and detrending instrumental errors in the data, the time series light curve (bottom panel) shows several transits with different depths (corresponding to different sized planets). **What causes the depth of the light curves to change for each planet? Do you see a trend in the width of the transit dips going from the innermost planet (top) to the outermost planet (bottom)? What causes this?**

After detrending the light curve to remove instrumental errors in the photometry, transits of the individual planets are phase-folded and shown below.

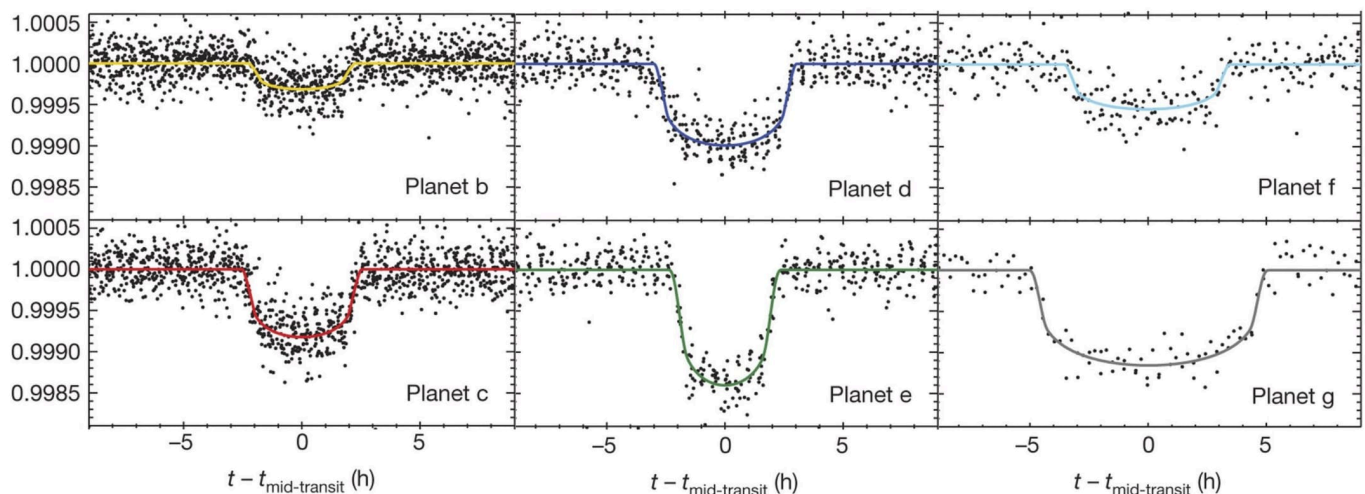


Figure 19 – Transit curves for a multi-planet system. **What causes the difference in the transit depths? What causes the difference in the transit widths?**

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Astrometry

Astrometry is one of the oldest methods that astronomers have for studying the motion of stars. The name says it all: astrometry is the process of measuring the positions of objects in the sky. This technique has been used to measure parallax (distances) to stars and proper motions (motions in the plane of the sky) and to detect binary star orbits. More recently, astronomers have tried to detect exoplanets with this technique. This technique has not been very successful from ground-based telescopes because distortions from the atmosphere (“twinkling” of stars) blurs out the spatial position of stars and therefore does not permit precise enough measurements. However, astrometry as a planet detection technique is about to explode onto the scene with the launch of the space-based Gaia spacecraft in 2013 by the European Space Agency. The [Gaia mission](#) is still collecting data and, as we’ll see, it takes many, many years of data to detect exoplanets using astrometry, so any new detections with Gaia may take some time. So far, Gaia has found one exoplanet.

To date, only three exoplanets have been found using this method: [DENIS-P J082303.1-491201 b](#), [GJ 896 A b](#) and [HIP 66074 b](#). The first two exoplanets are gas giants that were found using ground based instruments, while the third exoplanet is the exoplanet found with Gaia.

How It Works

The basic idea behind astrometry for exoplanet detection is depicted in the animation below. Astronomers obtain a series of pictures over time (a “time series”) and use at least three reference stars (which have nearly constant positions) to measure the changing position of the host star in the plane of the sky.

Astrometry

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=167>

Animation source: <https://exoplanets.nasa.gov/>

Side Note: How is astrometry (or the “astrometric technique”) different from Direct Imaging or from the Doppler method?

With direct imaging, astronomers obtain an actual image of the planet. To confirm that the planet is not a background source, additional observations are taken over time, verifying orbital motion of the planet candidate. But the big difference between direct imaging and astrometry is that with direct imaging the planet is observed.

Like the Doppler (or “radial velocity”) technique, astrometry is an “indirect” method – the planet is not seen, but its existence is inferred by the gravitation effect that the planet exerts on the host star. However, different things are being measured with Doppler and Astrometric techniques; the radial velocity technique measures the changing speed of the star along one dimension: the line of sight. The fundamental data are measurements of radial velocity over time. Astrometry measures the changing position (not speed) of the star projected onto the two-dimensional plane of the sky. The fundamental data are positions (angles and angular separations) over time.

Because astrometric measurements are made in two dimensions (x, y in the plane of the sky), the astrometric orbital model measures the true mass of the planet; it does not suffer from the unresolved inclination of orbital models from the radial velocity method.

It is straightforward to show that distance between the star and planet from the center-of-mass (a) and their masses are related by:

$$M_* d_* = m_p d_p \quad (5.3.1)$$

where

M_* is the mass of the star

m_p is the mass of the planet

d_* is the distance of the star from the center-of-mass

d_p is the distance of the planet from the center-of-mass

Concept Check: Calculate the Astrometric Displacement of the Sun

1. The mass of Jupiter is 0.001 times the mass of the Sun and Jupiter orbits at a distance of about 5 AU from the Sun. What is the astrometric displacement (in AU) of the Sun because of Jupiter?
2. The mass of Saturn is 1/3 the mass of Jupiter and Saturn orbits at a distance of about 10 AU from the Sun. What is the astrometric displacement (in AU) of the Sun because of Saturn?
3. If Saturn (at 10AU) had the same mass as Jupiter, would the astrometric displacement of the Sun be smaller or larger than the displacement from the real Jupiter at 5AU?
4. Can you picture the total astrometric orbit of the Sun that is caused by both Jupiter and Saturn? Try to make a sketch.

Need help? For part 3, consider how the center-of-mass point changes. A good resource for this is: <https://astro.unl.edu/naap/esp/centerofmass.html>, where you can quantitatively investigate the center-of mass.

Notice that even if you know the mass of the star (from its spectral type) and the distance between the star and the center of mass with astrometry, there are still two unknowns in the equation above: the mass of the planet and distance between the planet and the center of mass. However, astrometric observations also reveal the orbital period, so we can use Kepler’s Laws to derive the semi-major axis of the planet (the semi-major axis a is the same quantity as the distance between the planet and the center of mass). Then, if you know the mass of the star and measure the astrometric displacement of the star, you can solve for the planet mass.

Astrometry of Multi-Planet Systems

Consider our solar system. The astrometric motion of the Sun is depicted in Figure 20 below. These positional measurements assume a bird’s eye view looking down onto the orbital plane of the solar system. The larger lighter yellow circle indicates the size

of the Sun, so you would need to wait long enough to see motion that is larger than this size.

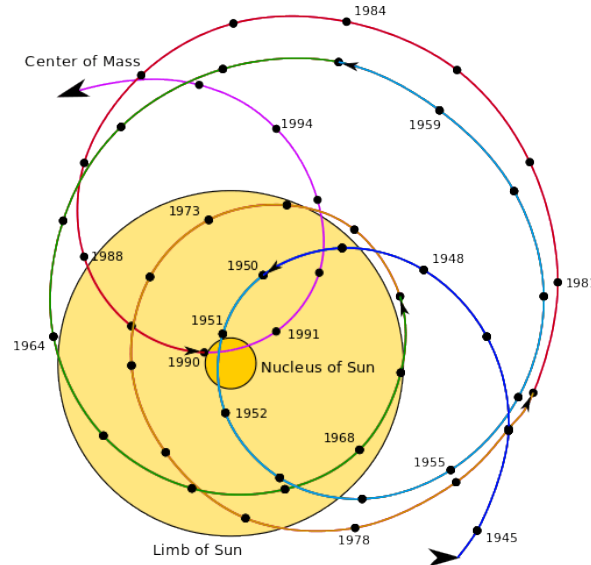


Figure 20 – The astrometric motion of the Sun from the gravitational tug of the orbiting planets from 1945 – 1997. **What is the approximate period of the largest loops of orbital motion for the Sun? What planet in our solar system has a similar period? Why doesn't the astrometric orbit of the Sun look like a single closed circle or ellipse?**

The star wobble demo below will allow you to add in any planets in our solar system and see the gravitational effect on the Sun that might be observed if we were observing from above the orbital plane. However (this is an important point), to fully model an astrometric orbit, astronomers must observe the star through a large fraction of one complete orbit. This would require about 12 years of observations to detect Jupiter or 30 years for Saturn!

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Concept Check: Astrometric Displacement of the Sun – a Bird's Eye view

Select **only** Jupiter and run the simulation. Then run the simulation selecting **only** Saturn. Does this match the result you expected from parts 1 and 2 of the first Astrometry question above?

Now select **both** Jupiter and Saturn. Does this result match your picture from part d of the first Astrometry question above?

How do the astrometric influences of Mercury, Venus, Earth, Mars, and Pluto compare to the influence of Jupiter, Saturn, Uranus, and Neptune.

Gravitational Microlensing

Gravitational Lensing Basics

Stars emit light in all directions. When we observe a distant star we see only the few rays of light from the star that are directed straight toward us, as shown in Figure 21. However, if a foreground star passes very close to our line of sight of the more distant source star, the light from the distant background star will be gravitationally focused. To understand this idea, remember that we are the observer. The distant star is the source. The intervening star is the lens, and it bends light from the source toward us. Figure 22 below show the light path from a distant source star with an intervening "lens" star. How does the path of light from the source star differ in Figures 21 and 22?

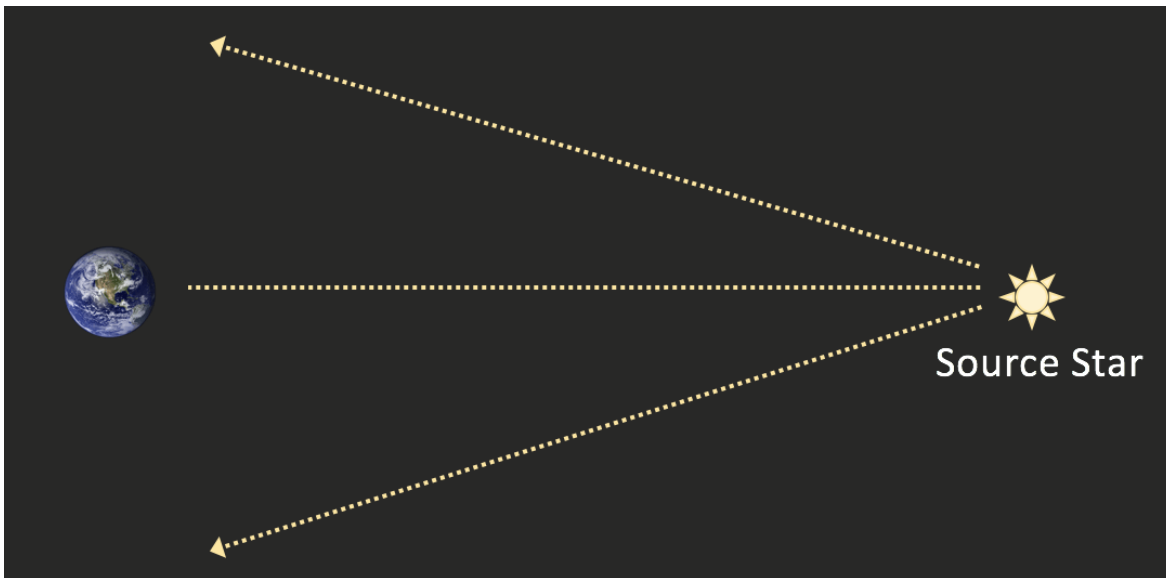


Figure 21 – A background source star emits light in all directions. **From Earth, can we see the top and bottom diverging rays of light from the source star?**

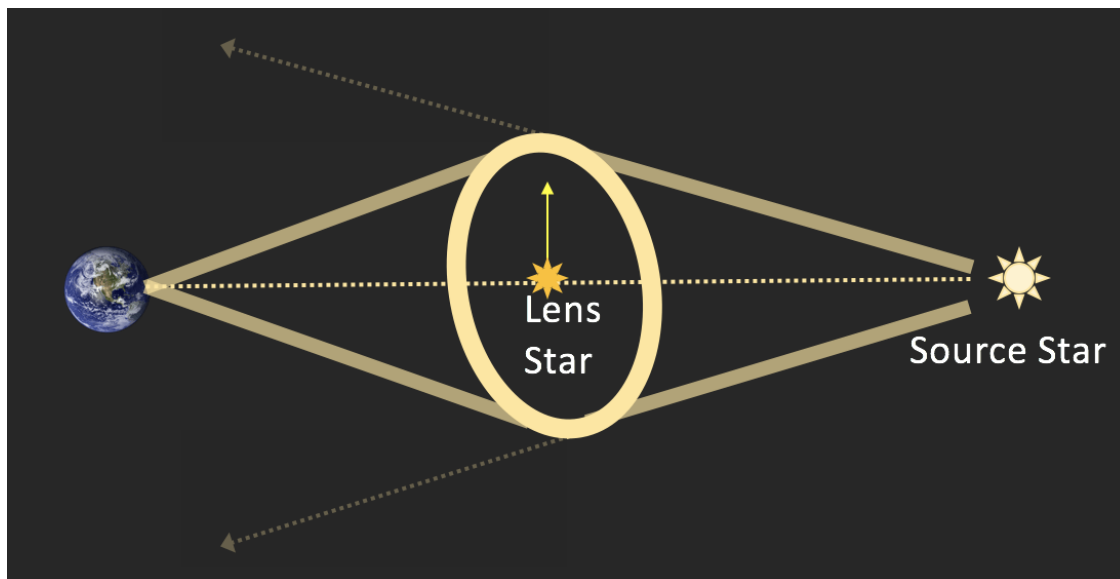


Figure 22 – When another star passes between our line of site to the source star, the gravity of that star bends spacetime and directs light from a gravitational “Einstein ring” toward the observer on Earth. Now, can we see the all of diverging rays of light from the source that pass through that ring. **What effect will that have on the brightness of the source star?**

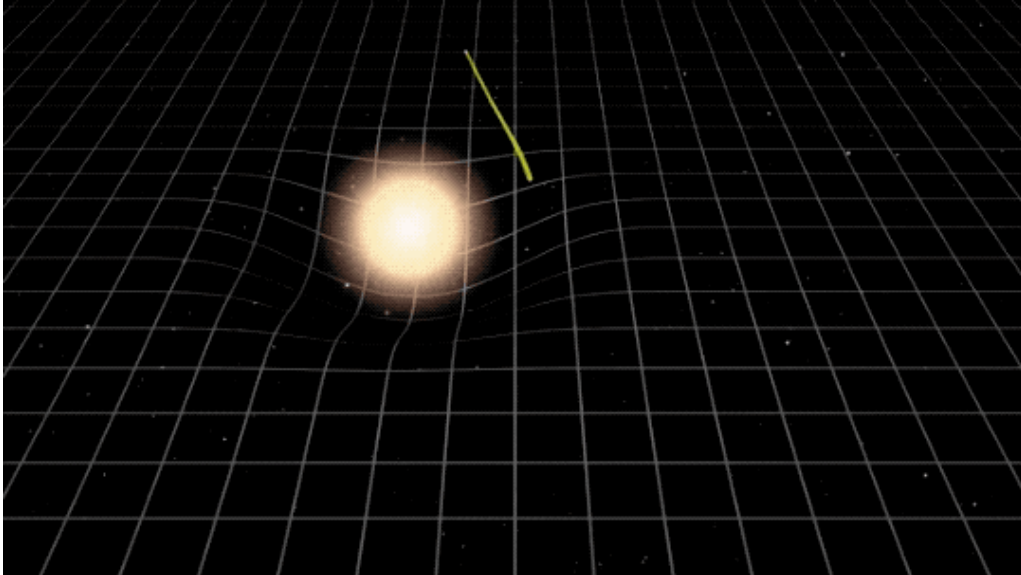
The bending and focusing of the light from the source star in Figure 22 is caused by the gravity of the lens star. To fully understand why light is bent, gravity needs to be considered as described by Albert Einstein. In 1915, Einstein published his General Theory of Relativity, treating gravity as a warping of the space-time continuum and an attribute of mass rather than a fundamental force. Einstein’s description of gravity employs mathematically sophisticated metric tensors, while Newtonian gravity uses more commonly accessible algebra or calculus. On the scales that humans are accustomed to working with, general relativity converges to the Newtonian description, so we are justified in defaulting to the Newtonian description of gravity. More details are given in the box below.

Want to know more: General Relativity and Space-Time

Sir Isaac Newton considered space and time to be completely independent, and that continued to be the accepted view until the beginning of the twentieth century. But Albert Einstein showed that there is an intimate connection between space and time, and that only by considering the two together—in what we call **spacetime**—can we build up a correct picture of the physical world. We examine spacetime a bit more closely in the next subsection.

The gist of Einstein's Theory of General Relativity is that the presence of matter curves or warps the fabric of spacetime. This curving of spacetime is identified with gravity. When something else—a beam of light, an electron, or the starship *Enterprise*—enters such a region of distorted spacetime, its path will be different from what it would have been in the absence of the matter. As American physicist John Wheeler summarized it: "Matter tells spacetime how to curve; spacetime tells matter how to move."

Spacetime is four-dimensional: three space dimensions and one time dimension (the time dimension is also warped but will do not discuss that further here). The warping of the space dimensions is sometimes shown as a 2-D grid, as in the part of the animation below that shows light bent around a single star. The light is following the shortest path it can take as it encounters the warped space.



Animation: As light from a far away source passes by a massive object, like a star or cluster, it is bent as it moves through the warped spacetime around the massive object. The warped spacetime deflects the light and causes it to converge at a point.

The amount of distortion in spacetime depends on the mass of material that is involved and on how concentrated and compact it is. Terrestrial objects, such as people and buildings, have far too little mass to introduce any significant distortion.

Every object with mass has a gravitational field. We can think of objects with spherical symmetry, like stars or planets, as point sources. The strength of the gravitational field for a star is spherically symmetric. The strength of the gravitational field for a star is spherically symmetric. When we consider gravitational lensing, the relevant geometry is a slice through the sphere – a disk perpendicular to our line of sight.

Qualitatively, the gravitational field at small radii on this disk (i.e., closest to the lensing object) is strongest, and light from the source is bent so strongly that it converges closer to the lensing object and does not reach the observer. Likewise, the gravitational field is weakest at large radii on the disk, and light from the source star focuses at distances beyond the observer. However, there is a ring on the disk where the bending of light from the source is "just right." Those rays of light from the source are brought to a focus at the position of the observer. This effect is called gravitational lensing, and it was predicted in 1936 by Einstein. The ring of light that is focused at the position of the observer is given the special name of an **Einstein ring**.

The size of the Einstein ring depends on the mass of the lens star, the distance between the observer and the lens, D_L , and the distance from the observer to the source, D_S . Gravitational lensing occurs even if the lens star is too faint for us to see. Since most of the stars in galaxy are low mass, low luminosity M dwarf stars, this is often the case!

Since lensing is an effect of gravity, it might seem like the amplitude of brightening will depend on the mass of the lens. While the mass of the lensing star determines the radius of the Einstein ring and the duration of the lensing event, the amplitude of brightening is most sensitive to something called the impact parameter. When the lensing object precisely lines up with our view to the source star, the impact parameter is zero and the magnification in brightness is a maximum. If the lensing object is not as tightly aligned with our view of the source – perhaps it just clips our view of the source star, then the impact parameter is larger and the brightening of the source star (the "magnification") is not as strong. This is depicted in Figure 23 below. The Einstein ring is represented by the dashed circle. The red, yellow, green, and blue lines show the path that the lensing object travels with

different offsets from precise alignment with our view of the source. The closer the alignment of the Einstein ring of the lens to our view of the source, the stronger the magnification that we observe. Gravitational lensing from objects as massive as stars typically lasts for several days.

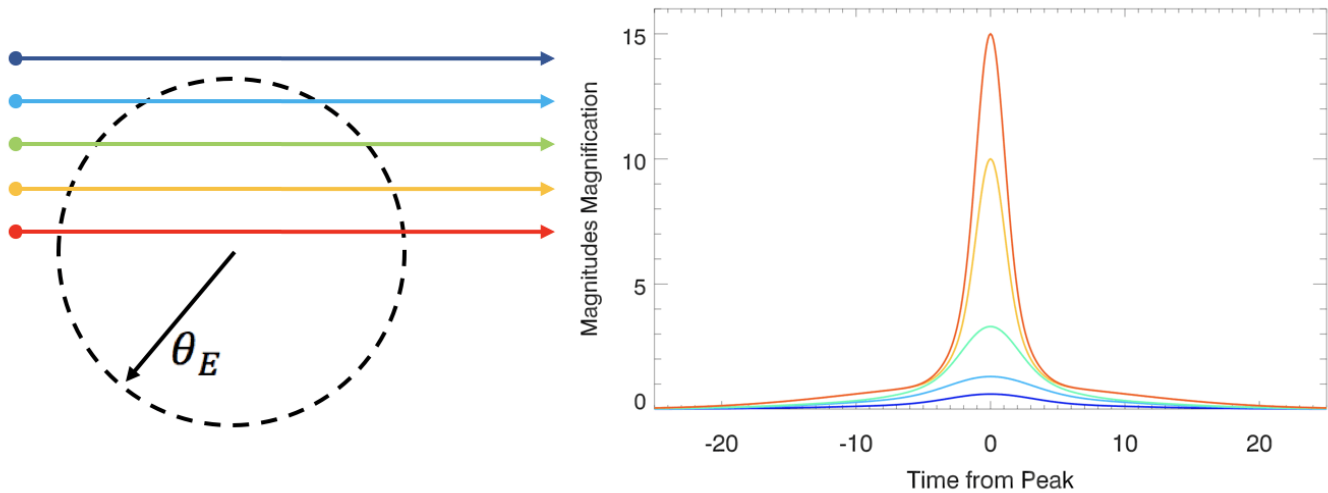


Figure 23 – The impact parameter defines the alignment of the observer, lens and source. The path of the lens star is represented by a straight-line vector. **Which line is closest to our view of the source star?** When the lens is very closely aligned with our line of site to the source, the lens has a small impact parameter and the amplification (red line) is a maximum. When the lens is not exactly aligned with our line of site to the source, the Einstein ring will not be centered on line of site–this is equivalent to having a larger impact parameter (blue line) and the source light magnification is weaker.

Microlensing

Gravitational lensing generally refers to the bending of light from a background source by a foreground massive object. Gravitational **microlensing** is a special case where multiple lensing events occur from more than one foreground mass. If the lensing star happens to have orbiting planets that also cross our line of sight to the source star, then a second microlensing event is superimposed during the brightening event. Because the mass of the orbiting planet is smaller than the mass of a star, the duration of the second microlensing event will be shorter, but since the alignment of the lensing planet is essentially the same as the alignment of the host star, the amplitude of the magnification will be similar.

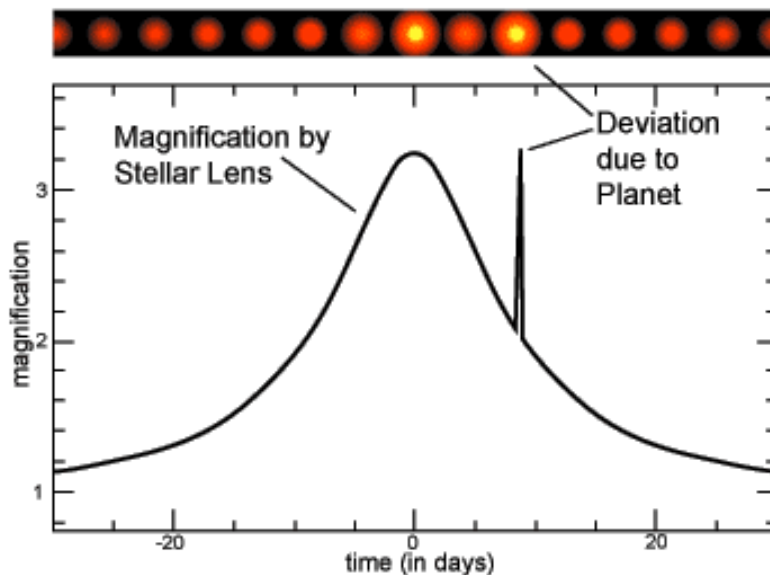


Figure 24 – The photometry of the source star is shown in the top row. Two brightenings occur because the lensing mass is a star with an orbiting planet. The time series light curve is shown in detail in the larger panel with a longer duration brightening caused by the foreground star and a short duration brightening caused by the planet.

The video below summarizes how microlensing searches are carried out at observatories today this clip highlights the [Las Cumbres Observatory Global Telescope Network](#)).

Detecting exoplanets with microlensing.

Video source: [Las Cumbres Observatory](#) CC BY-NC 2.0

Key Concepts and Summary

In this chapter, we reviewed three more methods for detecting exoplanets: transit photometry, which records a dimming in brightness as the orbiting planet crosses our line of sight to the host star; astrometry, where the orbiting planet changes the location of a host star in the plane of the sky; microlensing, where planets (often orbiting a star) pass in front of a background host star, bending spacetime to brighten the background source. Each of these observational methods is sensitive to different masses and orbital distances of planets.

Review Questions

Summary Questions

1. How is the transit method used to detect exoplanets? Describe what a light curve is.
2. How does the dip in brightness caused by a transiting planet vary with the size (radius) of the planet? What if there are multiple planets transiting a star at the same time?
3. Why don't each of the eight planets in our solar system induce identical astrometric movements of the Sun? Which planet causes the largest "wobble" of the Sun? Why?
4. What is gravitational lensing and microlensing?
5. How can gravitational microlensing be used to detect exoplanets?

Exercises

1. Consider the three methods for detecting exoplanets discussed in this chapter and make a chart listing the advantages of each method and what type of exoplanet it is best at finding.
2. Transit simulator activities

This page titled [5.3: Transit, Astrometry and Gravitational Microlensing Techniques](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

5.4: Exoplanets- Statistics and Modern Discoveries

The existence of planets around other stars is now commonplace — over 5,000 exoplanets have been detected in 2023 compared to less than 50 known exoplanets in 2000. That's a 100-fold increase in just over 20 years! The solar systems that have been found look very different from our own. Some have gas giants orbiting very close to their host stars and others have planets with sizes that are not seen amongst our eight planets. Let's take a look at the diversity of planets and planetary systems that have been found and highlight a few of the most interesting, with respect to the possibility of finding life.

Learning Objectives

By the end of this chapter, you will be able to:

- Discuss the ranges for mass and radius that the exoplanets discovered have
- Describe the types of exoplanets being detected and how they compare to the planets in our solar system
- Explain how different detection techniques are sensitive to detecting different kinds of exoplanets.
- Discuss the TRAPPIST-1 planetary system and how it is similar to and different from our own

Exoplanet Statistics

Now that a large sample of exoplanets have been detected, we can take stock of the types of planets that have been found. Overall, the transit method and radial velocity methods have yielded the most exoplanets to date, with 75% of known exoplanets found with the transit method and 19% with the radial velocity method. The animation below summarizes the history of planetary discovery. Note the explosion of detections with the transit method in 2014 when data from the Kepler mission started coming in.

Timeline for exoplanet discoveries.

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=174>

Figures 1 and 2 below summarize the detections by the plotting the planet's mass (Fig. 1) and the planet's radius (Fig. 2) against how long it takes the planet to orbit its host star (its period). Each planet has a symbol to indicate which detection method was used to discover the exoplanet. There are two methods included in Figures 1 and 2 which we did not discuss in the previous chapter, as neither are one of the primary methods used: timing variations and orbital brightness modulations. We'll explain these two methods briefly.

When and for how long a planetary transit will occur is predictable. If there is another unseen planet in the system that is gravitationally tugging on the transiting exoplanet, this can cause the transiting planet to either speed up or slow down and it will cross in front of the star sooner or later than predicted. This is the theory behind **timing variations**. This method has the advantage of allowing you to determine the mass of the exoplanets. A couple dozen of the more than 5,000 exoplanets detected to date were found via timing variations. In our own solar system, the planet Neptune was discovered via timing variations in 1846 — Uranus was found to orbit slower or faster during its orbit around the Sun than predicted, and the presence of another planet was hypothesized to be the reason. Neptune was found exactly where it would be if it was the cause for the changes in Uranus' orbital motion. Science!

A massive planet orbiting very close to its host star can cause the brightness of the star to change due changes in the amount of light reflecting off the planet's surface. Of course, the planet needs to have a reflective surface for this effect to be measured, and these changes in reflected light will manifest as a change in the brightness of the star that is in lockstep with the orbital period (or phases) of the close-in planet. These types of changes are called **orbital brightness modulations**. Changes in brightness can also be caused by distortions to the star's shape caused by the massive, close-orbiting planet.

There are a few important limitations to Figures 1 and 2 that we need to note. Figure 1 shows the mass and orbital period of the exoplanets. The mass of an exoplanet can be found with the radial velocity (RV) method and the timing variations method. The transit method allows you to determine the radius of an exoplanet but not the mass. Not all exoplanets found with the transit

method have been followed up with the RV method to get the mass, so not all planets found with the transit method are included in Figure 1.

In a similar way, not all exoplanets detected with the RV method are shown in Figure 2, and this is clear to see as Figure 2 is dominated by green squares for the transit method. Of the 1000+ stars found with the RV method, only the 66 that have had their radii determined are included in Figure 2.

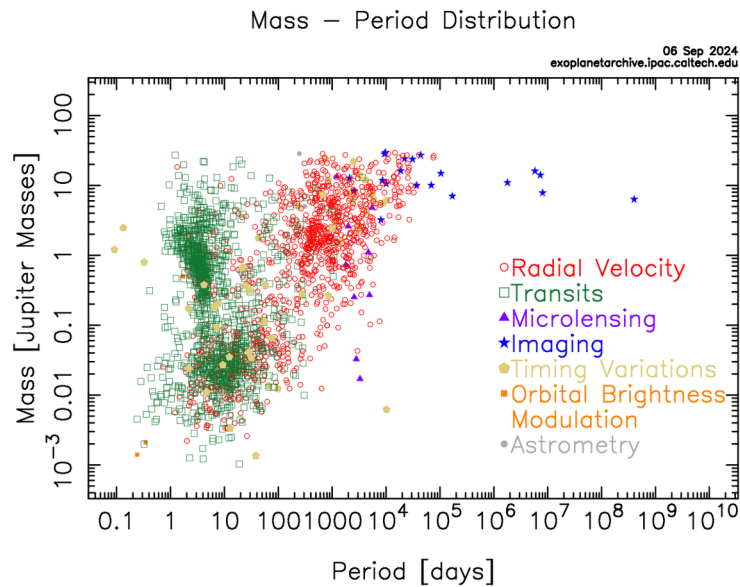


Figure 1 – The mass and period for the exoplanets detected to date, with the colored symbols corresponding to the different detection methods. Note that the masses are relative to the mass of Jupiter. The Earth's mass and period are indicated for reference. Click [here](#) to get most recent plot from the NASA exoplanet archive.

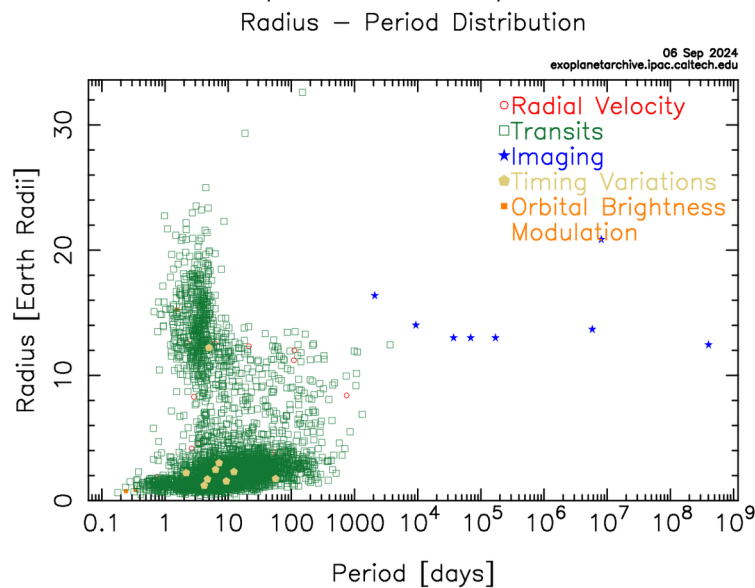


Figure 2 – The radius and period for the exoplanets detected to date, with the colored symbols corresponding to the different detection methods. Note that the radii are relative to the radius of the Earth. The Earth's radius and period are indicated for reference. Click [here](#) to get most recent plot from the NASA exoplanet archive.

Concept Check: Analyzing the Mass-Period and Radius-Period Plots

There are trends in the mass-period plot in Figure 1 that are especially prevalent when looking at the distribution for each detection method.

Transit Method

What trends do you see in radius and period for the exoplanets found using the transit method?

Looking at the radius-period plot (Figure 2), the transit method overwhelmingly finds planets with periods less than a hundred days. There are very few exoplanets found with periods greater than the Earth (365 days) — the limit is about 1,000 days which is a little less than 3 Earth years.

The transit method picks up planets with a wide range of radii, from planets smaller than the Earth all the way up to planets 30 times larger than the Earth. However, intuitively, planets with very small radii will be the most difficult to detect, as these cause the least amount of dimming in the brightness of the host star.

Summary: Most radii, short periods (a few hours up to a few years).

What types of exoplanets will the transit method preferentially detect?

The transit method is best at finding exoplanets that are fairly close to their host stars. This is not a technical fault of the method itself but rather the fact that to collect data for planets on long period orbits, you need measurements spanning decades or centuries, and we've only been observing transits for about 20 years so far.

Radial Velocity Method

What trends do you see in mass and period for the exoplanets found using the radial velocity (RV) method?

The RV method is able to find exoplanets with a wide range of orbital periods, from less than 1 day up to 10^5 days (275 years). The method is sensitive to a wide range of masses, from planets slightly less-massive than the Earth up to planets with 30 times the mass of Jupiter; there are more higher mass planets found than lower mass planets. As for the period range, there is a generally diagonal trend, where low-mass planets with shorter orbital periods and high-mass planets with longer orbital periods are found. There is some scatter whereby higher-mass planets with shorter orbital periods are found, but virtually no planets with lower masses and periods greater than 1 year are detected.

What types of exoplanets will the RV method preferentially detect? What are the limitations?

The RV method is best at finding massive planets with periods out to about 50 years. The method can also find lower mass planets orbiting close to their host stars but cannot detect low-mass planets orbiting further away.

Direct Imaging

What trends do you see in mass and period for the exoplanets found using the imaging method? How about trends in the radius and period?

What types of exoplanets will the direct imaging method preferentially detect?

Gravitational Microlensing

What trends do you see in mass and period for the exoplanets found using the microlensing method? Why do you think the radius cannot be determined for the microlensing method (that is, why is this method not shown in the radius-period plot)?

What types of exoplanets will gravitational microlensing preferentially detect?

In Figure 2, there is a gap for planets with radii of 4 to 10 times the radius of the Earth ($4 R_{\text{Earth}} < R < 10 R_{\text{Earth}}$), especially for planets that are orbiting close to their host stars. There are Jupiter-sized planets found orbiting close to their stars (hot Jupiters) as well as Earth-sized planets orbiting close — where are the planets with radii $4\text{--}10 R_{\text{Earth}}$? Is this gap due to limitations of the detection methods or is this telling us something more general about the types of planets that form around stars? Since the transit method can detect stars below and above this range of radii, this suggests the gap (sometimes called the “hot Neptune desert”) is real and that it has something to do either with the formation or longevity of planets this size. One idea is that planets of this size are most susceptible to having their atmospheres completely evaporate away, making them planets that can transition in size and eventually have a smaller radius (where planets are found).

The types of planets being found are commonly divided into four main categories: gas giants, Neptune-like, super-Earths, and terrestrial. Figure 3 summarizes the fraction of each of these types that has been found. Super-Earths have radii between 1.2 and 2.8 that of Earth, and the other large group with sizes between 2.8 and 4 that of Earth are often called Neptune-like or mini-Neptunes. While the totals are split into about one-third each for gas giants, super-Earths, and Neptune-like exoplanets, we need to keep in mind that the transit method, which has found 75% of all exoplanets, is biased against very small, terrestrial-size planets. The smallest radius for an exoplanet detected with the transit method is about 30% the size of Earth's radius ($R = 0.30 R_{\text{Earth}}$), so we cannot immediately conclude that there are fewer terrestrial exoplanets out there compared to the other types. In our own solar system, two of the eight planets, Mercury and Mars, have radii of $0.39 R_{\text{Earth}}$ and $0.53 R_{\text{Earth}}$.

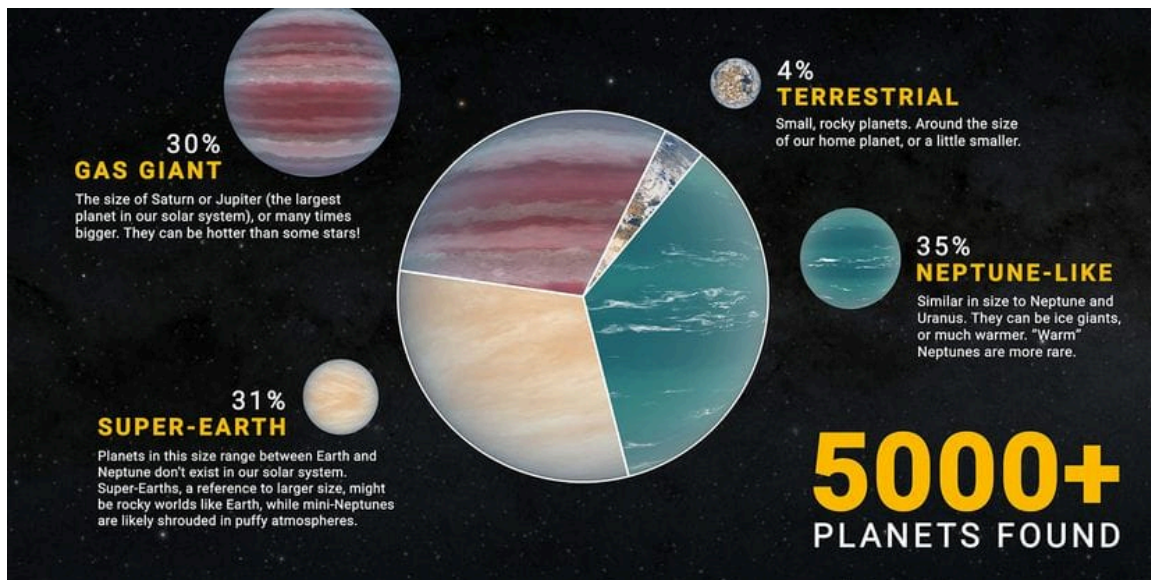


Figure 3 – Breakdown of exoplanets detected to date by type.

For a clearer look into the distribution of exoplanets by their size, Figure 4 shows a histogram of the radii for known transiting exoplanets that have a measured radius. This does not eliminate the bias against terrestrial planets but some interesting trends emerge.

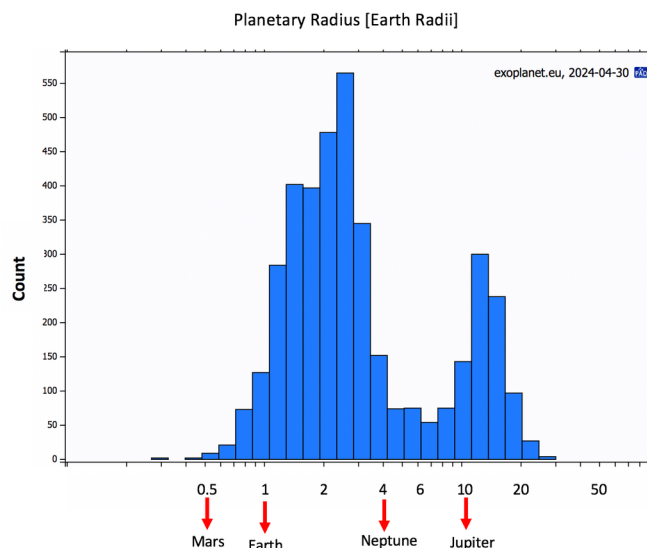


Figure 4 – The distribution of radii for exoplanets.

First, we see a preponderance of exoplanets with sizes intermediate between the Earth and Neptune — the super-Earths and mini-Neptunes. No planets of these types exist in our solar system! What a remarkable discovery it is that the most common types of planets in the Galaxy are completely absent from our solar system and were unknown until the Kepler mission. Further, this idea that planets of roughly Earth's size are so numerous is surely one of the most important discoveries of modern astronomy.

The TRAPPIST-1 system

One of the most intriguing planetary systems detected so far is TRAPPIST-1. The TRAPPIST-1 planetary system is about 40 light years away from Earth and has seven known exoplanets. The TRAPPIST-South survey program came online in 2010 by European astronomers using a small, ground-based telescope at La Silla Observatory in Chile to search for exoplanets with the transit method. The acronym TRAPPIST stands for “Transiting Planets and Planetesimals Small Telescope,” and is a nod to the popular Trappist beers found in Belgium.

In 2015, three planets were detected with this survey around a star that the team renamed TRAPPIST-1 (catchier than the original catalog name for this star, 2MASS J23062928-0502285). TRAPPIST-1 is a small, cool, red dwarf; the advantage of looking at small stars is that the transit signals from smaller, terrestrial planets are stronger. In 2017, four more planets were discovered around TRAPPIST-1, with the transit timing variations, bringing the total to seven planets orbiting TRAPPIST-1. All of these planets orbit at a distance that is smaller than the distance from the Sun to Mercury in our solar system and are similar in size to the rocky terrestrial worlds. Figure 5 below summarizes the properties of the seven known planets orbiting TRAPPIST-1 compared to our solar system’s four terrestrial planets.

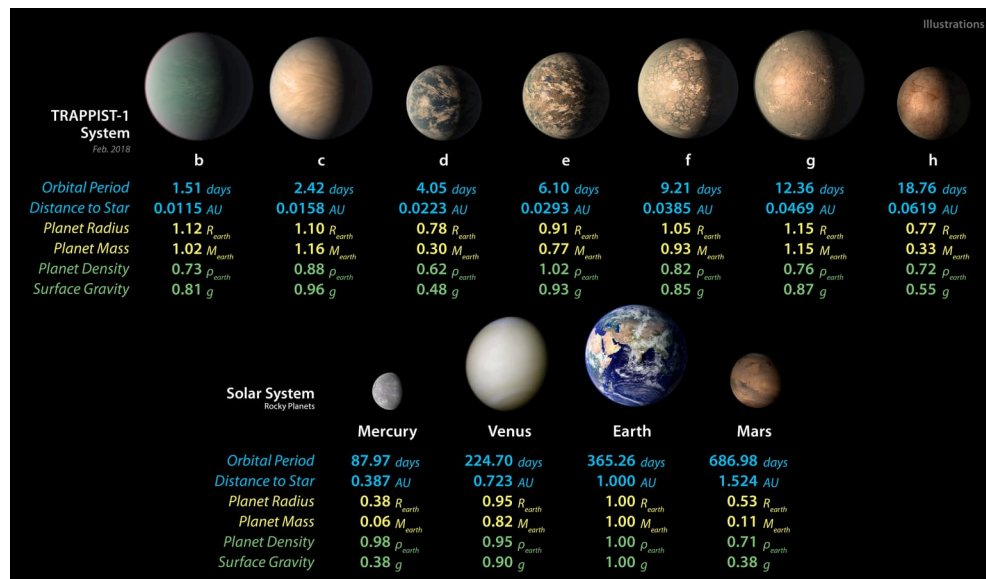


Figure 5 – The TRAPPIST-1 planetary system compared to terrestrial planets in the solar system.

Worked Example: Finding an Exoplanet’s Density

If both the mass and radius are known for an exoplanet, we can find its density. We explored this quantitatively in the chapter on the Earth but revisit it here with an emphasis on finding densities for exoplanets.

Notice that the masses and radii for the TRAPPIST-1 planets are reported relative to the Earth. For example, TRAPPIST-1 b has a radius that is 12% larger than the Earth’s radius ($R_{\text{T1b}} = 1.12 R_{\text{Earth}}$) and a mass almost identical to the Earth’s ($M_{\text{T1b}} = 1.02 M_{\text{Earth}}$). Since these values are relative to the Earth, we can immediately estimate TRAPPIST-1 b’s density:

$$\rho = M/V = M/\left(\frac{4}{3}\pi R^3\right) \quad (5.4.1)$$

Now, by taking the ratio of the density of any exoplanet to that of the Earth’s density, we can find the density of the exoplanet relative to Earth:

$$\rho/\rho_E = (M/\frac{4}{3}\pi R^3)/(M_E/\frac{4}{3}\pi R_E^3) = (M/M_E)/(R/R_E)^3 \quad (5.4.2)$$

Putting in the values of the mass and radius for TRAPPIST-1 b, we find:

$$\rho_{T1b} = (1.02 M_{\text{Earth}}/M_{\text{Earth}})/(1.12 R_{\text{Earth}}/R_{\text{Earth}})^3 \rho_{\text{Earth}} = (1.02)/(1.12)^3 \rho_{\text{Earth}} = 0.73 \rho_{\text{Earth}}$$

This agrees perfectly with the value reported in Figure 5.

Question 1

The exoplanet 51 Peg b has a radius that is 1.27 times greater than Jupiter's radius ($R_{51P} = 1.27 R_J$) and a mass that is 0.46 (46%) the mass of Jupiter ($M_{51Pb} = 0.46 M_J$). Find the density of 51 Peg b relative to Jupiter's density.

Show Answer

This can be solved exactly the same way as the example above (the only difference is that the mass and radius are now reported relative to Jupiter, rather than Earth).

$$\rho/\rho_J = (M/\frac{4}{3}\pi R^3)/(M_J/\frac{4}{3}\pi R_J^3) = (M/M_J)/(R/R_J)^3 \quad (5.4.3)$$

Putting in the values of the mass and radius for 51 Peg b, we find:

$$\rho_{51Pb} = (0.46 M_{Jup}/M_{Jup})/(1.27 R_{Jup}/R_{Jup})^3 \rho_{Jup} = (0.46)/(1.27)^3 \rho_{Jup} = 0.22 \rho_{Jup}$$

This tells us that 51 Peg b is a very light planet with a density that is just 22% the density of Jupiter. You can look up the density of Jupiter ($\rho_{Jup} = 1.3 \text{ g/cm}^3$) and calculate the value for the density of 51 Peg b:

$$\rho_{51Pb} = 0.22 \rho_{Jup} = (0.22) \times (1.3 \text{ g/cm}^3) = 0.29 \text{ g/cm}^3$$

Question 2

The exoplanet Kepler 22 b has a mass that is 9.1 times greater than Earth ($M_{K22b} = 9.1 M_E$) and a radius that is 0.19 (19%) the mass of Jupiter ($R_{K22b} = 0.19 R_J$). Find the density of 51 Peg b relative to Earth's density.

Show Answer

For this problem, we are given the radius relative to Jupiter and the mass relative to Earth. We will need to convert the radius so that it is relative to Earth as well, so that a direct comparison can be made to the Earth's density.

Jupiter's radius is 11.2 times the radius of Earth: $R_J = 11.2 R_E$. Using this conversion factor:

$$R_{22Kb} = 0.19 R_J = 0.19 \times (11.2 R_E) = 2.1 R_E$$

The density relative to Earth is:

$$\rho_{K22b} = (9.1 M_{\text{Earth}}/M_{\text{Earth}})/(2.1 R_{\text{Earth}}/R_{\text{Earth}})^3 \rho_{\text{Earth}} = (9.1)/(2.1)^3 \rho_{\text{Earth}} = 0.98 \rho_{\text{Earth}}$$

Kepler 22 b has almost the same density as Earth so Kepler 22 b is also a rocky planet, with a density slightly less than 5.5 g/cm^3 .

Three of the seven planets orbiting TRAPPIST-1 are in the habitable zone — the distance from TRAPPIST-1 where the temperature is just right for liquid water. If there are three habitable worlds around the TRAPPIST-1 star and if technological civilizations evolve on any of these planets, there will be a powerful incentive to build spaceships that travel to these neighboring worlds. All of the TRAPPIST-1 detected to date are either terrestrial or super-Earths. We know this based on both the mass and radii of the TRAPPIST-1 planets, which together allow us to find their densities. In Figure 5, we see that most of the TRAPPIST-1 planets have densities close to Earth, with TRAPPIST-1 d having a lower density at 62% the density of Earth ($\rho_{T1d} = 0.62 \rho_{\text{Earth}}$).

Figure 6 – The TRAPPIST-1 system compared to our solar system. Notice that the TRAPPIST-1 system is zoomed in 25 times to be on the same scale as our own; the entire TRAPPIST-1 system fits inside the orbit of Mercury. Credit:

The orbits of the seven TRAPPIST-1 planets are close to resonance, meaning that the ratio of the orbital periods are close to

integer values. Planet e is in a 6-day orbit and planet g is in a 12-day orbit; planet f is in a 9.2-day orbit and planet h is in an 18.7-day orbit.

TRAPPIST-1 System with JWST

Since its discovery, the TRAPPIST-1 system has been followed up with several different telescopes, including Spitzer, Kepler, Hubble and the James Webb Space Telescope. Hubble collected spectra for all seven exoplanetary transits in the TRAPPIST-1 system to study the compositions of any atmospheres that these planets may hold.

The mid-Infrared Instrument (MIRI) on JWST was used to find the temperature of TRAPPIST-1 b by measuring the brightness of TRAPPIST-1 when planet b was behind the star (this is called a secondary transit or an occultation). The temperature of TRAPPIST-1 b was found to be about 500 K and further analysis showed that it is unlikely that this exoplanet has an atmosphere (read more at <https://exoplanets.nasa.gov/news/1756/webb-measures-the-temperature-of-a-trappist-1-exoplanet/>). The data was fit to models for a planet with an atmosphere and without, and the observations were a better match to the model for a planet that is bare rock (with no atmosphere).

The NIRISS (Near-Infrared Imager and Slitless Spectrograph) instrument aboard JWST was also used to study TRAPPIST-1 b. The NIRISS results confirmed that there is little evidence that TRAPPIST-1 b has an atmosphere — although it still cannot be ruled out — but also found that contamination from the star TRAPPIST-1 can compromise the analysis of spectra^[1]. M dwarfs like TRAPPIST-1 experience frequent flares of X-rays and other surface phenomenon, such as star spots and faculae, both of which are caused by magnetic fields. If a flare occurs during the measurement of the planet's spectrum, this can create “ghost lines” that could be misinterpreted as an actual chemical signature in the spectrum. This study underscores the importance of taking into account the effects of outbursts of radiation from the host star, especially M dwarfs, which are the most plentiful type of star in our Galaxy.

Assessing Exoplanetary Habitability

With some basic properties of an exoplanet — its mass, radius, semi-major axis and orbital period — and its host star, a critical assessment can be made as to whether the exoplanet is habitable. In this section, we look at some of the key questions to ask in carrying out such an assessment.

Density

The density of a planet is important in understanding its composition and surface: this tells us if the world is a gas giant, a rocky planet or something in between. To find the density of an exoplanet, both its radius and mass are required. Some examples are shown above on how to carry out density calculations for exoplanets.

A good resource for finding the mass and radius of an exoplanet, reported relative to Earth or Jupiter, is NASA's Eyes on Exoplanets site: <https://exoplanets.nasa.gov/eyes-on-exoplanets/>

Once you know the approximate density of a planet (we say approximate here as the exact mass is usually not known, but instead the minimum mass found through the Doppler technique), you can then assess the surface properties by comparing to some benchmarks, such as the density of Earth and the density of Jupiter. Figure 7 below shows the densities for the planets in our solar system compared to those of exoplanets detected.

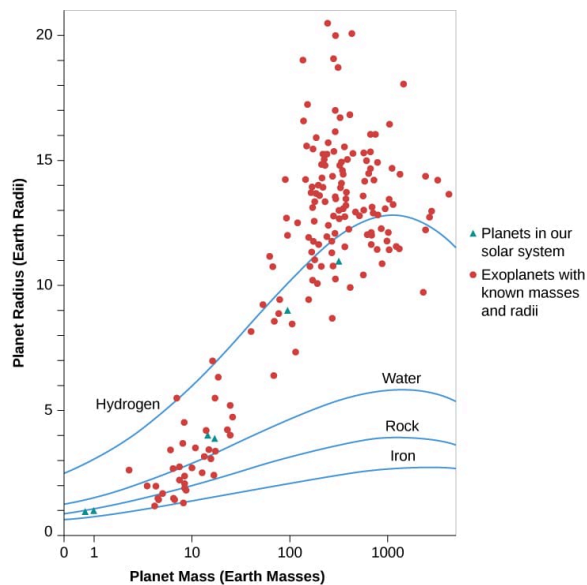


Figure 7 – Exoplanets with Known Densities. Exoplanets with known masses and radii (red circles) are plotted along with solid lines that show the theoretical size of pure iron, rock, water, and hydrogen planets with increasing mass. Masses are given in multiples of Earth's mass. (For comparison, Jupiter contains enough mass to make 320 Earths.) The green triangles indicate planets in our solar system.

Surface Gravity

The surface gravity of a planet can give insight as to whether the world can hold onto an atmosphere. Without an atmosphere, life on any other world seems improbable. On Earth, the atmosphere keeps the surface temperature (relatively) stable and enables the Greenhouse effect; without our atmosphere, Earth would be a frozen ice planet.

The surface gravity of a world depends on its mass and radius as:

$$g = G \frac{M}{R^2} \quad (5.4.4)$$

An example of how to find the surface gravity of a world is shown [here](#) in the chapter on Laws of Motion and Gravity.

How much gravity at the surface is enough to retain an atmosphere? Atmospheric retention depends not only on the surface gravity but also the temperature. Gas molecules move faster in hot environments compared to cold ones.

Habitable Zone (liquid water?)

A key question to ask is whether or not the planet is in the habitable zone around its host star. The distance from the star to the edges of the habitable zone depend entirely on the parameters of the host star (and can be estimated using just the luminosity of the star, as shown in [this example](#) in the Habitable Zone chapter). With the inner and outer boundaries (that is, the distance from the star to the closest part of the HZ and the distance to the furthest part), the distance from the exoplanet to the star (its semi-major axis, a) can be compared to see if it falls within these boundaries. An example is shown in the box below.

Worked Example: Is a planet in the HZ around its star?

Using the luminosity of its host star, estimate whether the rocky exoplanet Kepler 452 b is in the habitable one around its host star.

First, a quick search* reveals that the luminosity of Kepler 452 (the star around which Kepler 452 b orbits) is about 1.2 times the luminosity of the Sun: $L_{K452b} = 1.2 L_{Sun}$

Now, we can use the relationships for the inner/outer bounds of the conservative HZ:

$$d_{inner,star} = 0.95 \sqrt{L_{star}/L_{Sun}} \quad (5.4.5)$$

$$d_{outer,star} = 1.37 \sqrt{L_{star}/L_{Sun}} \quad (5.4.6)$$

$$d_{inner} = 0.95 \sqrt{1.2} \text{ AU} = 1.04 \text{ AU}$$

$$d_{outer} = 1.37 \sqrt{1.2} \text{ AU} = 1.50 \text{ AU}$$

Next, let's compare these bounds with the average distance of Kepler 452 b from Kepler 452:

$$d_a = 1.05 \text{ AU}$$

Kepler 452 d is very close to the inner edge of the HZ around Kepler 452. Yes — it is in the habitable zone.

Note that Kepler 452 has nearly the same luminosity as the Sun and the planet Kepler 452 is orbiting at about the same distance from Kepler 452 as the Earth does from the Sun. This is an intriguing example of an Earth-sized planet orbiting a Sun-like star...could Kepler 452 b have water on its surface?

*For a more precise estimate of the luminosity, look up the star on [NASA's exoplanet archive](https://nasa.gov/exoplanet/archive). You can also estimate a star's luminosity relative to our Sun by using the Stefan-Boltzmann law, if you know its radius and temperature.

For a graphic image, you can use the HZ generator tool at <http://astro.twam.info/hz/>

Try it! Generate the HZ around TRAPPIST-1.

To start, open the HZ sim link and click on the "Multiple" tab. Remove Star B (by clicking on the red X next to it). Now you can input the values for TRAPPIST-1 — note the units! You will need to adjust both the Plot Radius and the Resolution for the figure.

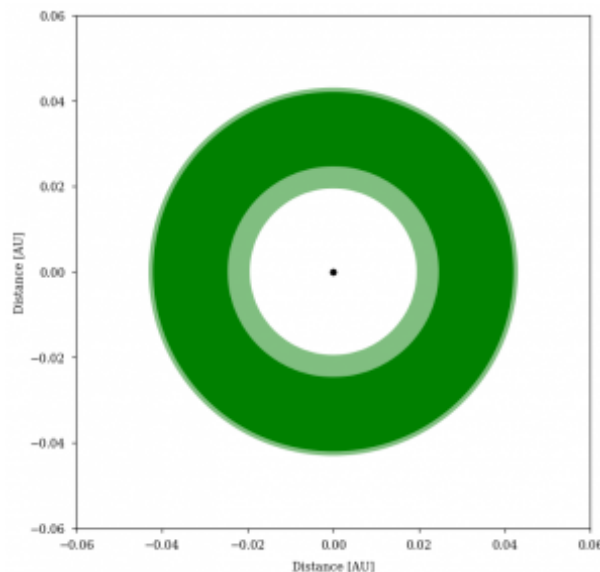


Figure generated at <http://astro.twam.info/hz/>. Citation: Müller & Haghighipour 2014 ApJ 782 26

Go Deeper

Look back at the distances for each of the seven detected in exoplanets in the TRAPPIST-1 system. Which of these planets fall into the HZ around TRAPPIST-1, as shown in the figure above?

[Ans: c, d, e]

Tidal Locking

A world is tidally locked if the same side of it always faces the object that it orbits. The Moon, for example, is tidally locked to the Earth and we always see the same side of the Moon from Earth. It takes the Moon the same amount of time to make one rotation on its axis as it does to make one orbit around the Earth. Pluto and its moon Charon are also tidally locked.

Planets can be tidally locked to their host stars as well. This is a gravitational effect and depends on the mass of the star and the planet's distance to the star. Over time, the rotation (spin) of the planet is gravitationally synchronized with the orbit. Figure 8 below summarizes where the “tidal lock radius” line (to the left of this line a planet is tidally locked) for stars by their spectral type and the distance that a planet is from the star. Our solar system is shown for comparison. Mercury is shown to the left of the tidal lock radius line, and Mercury is in a 3:2 orbital resonance with the Sun (because it also experiences gravitational pulls from Venus, Mercury is not in a 1:1 orbital resonance, which is the case for tidal locking).

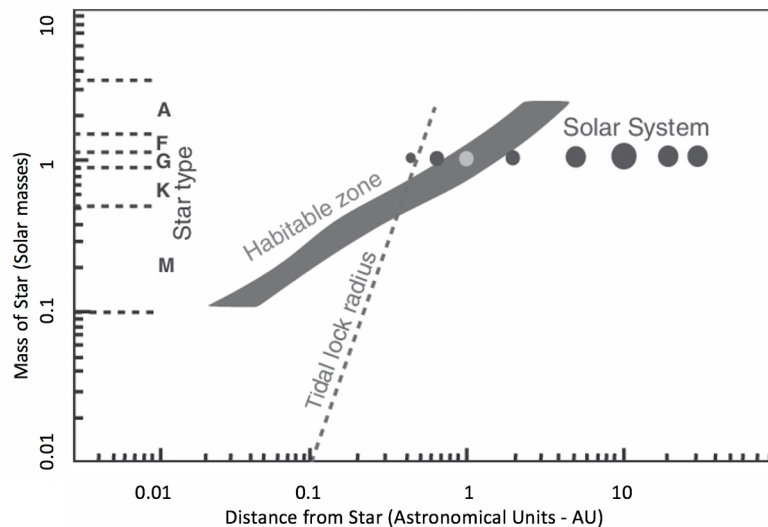


Figure 8 – The Habitable zone distance depends on spectral type of the star. The tidal locking distance is indicated with a dashed line.

Concept Check: Tidally locked?

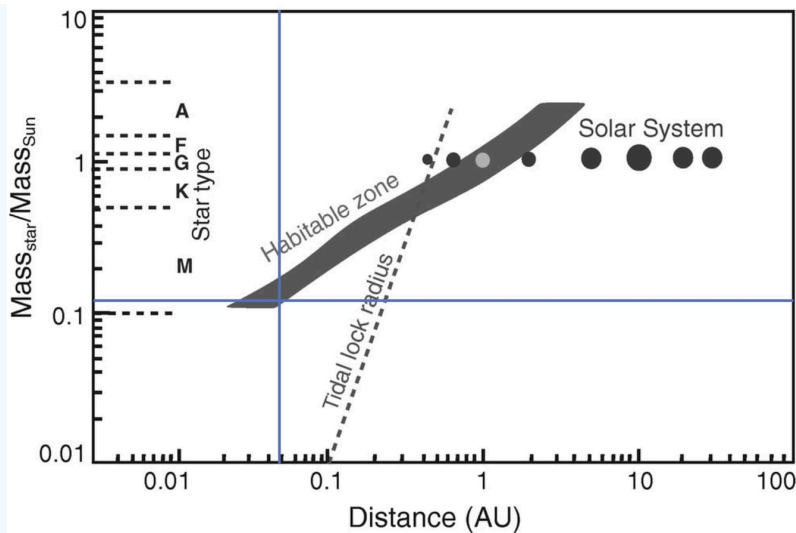
Using Figure 8, determine if the exoplanet Proxima b tidally locked around its host star Proxima Centauri.

To answer this using Figure 8, we need the mass of Proxima Centauri and the distance that Proxima b is from this star; these are ([source](#)):

$$M_{\text{ProxCen}} = 0.12 M_{\text{Sun}}$$

$$a_{\text{Proxb}} = 0.0486 \text{ AU}$$

Finding the intersection of these points on Figure 8, the exoplanet is close enough to be tidally locked. This figure also shows that Proxima b falls into the habitable zone around its star. The marked figure is shown below for clarity.



Key Concepts and Summary

Each of the observational detection methods for exoplanets is sensitive to a different range of masses and orbital separations between the host star and planet. In general, larger or more massive exoplanets around the nearest stars present the largest — and therefore most easily detectable — signals. Over the past decades, significant effort has gone into reducing instrumental errors and improving observing strategies with the goal of pressing measurement precision down to Earth-detecting sensitivities. The ensemble of exoplanets and exoplanet architectures reveal some differences from the solar system, however, there are still biases that prevent us from obtaining a full view of planetary systems around other stars. In the TRAPPIST-1 system, seven orbiting planets have been detected. These TRAPPIST-1 planets orbit close to the low mass host star and appear to be terrestrial, or rocky worlds.

Review Questions

Summary Questions

Exercises

1. **Assessing the habitability of an exoplanet.** Choose an exoplanet that you are interested in learning more about, especially the possibility of it hosting life. Begin by hypothesizing whether or not you think the exoplanet is habitable. Go through the four factors (density, surface gravity, habitable zone estimation, and tidal locking) in the Assessing Exoplanetary Habitability section above and discuss whether the evidence you found supports the hypothesis.

1. read more at <https://phys.org/news/2023-09-james-webb-space-telescope-spectrum.html> 5.04: Exoplanets- Statistics and Modern Discoveries#return-footnote-174-1

This page titled 5.4: Exoplanets- Statistics and Modern Discoveries is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by .

6: SEARCHES FOR TECHNOLOGICAL AND INTELLIGENT LIFE IN THE UNIVERSE

6.1: The Challenges of Space Flight

6.2: The Search for Extraterrestrial Intelligence (SETI)

6.3: Biosignatures and Technosignatures

6.4: The Drake Equation and the Fermi Paradox

6.5: Speculations and Skepticism

This page titled [6: Searches for Technological and Intelligent Life in the Universe](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

6.1: The Challenges of Space Flight

The vast distances between planets and stars mean that any plans for life forms to travel between potentially habitable worlds will meet with the challenges that come with the difficulties of prolonged space flight. In this chapter we explore why sending living things to space is expensive, time-consuming, and dangerous.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain why it is challenging to launch objects from Earth
- Discuss microgravity and why it poses challenges for astronauts
- Discuss why radiation from space is harmful to life
- Discuss why the large distance to other stars poses a challenge in traveling to them
- Describe some ways to deal with the vast distances involved in interstellar travel

Space flight began in 1957 when the Soviet Union launched their Sputnik-1 satellite into orbit. A space race quickly followed with the nominal conclusion being in 1969 when the United States civilian space program run by the National Aeronautics and Space Administration landed a human on the Moon. The excitement about space flight at the time was enough that optimists were soon predicting that by the end of the century, flights to space would be routine and humans might be living and working in space as casually as they live and work in remote locations on Earth. Some very ambitious promoters of the promise of human space flight thought that perhaps even entire colonies of people might be living in Earth orbit, on the Moon, or on Mars within a few decades time.

That this did not happen is a testament to how difficult and dangerous an endeavor space flight is. As a profession, astronauts have a fatality rate that is twice as high as mountain climbers attempting to summit Mount Everest. Nevertheless, the promise and allure of space flight holds such appeal that NASA routinely gets more than 1,000 applicants for every new astronaut position. Trips to space have even become vanity projects for the extremely wealthy. It seems that space flight holds an appeal that captures our human imaginations in profound ways.

But what makes space flight so difficult? What are the possibilities for solving the challenges that keep space so elusive for our human endeavors. This chapter will try to provide some answers to these questions while exploring ideas related to the mechanics of space flight as it pertains to humanity as well as potential (intelligent) life forms in the universe who may be similarly reaching for the stars.

Challenge #1: Getting to Orbit

The biggest barrier to getting to space is the gravitational potential well we live inside. Although a mere 100 km away, in order to reach low-Earth orbit, a rocket needs to accelerate to more than the awesome speed of 8 km/s or about 30,000 km/hour. This is no small feat: even with the most efficient fuels available rockets require fuel ratios of 8:1 meaning that for every 1 kg of payload you want to send to orbit you need 8 kg of fuel.

For all that, because rockets require such tremendous speed and space is relatively nearby (closer than some people's daily commute!), this means that launches to low-Earth orbit are incredibly fast. You can see an example of how little time it takes for a rocket to go from standing on the ground to releasing a payload into orbit below.

Escaping Earth

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=180>

The challenge, of course, is that the payload is sitting on the equivalent of eight times its weight in high-grade explosives. Of course, the explosives are carefully engineered to not all go off at once so that the smooth and controlled rocket flight is obtained

(this is the rocket science of rocket science!), but it is still dangerous enough that most rocket launches require observers to be several kilometers away when viewing in person.

Are there more efficient ways to get to orbit? Some enterprising inventors and designers have proposed such fanciful ideas as using a device called a mass driver which uses electromagnetic forces to accelerate a payload in a sort of souped up version of a maglev monorail. Sometimes combined with an engine called a scramjet which combusts the air at supersonic speeds, such systems would conceivably allow for launch into space without the need for enormous amounts of fuel, though, due to conservation of energy, it would require the same amount of energy although it could perhaps be sourced more sustainably than rocket fuel. These technologies are still in the development phase and it is not clear whether or how they might be practical for launches.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Another proposal which is perhaps even more fanciful but may be maximally efficient is a space elevator. This idea requires building a tower or cable that stretches well beyond low-Earth orbit to some 35,000 km distance in order to be dynamically stable. At the top of the space elevator, a counterweight would have to be fastened which would keep the structure secured through tension through the centrifugal force on the counterweight. Payloads could be sent up the space elevator using a variety of techniques — one more astonishing proposal suggests using lasers to push the payload up the space elevator. Once the payloads reached the top, they would be traveling at orbital speed simply by following the rotation of the Earth (this is so-called “geostationary orbit”). At that point the payload could be released and given a little push away from the space elevator and it would remain in space. The material stress on the space elevator’s material would be so strong, however, that there is currently no known substance that could withstand it.

For the last twenty years, the International Space Station has been serving as a laboratory for testing what it is like to have a permanent human-presence in orbit. Using orbital platforms as staging grounds for further space flight beyond the Earth has often been suggested as a way to assemble larger space ships without having to launch them all at once from the ground. While the ISS has not yet been used for this purpose, there are those who think it could serve such a function for future crewed missions to the Moon or Mars.



Figure 1 – The International Space Station (ISS) photographed by Expedition 56 crew members from a Soyuz spacecraft after undocking. NASA astronauts Andrew Feustel and Ricky Arnold and Roscosmos cosmonaut Oleg Artemyev executed a fly around of the orbiting laboratory to take pictures of the station before returning home after spending 197 days in space. T Along with China’s Tiangong space station, the ISS is the means by which humans are maintaining continuous presence in low-Earth orbit.

Challenge #2: Microgravity

The weight of anything in orbit which is the force felt in response to the force of gravity is zero. This is because when in orbit the force accelerates the orbiting body and is not opposed (this is so-called “free fall”). In a massive space ship, there is a tiny amount of gravitational pull that the ship will exert, but those forces are small enough that NASA calls such a scenario “microgravity” to indicate that the gravity is reduced by factors of millions compared to what is experienced on Earth.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

Microgravity requires certain practical considerations to be taken into account. Astronauts need to alter their eating, drinking, and hygiene habits to account for a lack of gravity. Simple manual tasks become more of a challenge: tools need to be tethered or they drift away, people need to strap in to sleeping sacks or they can hurt themselves or smash into equipment when they're unconscious. While the story of NASA spending one million dollars to develop a pen that would work in microgravity because most pens use gravity to pull the ink from the well is not true (the story goes that when the Soviets were asked how they solved the problem and replied that they used a pencil), it is true that a private company developed a "space pen" in the 1960s that forces the ink out using pressurized gas, these pens can be bought for a few dollars each.

Living in Microgravity

A video element has been excluded from this version of the text. You can watch it online here: <https://pb.libretexts.org/astrobio/?p=180>

The health effects of microgravity have now been extensively studied as astronauts have spent months and years aboard the International Space Station. One key issue is loss of muscle tone and bone density since those are both maintained on Earth by bodyweight resistance that does not exist in microgravity. Thus, astronauts are tasked with exercising for some 2.5 hours each day and even still will experience adverse effects that make it difficult to adjust to the living with their weight upon their return.

Challenge #3: Life Support

Once leaving the confines of Earth, certain requirements of life become scarce. At minimum, we require provisions of 4 kg of water, 3 kg of oxygen gas, and 2 kg of food each day to survive. Water and oxygen can be recycled after they are consumed by humans by condensing and distilling water humans release through sweat and urination and recreating oxygen from the carbon dioxide that humans exhale. The ISS is able to recycle 90% of the water and 40% of the oxygen consumed by humans with the losses re-provisioned by the periodic resupply missions that also bring new food, replacement crews, and equipment. On proposed long duration space flights well beyond low-Earth orbit, such resupply missions will not be so routine, so care will need to be taken in closing the recycling loop as much as possible. In principle, food can be grown on trips that are long enough and plants also provide an efficient ecosystem balance for repurposing the carbon dioxide that humans exhale. Optimizing plant growth for food and air recycling remains one of the research challenges taken on by current space missions and may be useful for future crewed missions to Mars or beyond.

To survive, humans also require temperature and atmospheric pressure control. The ISS maintains a temperature of normally 22° C (72° F) with temperatures adjustable between 18° and 26° C, well within the range of 4° to 35° C range to which humans can survive prolonged exposure. The space station has an atmospheric pressure the same as Earth's, but as long as the gaseous concentrations are adjusted, there is, in principle, no upper limit for humans to survive higher pressures since it is pressure differences rather than pressure itself which causes harmful effects. For low-pressure environments, the famous Armstrong Limit (named after US Air Force Major General Harry Armstrong) at 1/16 standard atmospheric pressure is the point below which the water in a human's circulatory system will start to boil with bruising and embolisms resulting. Experiments with test pilots suggest that without a pressure suit or oxygen, you would only have "useful consciousness" for a matter of seconds if in an environment below the Armstrong Limit. Contrary to some popular accounts in fiction, human beings do not explode if exposed to the vacuum of space.

Challenge #4: Radiation

The Earth's atmosphere and its magnetic field provide a shield against much of the most harmful radiation that permeates outer space. In the immediate environment surrounding the Earth, the major source of such danger comes from the solar wind and cosmic rays which have their origin in highly energetic processes beyond the solar systems and, for the most energetic particles, even beyond our galaxy. The most harmful form of radiation is so-called "ionizing radiation" which consists of particles with enough energy to ionize atoms and therefore disrupt the chemical equilibrium in cells. Short term effects from exposure to such radiation include various acute radiation sicknesses and longterm carcinogenic and degenerative effects.

After about six months in low-Earth orbit with the same level of shielding as provided by the ISS, humans receive the equivalent dose of radiation to ten CT-scans which is close to five times the occupational safety level as recommended by health agencies. The increased risk associated with this exposure is one of the major long-term health risks of space flight. In addition to exposure rates, Space Agencies sets career limits on the total amount of radiation exposure which range from 5 to 20 years aboard the ISS

(younger people have lower career radiation limits as some of radiation's long-term cancer and central nervous system effects are shown to arise after many decades). This is at about the same level of radiation exposure that an astronaut would receive on one round-trip to Mars. For planned longer stays on the Red Planet, radiation levels are approximately three times that on the ISS.

For space travel outside of the Earth's protection, [space weather](#) becomes a major consideration.

It is for these reasons that developing technology that can shield humans from radiation for long space voyages is one of the major research programs of the world's space programs. A variety of ideas have been proposed, from a straightforward solution of placing a significant amount of mass between the astronaut and the radiation source to new specialized materials with radiation-absorbing properties or artificial magnetospheres that are mini-versions of the one that protect the Earth.

Challenge #5: Distance and Time

When the Apollo astronauts traveled to the Moon, the trip from Earth to the lunar orbital insertion took two days. The fastest space missions to Mars take months and follow somewhat surprising trajectories in the form of transfer orbits which allow the spacecraft to travel with the aid of two brief rocket thrust maneuvers. Other more energy-efficient proposals to get to Mars have included using the gravitational field of Venus to slingshot spacecraft at the expense of more time needed to travel slowly to Venus.

Using the existing technology to travel to other stars would result in travel times in excess of thousands of years. This is obviously prohibitive for human lives, and so a variety of possible solutions have been proposed to allow for people to get to the other potentially habitable places beyond our Solar System.

Interstellar Space Flight Proposals

In spite of the challenges mentioned above, there are some feasible ways to send spacecraft to nearby stars. The Voyager 1 and Voyager 2 spacecraft are the products of human technology that have travelled the farthest from the Earth — they are both currently coasting through interstellar space. However, they were not designed to study anything beyond our outer solar system. Although Voyager 1 has travelled over 15 billion miles away from the Earth, keep in mind that this is a distance of just 0.00255 light years and the nearest star to the Sun, Proxima Centauri, is 4.24 light years away.

Generation Ships

A concept sometimes used in science fiction are so-called “generation ships” which travel at speeds similar to those we can currently achieve and would arrive at distant stars many thousands of years in the future. The idea is to provide enough material and people to sustain a journey of such a length. Such ships are sometimes envisioned as entire self-sustaining ecosystems that are designed to last far longer than most human technologies survive. Whether such a generation ship would be feasible is a question that would need considerable input from ecologists and social scientists.

Light Sails

Solar Sails

Mass (and its relativistic twin, energy) poses a huge challenge for getting a spacecraft to travel close to the speed of light. One way to get around this is send out the absolutely smallest mass possible and use a sustainable source of energy to accelerate the craft, ideally with the energy source separated from the craft to minimize the total mass. An obvious choice for energy is the Sun, which will keep shining for another 5 billion years. In principle, the power from the Sun could be used to push on a sail that could travel out of our solar system. A good analogy for this, on an earthly scale, is a sailboat that is propelled by wind (which itself is powered by the Sun). With a solar sail, the light from the Sun (photons) directly push on the sail. While photons have no mass, they still impart a quantity called momentum that can speed up the sail. This force or push from the Sun is small — we don't feel it pushing on us — so it takes time for the lightweight sail to speed up. That's the basic idea, and it's really quite simple. In fact, Johannes Kepler thought of the idea as he contemplated how sunlight pushes on the material in a comet (Halley's comet, to be specific, which came by Earth in 1607) to create the comet's long tails. Kepler wrote in a letter to Galileo in 1610:

“Provide ships or sails adapted to the heavenly breezes, and there will be some who will brave even that void.”

Science-fiction writers thought about this, too. In 1964, American writer Arthur C. Clarke wrote the short story “The Sunjammer,” which imagines “sun yachts” racing through the sky as a sport. An idea for a solar sail to visit Halley’s Comet on its return in 1986 was seriously considered by NASA and the Planetary Society, which had been founded in 1980 by Carl Sagan. The first solar sail, called IKAROS (Interplanetary Kite-craft Accelerated by Radiation Of the Sun; also a nod to the mythological Icarus who flew a little too close to the Sun) was launched in 2010 by JAXA, the Japanese air and space agency. The goal of IKAROS was to show that solar sails could be launched and work successfully. The sail for IKAROS was only 7.5 μm thick and was a 14 meter by 14 meter square (196 m^2 total, which is about half the area of a basketball court); it weighed about 2.2 kg. Two successful solar sails were launched in 2015 and 2019 by the Planetary Society, named LightSail 1 and LightSail 2.

Using sunlight to propel sails works well for visiting places in our solar system. However, as you get further away from the Sun, the sunlight gets very spread out and the sail receives less and less of it. By the time sunlight reaches the Earth, we receive 1,360 watts for every square meter (1360 W/m^2). So, the 196 square meter sail for IKAROS received about 270,000 W of power from the Sun ($14 \text{ m} \times 14 \text{ m} = 196 \text{ m}^2$ and $196 \text{ m}^2 \times 1360 \text{ W}/\text{m}^2 = 266,560 \text{ W}$). A way around this is to use a different source of energy that can provide power at this level or much, much higher. This can in theory be done with lasers.

Breakthrough Starshot

The Breakthrough Starshot initiative has the goal of sending a light sail to take pictures of Proxima b, which is the closest known exoplanet to our solar system. The Breakthrough Starshot initiative received \$100 million in funding from entrepreneur Yuri Milner in 2016 (Milner also funded the Breakthrough Listen initiative, which is focused on SETI). The project aims to create a fleet of nanocrafts that are made up of a small chip to take images of the planet that is attached to an extremely thin and lightweight sail. The sails would be pushed on their journey by an array of powerful lasers that are on the Earth’s surface. The goal is to send as many nanocrafts as possible, as some fraction are expected to fail somewhere along the way or after arriving at Proxima b.

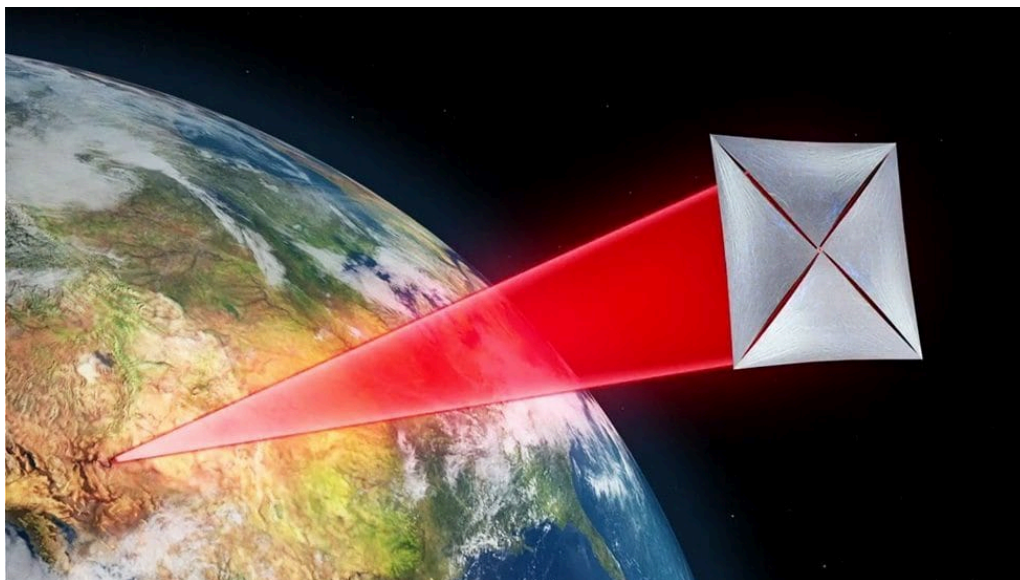


Figure 2 – Breakthrough Starshot’s light sail (artists’s impression). The sail is being pushed forward by an array of powerful lasers on the Earth’s surface. Credit: Breakthrough Initiatives

There are many engineering challenges to overcome, such as designing and building an array of lasers on the Earth that can send out up to 100 GW of power (a GW is a billion watts; sci-fi fans may recognize “1.21 giga watts” as the power needed for the flux capacitor in *Back to the Future*). The chips, which include the camera and other communication equipment, need to be extremely light, on the order of just a few grams (about the mass of a few paperclips), and around the size of a postage stamp. The sail also needs to no more than a few grams total and made of a reflective material, while still being large enough — perhaps a few meters on each side — to function properly. That all said, progress has been made on the project. Some prototypes, with much smaller sails than what the final version will have, were launched into the Earth’s atmosphere and took pictures of the Earth.

Relativistic Advantage and Constant Acceleration

Einstein’s Theory of Relativity offers a few fantastical but physically plausible mechanisms for allowing for interstellar flight in

ways that could in principle allow for a human being to explore the any location in the observable universe in one lifetime. To accomplish this incredible feat, a spacecraft could take advantage of time dilation and length contraction.

Einstein's Theory of Special Relativity is based on two assumptions: the laws of physics are universal in all reference frames (a reference frame is defined by a speed), and the speed of light is constant in all reference frames. Taking these two assumptions to their logical conclusions implied that different reference frames will measure lengths and durations differently depending on the speed. In particular, the faster a frame of reference is moving, the shorter the distances in the direction of motion will physically become. This is length contraction. Likewise, time moves more slowly in moving frames of reference such that less time will pass in a moving frame than in a stationary one.

A key factor used to calculate these relativistic effects is called the Lorentz Factor

$$\gamma = \frac{1}{\sqrt{1-(v/c)^2}}$$

where v is the speed and c is the speed of light. Relativity requires this factor be multiplied or divided by measurements of space or time in different reference frames to account for the effects of the constant speed of light in all reference frames. The Lorentz Factor, then, shows immediately why the speed of light is the universal speed limit since a speed of $v = c$ gives a gamma factor of infinity.

If you know the Lorentz Factor of a frame of reference, that is the amount by which times (t) are dilated (increased) can be related to each other by first considering all clocks that are moving with the reference frame to have a time (t_{proper}) and then asking how the time measured on the *same* clock would be observed to pass when compared to a clock that was not moving (t_{dilated}). Using the Lorentz Factor to relate these two *different* time measurements is done with the equation:

$$t_{\text{dilated}} = \gamma t_{\text{proper}}$$

Thus, if a spacecraft is traveling with a speed $v = 0.6c$, it will have a Lorentz Factor of $\gamma = 1.25$ and therefore a stationary observer would observe that one hour passes on the moving spacecraft for every one hour and fifteen minutes elapsed on their stationary timepiece.

The Lorentz Factor is likewise the factor by which lengths (L) are contracted (decreased). Similar to the consideration of time dilation, an observer can either measure the length between two locations when traveling at the same speed as the common reference frame between the two locations (L_{proper}), or one can measure the *same* length when traveling at a different speed ($L_{\text{contracted}}$). Using the Lorentz Factor to relate these two *different* length measurements is done with the equation:

$$L_{\text{contracted}} = \frac{L_{\text{proper}}}{\gamma}$$

Using the example of a 100-meter-long spacecraft traveling at 60% the speed of light, a stationary observer will observe the spacecraft's length to be a mere 80 meters. At the same time, passengers traveling on the spacecraft, will see the length contraction occurring in their direction of travel between two points in the same reference frame that appear to the observers in the spacecraft to be traveling at 60% the speed of light in the *opposite direction* compared to the spacecraft's travel. This means that there will be literally less space to traverse than if they had measured the proper distance to the destination from a reference frame where everyone was at rest.

This can have dramatic effects at speeds close to the speed of light. For example, traveling at 80% the speed of light will result in a Lorentz Factor of $\gamma = 5/3$ which means that distances to all the stars are reduced by 40% with a comparable reduction in travel time. Traveling at 99.5% the speed of light results in Lorentz Factor of $\gamma = 10$ with a 90% reduction in distances and travel times.

Although no spacecraft can reach the speed of light, spacecraft speeds, in principle, can get arbitrarily close to it with arbitrarily large Lorentz Factors. In such situations, the distances between two points become shorter and shorter and the amount of time it takes to get between two locations becomes smaller and smaller. Move fast enough and you can get to the edge of the universe in as short amount of time as you want.

The trade-off is that that the closer the speed gets to the speed of light, the more energy that is needed to increase the speed by the same amount. Roughly, for speeds that are close to the speed of light the amount of energy needed scales as the Lorentz Factor, so to boost a spaceship to the speed of light will require an infinite amount of energy. Therefore, these speculations require something close to an inexhaustible supply of energy to achieve these fantastic effects. Even the most conservative

approaches require energy budgets that are well in excess of the entire energy budget used by humankind each year. Getting to speeds that will make the journey between Earth and any distant location as short as possible will require developing entirely new sources of energy that are currently well beyond human technology.

Assuming that such an energy source (matter-antimatter, for example) could be harnessed, the planning of the journey has some additional complications. For example, speeding up to high speeds requires accelerating the spaceship which, as anyone who has ever been on a rollercoaster can attest, could result in some discomfort if the acceleration is large enough. A proposed solution that has been offered as a thought experiment is to accelerate the spaceship at a constant 1 g of acceleration at the equivalent to Earth's gravitational acceleration of 9.8 m/s^2 . In such a scenario, the spaceship occupants would feel the same gravity as they did on Earth, and after one year at this constant acceleration the spacecraft will reach 77.5% the speed of light.

Constant acceleration space travel, then is a fairly astonishing physical scenario. A successful trip might entail accelerating towards the destination for the half the trip, and then decelerating for the second half so that the spacecraft arrives in the same frame of reference as the target. This allows for arrivals at even incredibly distant destinations within human-realizable timeframes. Under such scenarios, it takes 3.5 years to reach Proxima Centauri, just under 20 years to reach the galactic center, and just under 50 years to travel the furthest distances observable from Earth.

One complicating matter is that as you start to move at relativistic speeds, the ambient electromagnetic radiation becomes blueshifted into energy realms that are lethal and highly destructive. Likewise, colliding with the particles that might be along the trajectory would be the equivalent of an atom smasher. The spacecraft would have to survive such an onslaught, so some sort of shielding would be required to protect from the exposure.

Returning home brings on other difficulties. While the time passed on the spacecraft is relatively short, the time elapsed on Earth will be substantially greater. This is due to the so-called "Twin's Paradox" which relates that in the context of relativity, an observer who accelerates to a boosted reference frame and then decelerates back will experience less time than an observer who remains in the initial reference frame. This has the effect that while the spacefarer in a constant accelerating spacecraft experiences a 7 year round trip to the nearest star, everyone who remains on Earth will experience 12 years. The astronaut who travels to the center of the Milky Way will return to an Earth that is 55,000 years in the future while the voyages that go to the edge of the universe will return to a Milky Way well after the Sun has become a white dwarf, billions of years in the future.

In spite of all these uncomfortable complications, the very fact that such a journey in a human lifetime is physically possible is somewhat remarkable. While we are nowhere near being able to build such constant acceleration spacecraft engines, there is no law of physics which prevents such a thing from being constructed, though it would be on a scale much more massive than anything currently possible.

Key Concepts and Summary

Exploring the solar system or other stars requires incredible amounts of energy. Energy is required to escape the gravity of Earth and humans evolved in the environmental niche of our planet – space travel has many challenges to the survival of humans including microgravity, radiation, and travel times that range from decades to centuries, for even the closest star systems. Sending robotic spacecraft is a good first start for exploring our sector of the galaxy.

Review Questions

Summary Questions

1. What is the main reason that space travel from Earth is so challenging?
2. What is microgravity?
3. Why does microgravity pose challenges for astronauts?
4. Why is radiation from space harmful to life?
5. How do the large distances to other stars pose a challenge in traveling to them?
6. What are some ways to deal with the vast distances involved in interstellar travel?

Exercises

This page titled [6.1: The Challenges of Space Flight](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

6.2: The Search for Extraterrestrial Intelligence (SETI)

The search for extraterrestrial intelligence inherently assumes that any civilization that we make contact with will have gained a certain level of intelligence. Intelligence can have many different connotations, and in the context of SETI it has a specific meaning: the civilization is able to send out signs or signals that indicate it is an advanced civilization. A good way to start thinking about this is by asking when the Earth became an advanced intelligent civilization that sent out signals indicative of technology. We will start there and see how SETI searches have evolved from our planet's first radio transmissions to modern surveys that scan the skies over broad bands to look for that cosmic needle in the haystack.

Learning Objectives

By the end of this chapter, you will be able to:

- Describe early searches for extraterrestrial life and why radio waves were chosen
- Discuss SETI searches in optical (visible) light
- Describe some of the signals that SETI scientists look for
- Discuss artifacts that humans have sent out into space

The Beginnings of Radio Astronomy

Huge strides in understanding the physical nature of electromagnetic radiation in the mid- to late-1800s led to the first electronic search for radio signals in 1899. Nikola Tesla was born in Croatia and emigrated to the United States in 1884. In 1899, while working on a project that could enable a global wireless communication network (indeed, Tesla was way ahead of the times in realizing that wi-fi networks were quite feasible), Tesla detected a repeating signal that he eventually attributed to Mars:

“Twenty-two years ago, while experimenting in Colorado with a wireless power plant, I obtained extraordinary experimental evidence of the existence of life on Mars. I had perfected a wireless receiver of extraordinary sensitiveness, far beyond anything known, and I caught signals which I interpreted as meaning 1-2-3-4. I believe the Martians used numbers for communication because numbers are universal.” — Nikola Tesla, 1923, *Albany Telegram*.

Another pioneer of radio communication and a contemporary of Tesla, Guglielmo Marconi received the first trans-Atlantic radio transmission in 1901. Marconi detected an anomalous radio signal in 1921 and believed that it was a Martian response to signals that Marconi had sent out earlier. Today, we know that neither Tesla nor Marconi's detections were of a Martian origin, but at that time Percival Lowell's sketches of Mars and claims of inhabitants, as well as H.G. Wells' publication of *The War of the Worlds*, were still fairly recent (the *New York Times* published a story on December 9, 1906 with the headline “There is life on the planet Mars: Prof. Percival Lowell recognized as the greatest authority on the subject, declares there can be no doubt that living beings inhabit our neighbor world.”).

Early SETI Searches

The year 1959 can be considered a turning point in the history of SETI, as that is when an article appeared in the journal *Nature* by Giuseppe Cocconi and Philip Morrison “Searching for Interstellar Communications.” In this paper, the authors suggested searching around the radio frequency of 1420 MHz, as this is the frequency associated with neutral hydrogen (the most common element in the universe), specifically the amount of energy emitted by the atom when the electron changes its spin state (this is called the hydrogen “spin flip”). The thinking was that other civilizations in an “early stage of the development of radio astronomy” would also recognize this frequency. Cocconi and Morrison end the paper with this memorable quote regarding searches for interstellar signs of communication:

“The probability of success is difficult to estimate; but if we never search, the chance of success is zero.”

SETI searches formally began in 1960, when Frank Drake used the 85-foot Green Bank radio telescope to observe the nearby stars Tau Ceti and Epsilon Eridani in a project that he dubbed “Project Ozma.” To minimize confusion from background sources, the radio search concentrated on a very narrow frequency band near 1420 MHz. Drake observed the stars for four months, always

with a negative result until he found a surprising signal on April 8, 1960. An extraordinarily careful scientist, Drake followed up and ultimately discovered that this was a false signal from a high-flying aircraft. But, the realization of what it would feel like to detect another civilization started him on a lifelong search for technological signatures (coined “technosignatures” by Jill Tarter). At this early stage, these searches were referred to as CETI, for Communication with ExtraTerrestrial Intelligence.

The original version of this chapter contained H5P content. You may want to remove or replace this element.

The early 1960s also marks the formal beginning of optical SETI searches. In 1961, a paper was published in *Nature* that proposed searching for signals from interstellar and interplanetary communication by looking for laser pulses. One of the authors of this paper, Charles Townes, had recently invented the maser, which then led to lasers (Townes and a colleague patented the laser in 1960). While there are legitimate reasons for conducting optical searches — for example, signal processing in the optical regime is far simpler and requires less computing power than in the radio, information can be transmitted more efficiently, and there is virtually no optical interference to deal with — the power needed for laser pulse transmissions made this idea seem technically unfeasible. However, thanks to the natural doubling in technology every 2 years (Moore’s Law), lasers caught up and several optical SETI programs were up and running by 2000, notably at Harvard and the University of California at Berkeley.

Other ways of searching for extraterrestrial intelligence were circulating and being published. In 1960, physicist Ronald Bracewell published a paper in *Nature*, “Communications From Superior Galactic Communities,” in which he suggested we look for probes in our solar system sent by other advanced civilizations. Bracewell considered a galaxy-wide network of advanced civilizations who cooperate and send these probes out to stars nearest to their own. Also in 1960, Freeman Dyson published an article in *Science* titled “Search for Artificial Stellar Sources of Infrared Radiation,” which proposed that a very advanced civilization could build a giant structure surrounding its planetary system. These mega-structures would be detectable via their waste heat in the infrared.

Searches were also underway in the U.S.S.R. at the same time. Nikolai Kardashev proposed a scheme for classifying intelligent civilizations based on how much energy they could use (the Kardashev Scale) and published results from a study of two quasars in 1964. One of the sources that Kardashev speculated was possibly from an extraterrestrial civilization, CTA-102, caught the attention of the U.S. rock band The Byrds and they recorded the song CTA-102 in 1967, reflecting the interest from the general public in SETI. A book co-written by Soviet astronomer Iosif Shklovskii and American astronomer Carl Sagan called “Intelligent Life in the Universe” was published in 1966. It is a remarkable collaboration considering that this was right in the middle of the Cold War. SETI collaborations between American and Soviet scientists continued, and an international conference was held at the Byurakan Astronomical Observatory in Soviet Armenia in 1971. At the 1971 Byurakan conference, CETI was renamed SETI, since no communication had actually been detected yet.

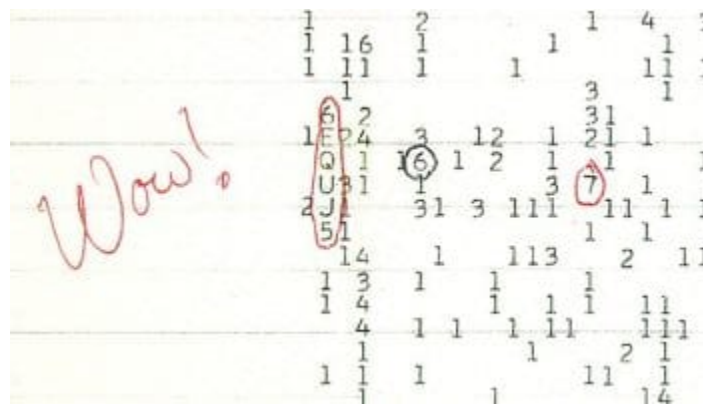


Figure 1 – The “Wow!” signal. Credit: [Wikipedia](#), public domain

More dedicated SETI searches were conducted in the 1970s using various radio telescopes, including the Arecibo dish in Puerto Rico and Ohio State’s Big Ear facility. On August 15, 1977, an astronomer observing at Big Ear picked up a very intense signal right around the hydrogen spin-flip frequency at 1420 MHz. When an astronomer working as a volunteer at Big Ear saw the print-out showing the signal, he wrote the word Wow! in red ink to convey his excitement (Figure 1). The “Wow! signal” has never repeated and the same region is still searched from time to time. This is a similar detection to Frank Drake’s “false alarm” in 1960, which turned out to be an aircraft, but the signal remains a subject of minor fascination to SETI scientists. In 1971, scientists Bernard Oliver and John Billingham, through their work for Project Cyclops, proposed extending radio SETI searches from 1420 MHz out to around 1700 MHz, to include radio emissions from the hydroxyl group (-OH); they eventually coined the term “water hole” for this range of frequencies. In 1979, on the 20th anniversary of the Cocconi and Morrison paper, Jill Tarter published the results of her radio search of 201 nearby stars using the 91-meter radio antenna at the National Radio Astronomy Observatory.

Want to know more: The Kardashev Scale

If we do detect communication signals from a technologically advanced extraterrestrial civilization, what should we expect? This is a vague question with really anything you can imagine being a valid answer, as we have never encountered any life outside of Earth. However, Nikolai Kardashev thought about this seriously and came up with a classification system for extraterrestrial civilizations based upon how much energy they are able to harness from their surroundings. Energy consumption can be thought of as a proxy for how technologically advanced a civilization is. He published this in 1964 in the journal *Soviet Astronomy* in a paper titled “Transmission of Information by Extraterrestrial Civilizations” and suggested three civilization types:

- **Type I** civilizations use all of the energy available from their **planet**. Kardashev initially wrote that a Type I civilization would be at a technological level close to Earth’s level in 1964. This would include energy from natural phenomena such as weather. Residents of planet Earth are not yet at Type I.
- **Type II** civilizations use all of the energy available from their **star**. One example of a technosignature from a Type II civilization is a Dyson Sphere.
- **Type III** civilizations use all of the energy available from their **galaxy**. How to imagine this again brings us back to speculation, but we could think about a civilization harnessing massive amounts of energy from a supermassive black hole at the center of a galaxy.

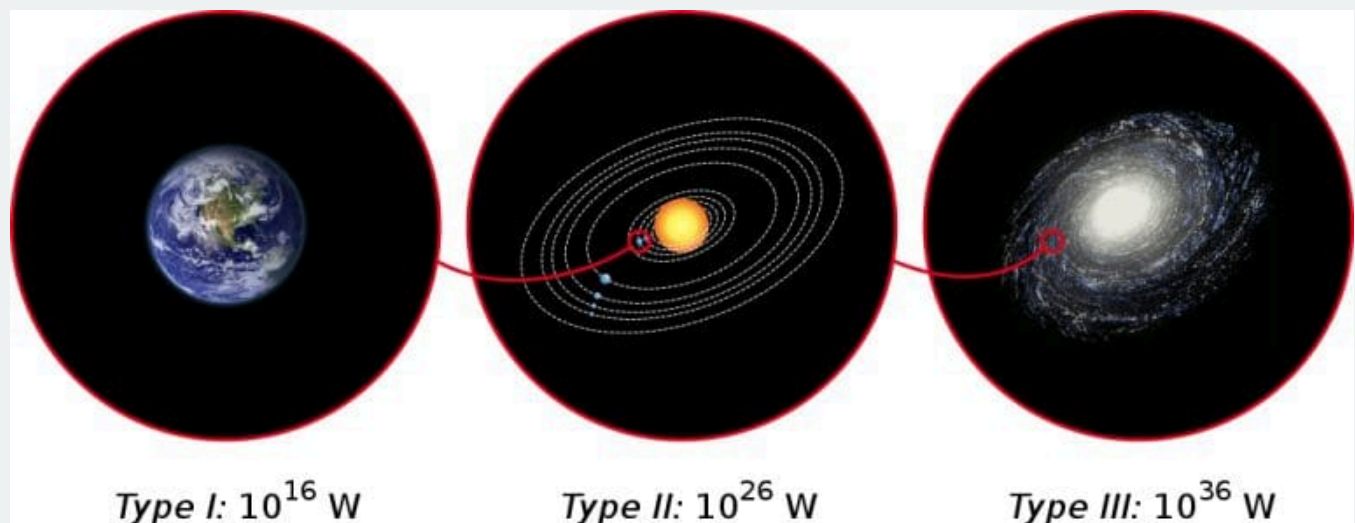


Figure 2 – The Kardashev scale for civilizations, with the amount of energy consumption for each indicated in Watts (W). A type II civilization uses all of the available energy from its host star, and 10^{26} W is the power output for a star like the Sun. Credit: [Wikipedia](#), CC BY-SA

More recent modifications to the Kardashev Scale also encompass Type 0 to Type 4, which includes biological civilizations (Type 0) and civilizations that can use all of the energy available in the universe (Type 4).

In 1980, The Planetary Society was founded — Carl Sagan was one of the original co-founders and today the Society is led by the American science communicator Bill Nye — to promote SETI to the public and to provide a means for non-governmental funding for SETI projects. By 1981, The Planetary Society had funded “Suitcase SETI,” a portable observing kit that was used at Arecibo and eventually morphed into the META SETI projects. The SERENDIP (Search for Extraterrestrial Radio Emissions from Nearby Developed Intelligent Populations) searches began at the University of California Berkeley in the early 1980s and took advantage of commensal observations, meaning that several projects could run at the same time. The non-profit SETI Institute was founded around this time, in 1984, with Jill Tarter conducting its first SETI search.



Figure 3 – Jill Tarter served as project scientist for NASA's SETI program before funding was terminated in 1994. She is now Chair Emeritus for SETI research. Credit: Seth Shostak/SETI

Modern SETI

The 1990s were a challenging time for SETI, as government funding for SETI was axed in 1993. Given the fascination with aliens in Hollywood science fiction movies, one would think that the public and the government would consider SETI to be an exciting and worthwhile area of research. This is a research field that costs little but has faced active opposition. The federally funded SETI program was started in 1992, and just a year later, Senator Richard Bryan of Nevada argued that funding should be shut down: "...As of today, millions have been spent and we have yet to bag a single little green fellow. Not a single Martian has said 'take me to your leader,' and not a single flying saucer has applied for FAA approval."

Scientists have been funded to search for elusive dark matter particles for more than 40 years, but not a single one has yet been found. Searches for black holes went on for decades before the first candidates were found. The hunt for exoplanets dates back a few centuries, but the first exoplanet was not detected around a Sun-like star until 1995. Transformational science requires that we push the boundaries.



Figure 4 – The Allen telescope Array in northern California. The array currently has 42 radio dishes built. Credit: Seth Shostak, SETI Institute

Fortunately, SETI pursuits continued on in the wake of the federal funding being cancelled. UC Berkeley's SERENDIP programs continued into the 2010s and the SETI@home project came online in 1999. After more than a decade of planning, the Allen Telescope Array (ATA) came online in 2007 at the Hat Creek Radio Observatory in northern California, and this was the first telescope built from the ground up to perform SETI searches. Major funding for the ATA came through a donation from Paul Allen, the co-founder of Microsoft.

Breakthrough Listen

A major milestone for SETI came in 2015 with the launch of the Breakthrough Listen initiative. Like the SETI Institute, Breakthrough Listen is supported by both scientists and other philanthropists. Funding was provided by Yuri Milner, and physicist Stephen Hawking was involved in the initial launch. Breakthrough Listen is being funding \$10 million a year for 10 years. The

goals of Breakthrough Listen are to search for technological signatures of intelligent life in the universe and answer the question “Are we alone?”. The science goals are to observe 1 million nearby stars and 100 galaxies using radio telescopes. Breakthrough Listen is also surveying the Milky Way’s dense disk and central region in both the radio and optical parts of the electromagnetic spectrum (the central region includes the Milky Way’s super-massive black hole, which a Type III civilization could potentially be pulling energy from). All of the data collected is immediately made available to anyone interested in using it. This is the largest dedicated SETI search to date and is currently centered at the University of California at Berkeley and led by Andrew Siemion.

Breakthrough Listen is a global collaboration using the largest and most sensitive radio telescopes in the world, including: the 64-m Parkes Telescope (Murriyang) in Australia; the MeerKAT array in South Africa, which has sixty-four 13.5-m dishes that can be used in tandem; the 100-m Green Bank Telescope in West Virginia; and the 500-m FAST Telescope in China. Some of the telescopes used by Breakthrough Listen, such as the Green Bank and Parkes, are steerable, meaning that they can point directly at any object of interest that may arise at some point. Observations for Breakthrough Listen began right away in 2016 using the Parkes and Green Bank telescopes.

Optical SETI (OSETI) searches are being conducted by Breakthrough Listen through a partnership with the VERITAS telescope array in Arizona. VERITAS is looking for high-energy gamma ray pulses that arrive as very short bursts, and this strategy can be co-opted to search for optical SETI signals. This is aligned with OSETI’s strategy of looking for fast (on the order of nanoseconds) laser pulses that were sent as beacons from another technologically advanced civilization. Breakthrough Listen is also doing a deep search for laser beacons by analyzing spectra of stars taken by the Automated Planet Finder telescope at Lick Observatory. These signals would appear in a star’s spectrum as a very strong emission line.

In addition to stars and galaxies, Breakthrough Listen is also looking at other objects of interest to SETI. For example, an object, now known as Oumuamua, passed close to Earth in October 2017 and was detected by the Pan-STARRS mission. Oumuamua immediately garnered attention from the astronomical community as its motion suggested that it did not come from our solar system but was an “interstellar” visitor. What kind of an interstellar visitor, exactly? Natural origins, such as an asteroid or comet, were considered, but it was also suggested that Oumuamua was a piece of alien technology sent our way for communications. Breakthrough Listen used the Green Bank Telescope to observe Oumuamua in December 2017 and found that there were no radio transmissions emanating from the object, evidence against the alien artifact theory.

Earth’s Technosignature: Voyager Spacecraft

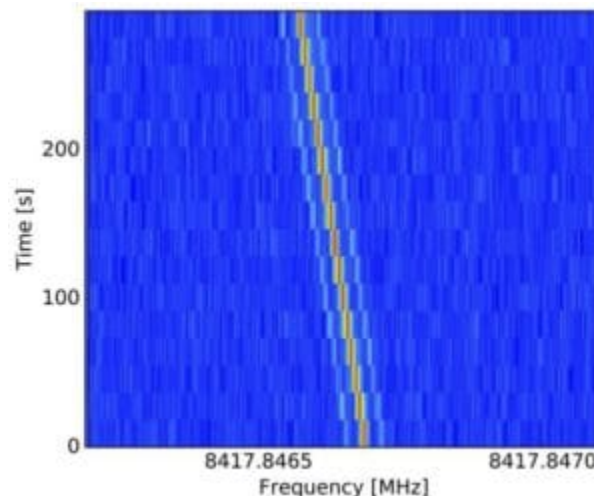


Figure 5 – Evidence of a technologically advanced civilization. In this case, the civilization is Earth. The figure shows radio transmissions from the Voyager 1 spacecraft. Credit: [PASP](#), IOP Science

How can we have confidence that any search strategy will succeed if we haven’t yet detected any signs of extraterrestrial intelligent life? One proof-of-concept for our SETI radio searches is to detect one of our own artifacts. The Voyager I spacecraft was launched in 1977 and is currently coasting through the Kuiper Belt out beyond 160 AU. In 2020, as part of Breakthrough Listen, the Green Bank Telescope pointed at the Voyager I spacecraft and recorded data for 5 minutes. Figure 5 shows the signal as recovered in the GBT data. The receiver at the GBT picked up Voyager’s carrier signal which broadcasts at a frequency of 8415 MHz.

Many other technosignatures from Earth are regularly picked up, such as the rovers on Mars or the New Horizons mission that is

Extraordinary Claims: BLC1

An exciting moment for Breakthrough Listen came in 2020 with the announcement of a signal, dubbed BLC1 (Breakthrough Listen Candidate 1), that could potentially fit the criteria for a technosignature, which is a transmission from another technologically advanced civilization. The signal was found in radio data from a 2019 study of flares coming from the star Proxima Centauri, a cool red dwarf which happens to be the closest star from the Sun and has a planet orbiting in its habitable zone. The data was taken with the Parkes Telescope in Australia and was carried out to assess how much damage flares from red dwarfs like Proxima Centauri could cause for any life on a planet orbiting the red dwarf. To see if BLC1 was actually a technosignature, a rigorous analysis to eliminate a false positive was carried out.

The BLC1 signal was detected at a very specific frequency of 982.0 MHz, with a narrow bandwidth (meaning the signal was very concentrated around the main frequency) of around 3.8 Hz; signals that are this narrow have a high chance of being unnatural. If a signal originates from Earth, an easy way to verify that it comes from Earth is to point the telescope away from the source and see if the signal persists. This was done for the Parkes observations of Proxima Centauri, and the signal at 982 MHz did not show up when the telescope was pointed away from the star, suggesting that the signal did not originate on Earth. If a signal is coming from another planet, as opposed to some source of interference on the Earth, you should see a slight change in the frequency, or drift, over time due to the motion of the source with respect to the receiver on Earth. BLC1's frequency did indeed drift, although not exactly in the sense expected.

The next step in confirming that a signal is legitimate is to eliminate a false positive due to radio interference from the Earth. The pipeline used by the student analyzing the Proxima Centauri data should have eliminated all of these but a stray could still sneak through. We discussed Frank Drake's "false positive" detection in 1960 — ultimately, that signal was attributed to aircraft in the Earth's sky. This source was of course checked, and 982 MHz is reserved for aircraft in Australia. However, no aircraft reported any radio transmissions at or around this frequency. (Another memorable example of tracking down interference occurred when putative technosignatures at 1420 MHz were detected with the Parkes Telescope (these signals were referred to as "perlytons") that primarily showed up during daytime work hours turned out to be due to a microwave oven door at Parkes being opened while the oven was running. This particular story underscores the importance of thinking of every possible source of radio interference from Earth; it took 17 years to figure out the perlyton mystery!)

Ultimately, radio interference was found to be the cause of BLC1. Sofia Sheikh, now at the SETI Institute, looked through archival data taken with the Parkes Telescope and identified the same signal in other observations, some present when the telescope pointed away from the source. As a result, Sheikh developed a detailed procedure to help eliminate some of the more elusive sources of radio interference. This will help in assessing any future potential technosignatures that are found. You can watch Dr. Sheikh discussing the BLC1 signal in the video below.

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Video Credit: [Berkeley SETI Research Center](#)

Read More

Sheikh, S. Z. et al. (2021). Analysis of the Breakthrough Listen signal of interest blc1 with a technosignature verification framework. *Nature Astronomy*, 5(11), 1153-1162. doi.org/10.1038/s41550-021-01508-8

Wright, J.T. (2020, December 20). *BLC1: A candidate signal around Proxima*. AstroWright. <https://sites.psu.edu/astrowright/2020/12/20/blc1-a-candidate-signal-around-proxima/>

Messages from Earth

The discussion of SETI above has focused on searches that humans have carried out to look for signs of extraterrestrial intelligence. We can also ask what another civilization would find if they pointed their detectors and receivers at Earth. This is in many ways a speculative question as we must assume that another civilization will search in ways similar to our own, but it is still a valuable exercise to think about this in a broader context and consider how our technology — and our species — are evolving.

Unintentional

Earth has been a technologically advanced civilization for a little over a century. Radio and TV broadcasts have been traveling away from the Earth since the 1920s and 1930s. We can think of this type of transmission as part of Earth's radio leakage, or everyday radio transmissions that someone or something “eavesdropping” on Earth would detect. Other types of radio leakage from Earth include military radar, and, more modernly, cell phones. You can see exactly how radio frequencies in the U.S. have been allocated by the NTIA (National Telecommunications and Information Administration) [here](#) — the takeaway is that we are sending out an immense amount of radio waves!

Intentional

Arecibo Radio Message

After the Green Bank Conference in 1961, Frank Drake began seriously thinking about what information to encode into a radio message sent out from Earth. Drake decided on a grid where each square was either filled or empty, meaning set to a binary value of either ‘0’ or ‘1’. This grid could be filled in to convey information (as much as can be done in this simple way) and then transmitted out toward another potentially technologically advanced civilization. Despite the fact that only one of his colleagues from the Green Bank Conference could decode a test message he put together (that person was Bernard Oliver), Drake got some input from others, including Carl Sagan, and in 1974 there was a fortuitous opportunity to actually send the message.

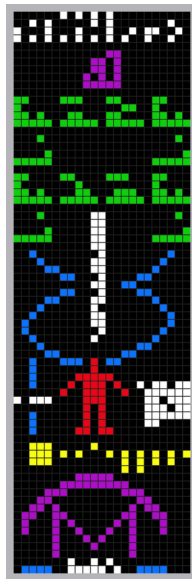


Figure 6 – The Arecibo message, broadcast from the Arecibo Observatory in 1974. Credit: [Wikipedia](#), CC BY-SA

The Arecibo Observatory in Puerto Rico opened in 1963 and the radio dish underwent a major upgrade in 1974. To help publicize the newly gained power of the Arecibo dish, Drake and Sagan decided to transmit a radio message in the direction of the globular cluster M13 using the refurbished dish. The message was a grid with 23 columns and 73 rows, leading to a total of 1679 binary digits (Figure 6). These numbers were deliberately selected because each is a prime number and it is thought that perhaps math is a universal language and other civilizations might recognize these as special numbers. To transmit the message, two frequencies very close to each other were used, with one of the frequencies representing zeroes and the second representing ones. It took a little under 3 minutes to send out this message using 3 TW (3 trillion Watts) of power. No response is expected and the event was really intended to show just how much power the transmitter at Arecibo could send out.

Artifacts

The 1970s were a prolific decade for exploring our solar system. The *Pioneer* and *Voyager* missions were both successfully launched to study the outer planets in our solar system. Pioneer 10 left the Earth in 1972 and Pioneer 11 in 1973. Pioneer 10 and 11 each had an identical metal plaque attached to it. The plaque was designed by Carl Sagan and meant to convey some information about Earth and show some symbols of our civilization. A sketch showing the spin-flip frequency of hydrogen was included, as this was the dominant frequency that humans used to search for other civilizations.

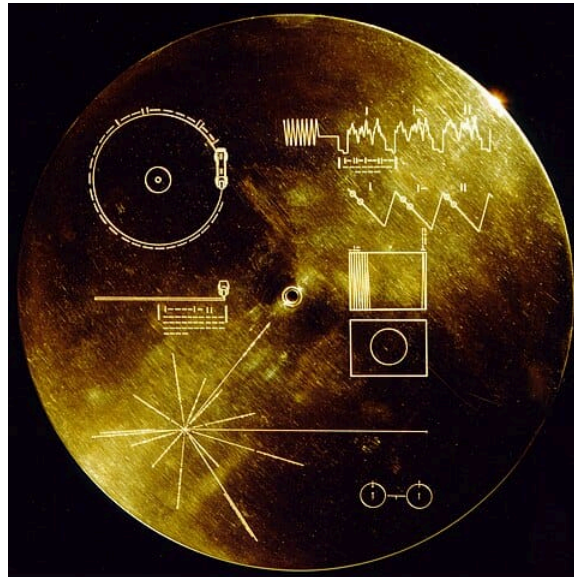


Figure 7 – The Voyager Gold Record cover. Credit: NASA, public domain

Voyager 1 and Voyager 2 were launched in 1977 and both included a relic of our humanity. This time, instead of a plaque, the object was a phonograph record. It may seem outdated now to send out a record, but this was the dominant musical medium at that time; the first commercial compact disc (CD) wasn't pressed until 1982. (If you are curious, the first CD was released in Japan and was American rock artist Bruce Springsteen's album "Born in the USA.") The records are made of gold, vinyl being deemed too flimsy to withstand the interstellar journey. The records also contain a stylus — the part of the arm on a record player that holds the needle — and instructions on how to play the record are etched onto the aluminum cover (Figure 7).

The information on the records describes our biochemical make-up, including and the nucleotide bases in our DNA, and provides iconic cultural images, *circa* 1977. In addition to greetings in more than 50 languages, the Voyagers are carrying 90 minutes of music, including the haunting "Dark was the Night, Cold was the Ground" by Blind Willie Johnson and Chuck Berry's "Johnny B. Goode." A 1977 Saturday Night Live skit portrays extraterrestrials finding the recording and sending back the message: "Send more Chuck Berry."

The Voyager spacecraft are now destined to orbit the center of the galaxy, like all of the stars in the Milky Way. They may well live outlive our Sun and our solar system. Perhaps in the very distant future, another civilization will intercept these spacecraft. If they decode our message, they will know that once upon a time in the galaxy, there was another civilization with the generosity to send a message: "You are not alone. Here is the code for carbon-based life that lived on our planet. Even if our species is never able to learn if we are alone in the vast universe, we hope that this message will give you some answers."

What to do if a signal is detected?

Extraordinary claims require extraordinary evidence! The first step is to confirm the detection using the most sensitive detector available, and the next is to be sure that the signal is not a false positive from an astrophysical source or human interference. This is explored more deeply in the next chapter on Technosignatures and Biosignatures.

Key Concepts and Summary

The Search for Extraterrestrial Intelligence (SETI) started in the 1960. The first SETI search targeted two of our closest

neighboring stars, Epsilon Eridani and Tau Ceti, for a radio transmission at 1420 MHz. Choosing one narrow bandpass (out of an almost infinite bandwidth) and pointing to the stars one led to a characterization of SETI as searching for a needle in a haystack. However, the searches continue and are being expanded from radio to optical and from pointing to single stars to full galaxies. An intentional message was transmitted by the Arecibo Observatory in 1963, creating considerable political backlash. However, our radio and television signals have been (unintentionally) transmitted for decades and are now washing over some of our nearest stars.

Review Questions

Summary Questions

1. Why did the first SETI searches focus on radio waves (as do the majority of searches today)?
2. Why is the frequency of 1420 MHz often searched around?
3. How do optical SETI searches differ from radio SETI searches?
4. How were SETI searches funded after the U.S. government cut off federal funding in 1993?
5. What are some of the strategies that modern SETI searches use?
6. Why is the Voyager I spacecraft considered a technosignature from Earth?
7. What are some artifacts that the Earth has sent out?
8. What information was contained in the Arecibo Message and why was it sent out?

Exercises

1. You can look through the Breakthrough Listen data and find the Voyager 1 signal! The Berkeley SETI Research Center provides a tutorial and link to the data at <https://seti.berkeley.edu/listen/data.html>.
2. Open the link to the [United States Frequency Allocation Chart](#). Look at the region around the “water hole” from about 1420 MHz to 1670 MHz (1.42 GHz to 1.67 GHz). What other types of transmissions are being broadcast around the water hole? Do you think this is a concern if other civilizations are listening to Earth?
3. Can you decode the Arecibo Message? The top 4 rows are a primer, showing the symbols for the numbers 1 through 10.
4. The Earth is estimated to be 0.72 on the Kardashev scale. What do you think it would take to bring Earth to 1.0?
5. Design your own “Arecibo message” using binary digits (0s and 1s). If you are including words, do you think another civilization will understand them?
6. Design your own artifact to send out. Describe what you would include and where you would send it.

6.3: Biosignatures and Technosignatures

Traveling to another solar system to visit exoplanets is impossible today and in the foreseeable future. SETI scientists instead search for signs of life by looking for its telltale signature. These signatures can take the form of biological signs of life, like oxygen (O_2) for Earth life, or signs of technology, such as the byproducts of chemical pollution from an industrial civilization. A more direct detection of technology could come in the form of an artifact that was deliberately sent our way. In this chapter, we will consider the different forms that these signatures of life can take and consider which type we may expect to find first.

Learning Objectives

By the end of this chapter, you will be able to:

- Explain the concept of a biosignature and describe Earth's biosignatures
- Discuss how biosignatures can be searched for in exoplanetary atmospheres
- Explain the concept of technosignatures and describe Earth's main technosignatures
- Discuss some technosignatures that astrobiologists search for from other planets
- Discuss the “cosmic haystack” and how much of it has been searched for signs of other intelligent life

Biosignatures

It is apt to begin a discussion of biosignatures with the adage “Extraordinary claims require extraordinary evidence.” As we look at some possible signatures of life, we cannot immediately conclude that the detection of just one biosignature means we have found other life in the universe. For any potential biosignature detected, absolutely all abiotic sources must be ruled out. False positives can be difficult to identify but an exceedingly critical analysis is in order.

A good place to start in pondering atmospheric biosignatures on other planets is to see what Earth's atmosphere looks like — and what another civilization would see if they took a spectrum of the Earth. Figure 1 shows the spectrum of Earth's atmosphere at optical and near-IR wavelengths. The first thing that you may notice is water...and recall “Follow the water!”. This is a sign of liquid water oceans on Earth, so we are indeed excited when H_2O is detected in an exoplanetary atmosphere, but water vapor can also be produced naturally (recall that Jupiter and Saturn have water vapor in their atmospheres).

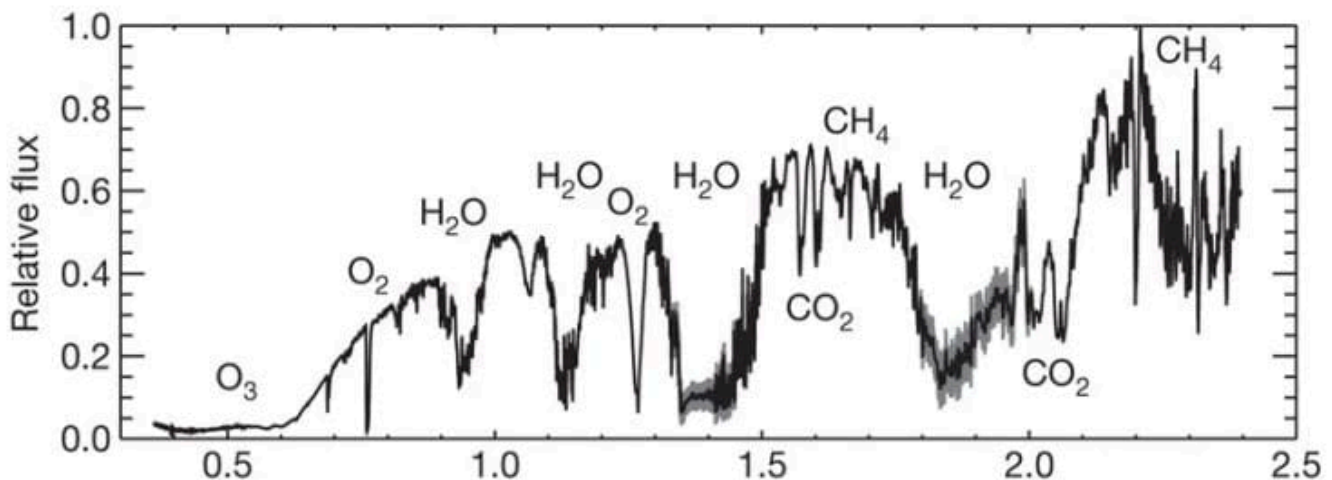


Figure 1 – A spectrum of the Earth's atmosphere. From the spectrum, we can detect oxygen (O_2 , O_3), carbon dioxide (CO_2), water (H_2O), and methane (CH_4) that indicate the presence of life on Earth. Credit: [Nature](#)

Carbon dioxide is also a byproduct of life on Earth — all animals exhale CO_2 and CO_2 is released when plants decay. But CO_2 can also be produced naturally through volcanic eruptions, and through anthropogenic (man-made) sources such as the burning of fossil fuels or deforestation. To further exemplify this point, Figure 2 shows the atmospheric profiles for Venus, Earth and Mars.

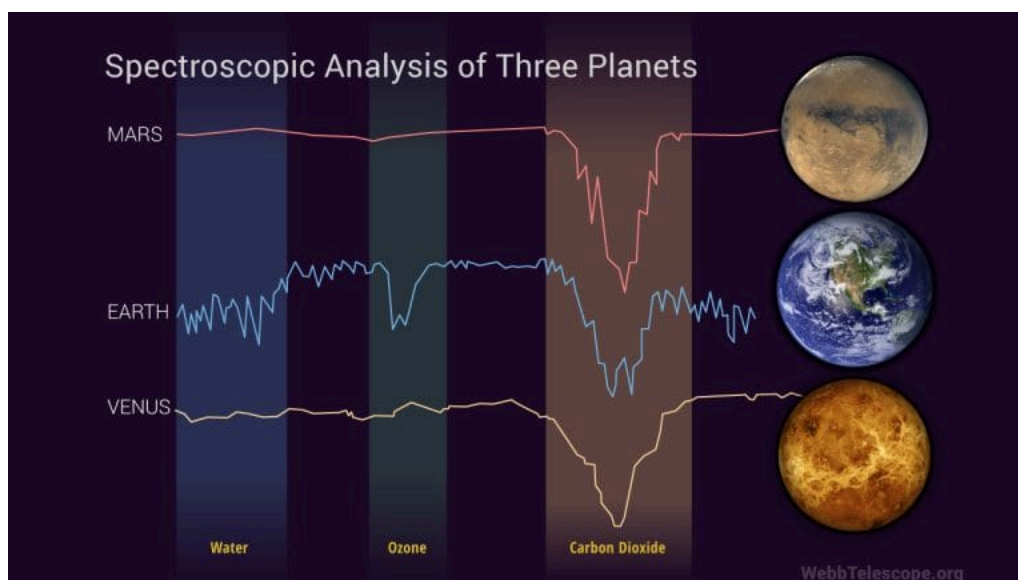


Figure 2 – A comparison of the atmosphere spectra from Mars, Earth, and Venus. The dips mean that the molecule is present in that planet's atmosphere. Note how water and ozone is only present on Earth, setting it apart from our neighboring planets. Credit: [NASA](#), [ESA](#), [CSA](#), [Christine Klicka \(STScI\)](#), public domain

Mars, Earth and Venus all have CO_2 in their atmospheres but only Earth shows water and ozone (which is related to O_2). Let's take a deeper look at some potential biosignatures and their abiotic sources. As we'll see, oxygen alone or methane alone would not be a good biosignature, but the combination of oxygen and methane is a smoking gun for life. At the temperatures and pressures in the atmosphere of Earth, oxygen would react quickly with methane, producing carbon dioxide and water. The simultaneous existence of these elements on Earth occurs only because life is producing these elements and maintaining chemical disequilibrium.

Oxygen

Life on Earth keeps O_2 in our atmosphere — it is a byproduct of life on Earth. Ozone — O_3 — is formed when an O_2 molecule is broken apart and one of the single oxygen atoms combines with another O_2 molecule. So, the presence of ozone is a proxy for O_2 . Of course, just O_3 is not enough and a more compelling case would be made if both O_3 and another biosignature were detected. For example, ozone and water in our atmosphere point to habitable conditions on the surface.

False negatives need to also be considered in addition to false positives. Just because an atmosphere does not show O_2 , this does not necessarily mean that there is no life. On Earth, cyanobacteria began putting O_2 into our atmosphere perhaps as early as 3.8 billion years ago (stromatolites) but it would not have showed up in a spectrum because it would immediately react with other elements. On a similar note, if there was a very low level of O_2 , our instruments may not be sensitive enough to detect it. Finding unambiguous signs of life is a tricky game to play.

Methane

Methane (CH_4) is another molecule that can be produced both biotic and abiotic sources. Methane is a natural byproduct of volcanic outgassing and, in our own solar system, the lifeless worlds Uranus and Neptune have methane in their lifeless atmospheres. Methane can also be produced geologically in hydrothermal vent communities through a process called serpentinization. Methane is created by life as a waste product. Some animal life, such as cows and sheep, produce methane through microbes in their stomachs. These ancient archaean microbes, called methanogens, take in H_2 and CO_2 and expel CH_4 as waste. Like O_2 , CH_4 will generally not stay in an atmosphere unless it is continuously replenished causing chemical disequilibrium. This is why the topic of methane in Mars' atmosphere often pops up in the news, as it could be interpreted as a sign of life (but, if the detection is real, is very likely due to permafrost evaporating at the Martian poles).

Methane on its own would not be a strong biosignature but if it was detected with O_2 then a very strong case could be made for life, as this combination would be hard to create without photosynthesis.

Sulfur gases

Gaseous molecules containing sulfur are produced by both abiotic and biotic sources on the Earth. Simple sulfur gases such as SO_2 and H_2S are produced by metabolism but also are produced naturally through volcanic outgassing. A less ambiguous biotic signature is from more complex sulfur cases such as dimethyl sulfide (DMS) and dimethyl disulfide (DMDS). DMS on Earth is created mainly by marine phytoplankton and is also found in microbial mats.

The atmosphere of the exoplanet K2-18 b — a Super-Earth orbiting in the habitable zone around its host star — was recently studied by JWST. When K2-18 b passed in front of its host star, K2-18, the infrared spectrograph on JWST collected the light from the planet and the star, and was able to separate out the atmosphere of just K2-18b, shown in Figure 3. DMS was detected but further studies will be needed to corroborate this detection.

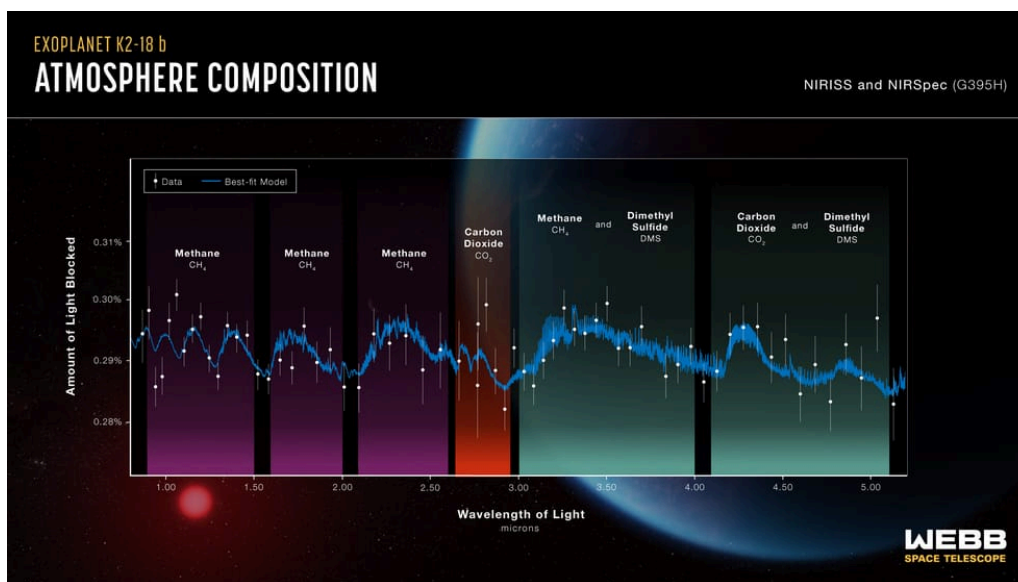


Figure 3 – Spectrum of the exoplanet K2-18 b, taken with the James Webb Space Telescope (JWST). Some of the chemicals in the atmosphere of K2-18 b are labeled, such as dimethyl sulfide (DMS), which is a sign of life on Earth. Credit: [NASA](#), [CSA](#), [ESA](#), [STScI](#), public domain

Nitrous Oxide

Nitrous oxide (N_2O) is detected in excess (chemical disequilibrium) in the Earth's atmosphere. It can be created from both abiotic and biotic sources. One biotic source is actually more of a technosignature — the amount of N_2O in our atmosphere has risen exponentially since the Industrial Revolution in 1750. It is a byproduct of agriculture and fossil fuel burning. It is also produced by bacteria that use nitrates in the soil and water to create N_2O . Nitrous oxide can also be formed naturally by lightning, but the bulk of N_2O in Earth's atmosphere is due to life.

If the detection of N_2O is to be considered a biosignature, it must be considered along with other factors about the exoplanet and other gases that are found in its atmosphere.

Vegetation Red Edge

In addition to biosignatures gases in the atmosphere, a planet's surface can also reveal the presence of life. On Earth, the red edge is a property of chlorophyll in photosynthetic plants, which becomes abruptly reflective at infrared wavelengths as shown in Figure 4. At wavelengths above about $0.7 \mu\text{m}$ (700 nm, which is right at the edge between red visible light and infrared light), leaves will reflect sunlight and show a sharp increase in the amount of light reflected at this point.

image

Figure 4 – The “red edge” created by plants on Earth. At wavelengths greater than about $0.7 \mu\text{m}$, the amount of sunlight reflected by the plant shows a dramatic increase. Credit [Astrobiology](#), Mary Ann Liebert Inc.

The specific wavelength for the red edge on Earth is a consequence of the properties of the Sun and the intensity of its light at different wavelengths (the most intense light from the Sun being yellow-green). For an exoplanet orbiting an M star, the most intense radiation is in the infrared part of the electromagnetic spectrum, so the red edge could actually be some different color or perhaps a black edge, depending how vegetation responds to the different intensity of light. H.G. Wells considered this and wrote in *The War of the Worlds* in 1895: “The vegetable kingdom in Mars, instead of having green for a dominant colour, is of a vivid blood-red tint.”

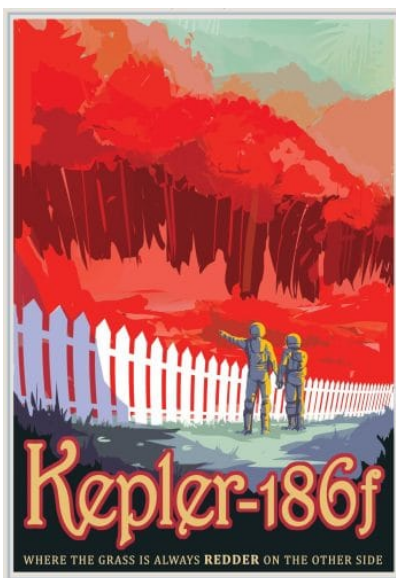


Figure 5 – A travel poster created by NASA for the exoplanet Kepler-186f. Kepler-186 is a red dwarf star so any vegetation on this planet could appear a different color. Credit, [NASA](#), public domain

This reflectance property of chlorophyll also keeps plants from overheating, but is also integral to plants growing efficiently through a process called shade avoidance. Plants are capable of detecting a lack of infrared light and this triggers a response in plants that produces longer stems and larger leaves, which allows a plant to capture more sunlight.

Photosynthetic organisms were a critical step in the evolution of aerobic respiration in eukaryotes. Therefore, a similar detection in the atmospheres of other planets would be an intriguing clue for bio-activity. The red edge is a strong signal in the atmospheric spectra from Earth and an easy feature to search for in the atmospheres of other planets.

The image to the left shows an example of an exoplanet, Kepler-186f, that is Earth-sized and falls into the habitable zone around its host star. Kepler-186 is a cool, red dwarf star so any plants that grow there could adapt to their star's light in such a way that leaves have different colors that differ from those on Earth.

Technosignatures

In 2007, Jill Tarter wrote in her proceedings from the International Astronomical Union's 2006 General Assembly meeting:

“If we can find technosignatures – evidence of some technology that modifies its environment in ways that are detectable – then we will be permitted to infer the existence, at least at some time, of intelligent technologists.”

This succinct definition encompasses a broad range of possibilities for technosignatures, including some of the biosignatures discussed above that are the result of a technologically advanced civilization's industrial output.

The first SETI searches were looking for technosignatures. Frank Drake's Project Ozma in 1960 focused on the 21 cm hydrogen line with the assumption that other technologically advanced civilizations would utilize this wavelength for communication. Dyson spheres and Bracewell probes are both signs of technology. It wasn't until exoplanets were first discovered in the 1990s that searches for biosignatures in their atmospheres could be realistically considered.

We can imagine many different types of technosignatures, from communication signals using light to entire planetary systems that have been re-engineered to better suit their inhabitants' needs. It is impossible to know what a [Type II or Type III civilization](#)

might have the technological capacity to do so we will stay focused on technosignatures that do not require any new laws of physics.

Radio Signals

In the chapter on SETI, we discussed the theory behind radio searches at certain “magic” frequencies such as the hydrogen spin-flip at 1420 MHz. Modern searches cover a broader range of radio wavelengths: Breakthrough Listen searches the range from 1-10 GHz. Of course, it is impossible to predict what frequency another civilization would choose, but we can gain some insight from Game Theory and the concept of Schelling Points. The basic idea of Schelling points, which is explored more deeply in the box below, is that it is worthwhile to consider what strategy your opponent will adopt before beginning a game. For SETI, this translates to thinking about how another civilization would communicate with us, especially if the civilization is also thinking about how we will try to communicate with them.

Want to know more: Schelling Points

Game Theory is the science of strategy. More often than not, the best way to win a game is to have a strategy. But what if you cannot communicate with the other player in a game and the goal is for you both to cooperate to win? There are still ways for the players to come up with a winning strategy by intentionally thinking about what the other person will do and realizing that they will also be thinking about what you will do. The American economist Thomas Schelling wrote about this in his 1960 book called “The Strategy of Conflict.” Schelling focused on political conflict resolution and how to navigate negotiations in a mutually beneficial way. He gave some examples that explicate the idea of what he called focal points.

Imagine you are to meet someone in New York City but you have no plan and you cannot communicate with the other person. You both need to guess where and when to meet. New York City is a big place so random guesses won’t be a successful strategy. [For fun, think about this specific question and come up with an answer. This specific game was carried out in NYC and was successful!] There are some obvious places and times and, so the thinking goes, two rational people will arrive at one of these points.

In his book, Schelling specifically mentions 1420 MHz as a focal point for SETI. Schelling’s book came out just one year after the Cocconi and Morrison paper that suggested 1420 MHz as a frequency that other civilizations like us should be aware of.

We can also consider a Schelling point (Jason Wright has called these Schelling points in the context of SETI, as not to be confused with the other meanings of the term focal point) in time for extraterrestrial communications. We learn a great amount of information by observing exoplanets when they transit their host star. For example, the radius of the planet and how long it takes to orbit its star can be determined with transit data. A civilization at or above our technological level would also realize that they can learn a lot about other planets through transits, so they may deliberately send a communication out in our direction when they know we can see their transit.

Thinking about Schelling points for SETI reminds us that there are also social scientific and psychological aspects to consider.

For more details, see Jason Wright’s Schelling Points page on his blog: <https://sites.psu.edu/astrowright/2017/07/09/schelling-points-in-seti/>

Activity 1 at the end of this chapter allows you to think about how to approach Schelling Points.

One way to identify an artificially created signal would be an extremely narrow bandwidth. There are limits to how narrow the signal from an astrophysical source can be so artificial signals could be distinguished in this way, although there is still some broadening of the signal that occurs as it travels to Earth (due to the relative motion of the extraterrestrial transmitter and our receiver on Earth).

The main signature of Earth’s technology is through our radio leakage. Earth’s radio and TV transmissions have been traveling away from the Earth for more than 80 years now — meaning any civilizations located up to 80 light years away from us could potentially have intercepted them — and we can add military radar and wifi signals to that list. Could we detect similar radio leakage from another civilization? Breakthrough Listen can get us close with the Meerkat, a precursor to the SKA (Square Kilometer Array).

Another possible way that a technologically advanced civilization might catch our attention is by changing the output of a natural radio source, such as a pulsar. An extremely advanced civilization could change the pulse rate — perhaps have the period flip back and forth between two different times — in a way that could not happen naturally.

Laser Pulse Beacons

Optical SETI searches for communication signals in the form of pulses of laser light. The pulses would need to be communicated in a way that we will notice, perhaps by having them come as bursts of prime numbers (two flashes, three flashes, five flashes, seven flashes, ...). This type of communication is a technosignature as the civilization sending the pulses must be advanced enough to have developed lasers.

Megastructures

In 1960, Freeman Dyson described a hypothetical construct that an advanced civilization capable of harnessing all of the energy from its star (a Type II civilization on the Kardashev Scale) would build. These “[Dyson spheres](#),” as they have come to be known, could be made of a network of materials that can absorb the star’s light and transform it into some other type of useful energy. A Dyson sphere could surround an entire solar system. Figure 6 shows an example of a Dyson sphere around Earth in our solar system — the radius of this megastructure would be 1 AU (astronomical unit)! Dyson recognized that a solid shell of material would not be possible to build so instead envisioned a ring of objects in a swarm surrounding the star. Dyson swarms are also frequently discussed, with the objects in the swarm sometimes comprising habitats that contain energy collectors.

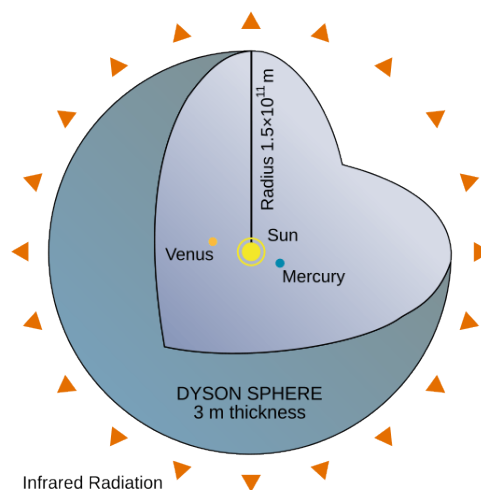


Figure 6 – A Dyson sphere. A Dyson sphere as shown here would only appear as infrared radiation, as all of the visible light would be blocked by the sphere. Credit: [Wikipedia](#), public domain

How could we detect such a megastructure? The light that is absorbed by the material making up the ring or swarm would be re-radiated as infrared radiation, just as sunlight absorbed by the Earth’s surface is re-emitted as heat. Thus, infrared telescopes could detect this “waste heat” as an otherwise unexplainable excess of infrared radiation. A star with a Dyson sphere around it could also appear to suddenly start flickering or have all of its visible light disappear. The sky can be monitored for this unusual type of star, which would be accompanied by the infrared waste heat.

Transiting Objects

We have already seen the power of using the [Transit Method](#) to detect exoplanets. The Transit Method works simply, by measuring the change in brightness of a star to look for dips caused when a planets passes in front of the star, and the results from NASA’s Kepler and TESS missions have been prolific. The same principle and techniques for finding exoplanets can be used to search for technosignatures. (The Transit Method can also be used to search for **exomoons**, which we discuss elsewhere.)

Megastructures

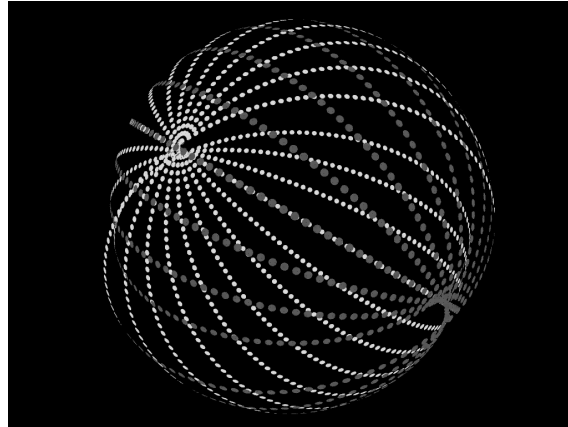


Figure 7 – A Dyson swarm. Credit: [wikimedia](#), CC BY-SA

If a technologically advanced civilization has built a Dyson swarm, this could show up as a series of dips in the brightness of a star. The details of the light curve would depend on how the Dyson swarm is constructed (Figure 7 shows an example of a swarm made of a series of polar rings) but we can easily imagine the brightness dips caused by the material passing in front of its star.

An excellent example of what a transit curve from a Dyson swarm might look like is seen by looking at the real light curve for the star KIC 8462852, also known as Tabby's Star for the astronomer, Dr. Tabitha Boyajian, who led the team that studied the star's behavior.

Tabby's Star

In 2016, a postdoctoral fellow at Yale University, Tabby Boyajian, published a paper on a strange light curve measured by Kepler. The signal was strong but irregular and unlike anything that had been seen before. The light curve for Tabby's Star was flagged by the citizen science project Planet Hunters, which presented the public with light curves from Kepler data to look for any anomalies. It worked! A paper draft sat on her desk for two years as the Planet Hunter team tried to reason out an explanation – the best idea was that this was a disintegrating asteroid or irregular dust cloud. Failing any certain ideas for the source of the signal, the paper was published in the hopes that someone else would have a better explanation. Almost immediately, Jason Wright at Penn State University suggested an idea: perhaps this was an alien megastructure, like the one suggested by Freeman Dyson. Of course, this idea was picked up in the popular press. Credible news stories with titles such as “I’m STILL Not Sayin’ Aliens. But This Star Is Really Weird” popped up immediately (this particular one is from [slate.com](#) and was written by astro-myth-debunker Phil Plait).

The light curve for Tabby's Star is consistent with a Dyson swarm. The figure below shows some follow-up data taken by an amateur astronomer in 2017-2018 and the large dips in brightness are easy to spot, with the largest dip occurring in March 2018.

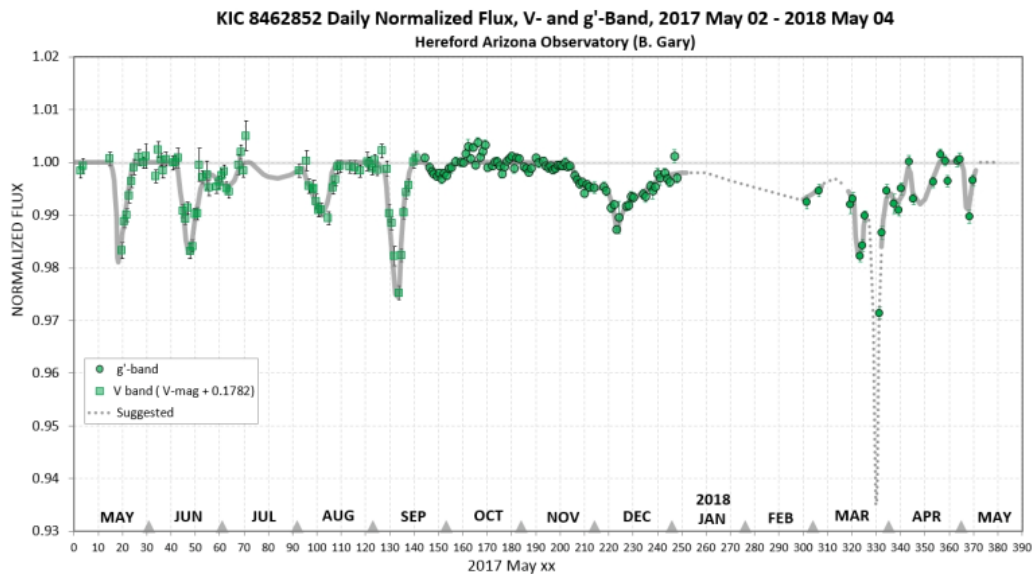


Figure 8 – The brightness of Tabby's Star over time. This data was collected for a year between 2017 and 2018. Irregular dips can be seen over time. Credit: [wikimedia](#), CC0

Once again, we note that this is certainly an extraordinary claim and extraordinary evidence is needed. Astronomers quickly applied Occam's Razor and identified several simpler, non-alien explanations to explain the light curve. An early idea was that there is a system of comets that occasionally occludes the light from Tabby's star. It is now widely accepted that the dips in Tabby Star's light can be explained naturally by dust but it is an interesting case study in how sensational claims are disproven. Dr. Boyajian gave a TED Talk about the discovery back in 2016:

Tabby's stare

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Video Credit: @TED

Structure Beacons

A Dyson swarm could reveal itself in a light curve found using the transit method. Another type of technosignature that could be detected this way is an object that is not round and has a shape that is designed to create a distinct signature in the brightness data. A couple of examples are shown here, from a 2005 paper that explored this idea:

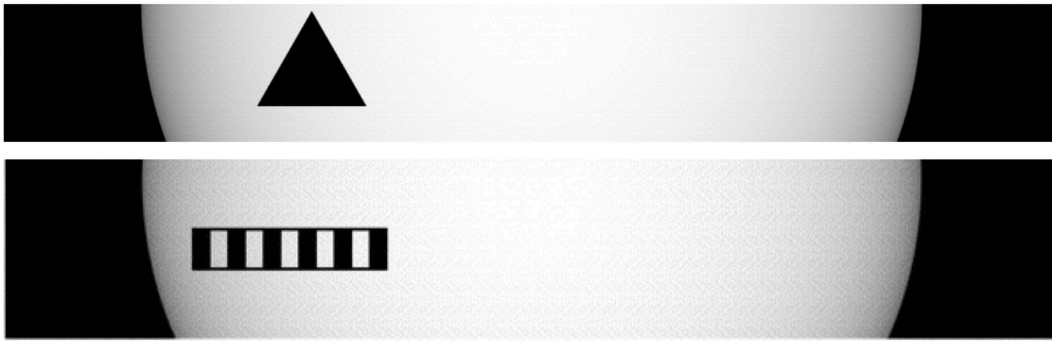


Figure 9 – Two different shapes, triangles and panels, for objects that an advanced technological civilization could place in orbit around their star to catch our attention. The light curves for these objects would be distinct from those created by a round object. Credit: [Astrophysical Journal](#)

The civilization could place the object deliberately in their transit zone so Earthlings would see it when we are observing their star. We can think of these deliberately placed objects as “structure beacons,” as opposed to laser pulse beacons. These structures could be light collectors, such as those in a Dyson swarm, or perhaps deliberate pulses in an attempt to cloak the civilization’s biosignatures.

Exobelts

The first artificial satellite launched into orbit around the Earth was the Soviet *Sputnik I* in 1957. Sputnik did not stay in orbit after completing its mission but humans have placed more and more satellites into Earth’s orbit since that time, and over 10,000 are orbiting Earth today. Not all of the satellites are active — some stay up after the mission is complete and these add to Earth’s growing collection of **space junk**. Objects orbiting Earth can be placed at different altitudes. Some objects, such as Space X’s StarLink internet satellites, are placed in low-Earth orbits. Others, such as weather satellites, are placed in geostationary orbits so that they orbit at the same rate that the Earth rotates and thus are “fixed” at one location over the Earth. These satellites are one of Earth’s technosignature and we can imagine that another advanced civilization could also have “belts” of satellites orbiting its planet.

If a belt of material around an exoplanet was sufficiently wide and far enough from the planet, it could be picked up using the transit method. Such **exobelts** have been speculated to be in geostationary orbits, placing them high enough to have a detectable signature in the light curve of the star the exobelt is around. Exobelts at geostationary orbit have been called Clarke Exobelts, as a reference to an article written in 1945 by science fiction writer Arthur C. Clarke that put forth the idea of using geostationary orbits for extraterrestrial communication.

Planet-Planet Occultations

Another artifact of our technological civilization that could be detected are the satellites we have sent out to explore our own solar system. Earthlings are in regular radio communication with satellites all over the solar system, from the rovers on Mars to the New Horizons orbiter in the Kuiper Belt, through the Deep Space Network. Perhaps another solar system is carrying out similar planet-to-planet communications with their own network and we could detect these transmissions. Note that the planets do not need to be in the habitable zone — we communicate with satellites orbiting Jupiter and on Mars, neither of which are habitable. The alignment needs to be such that from Earth, there is a straight line to the two planets as they both transit the star (these events are called planet-planet occultations). If you know exactly when these occultations happen, you can either search through existing data or schedule observations to see if any anomalous communication signals are mixed in.

Artifacts (SETA)

The search for extraterrestrial artifacts is sometimes referred to as SETA. The Earth has sent artifacts out to broadcast our existence. While the Voyager Gold Records and Pioneer plaques help to keep the public engaged in SETI, these are not expected to be collected and responded to by another civilization. What about artifacts sent our way from another civilization? These extraterrestrial artifacts of course could take on many different forms and could be intentionally sent for us or we may randomly pick up a piece of their technology.

Concept Check

List a few types of artifacts that an extraterrestrial civilization might send out. For each object, explain how it could be detected by Earthlings. If it cannot be detected, explain why.

Some ideas that have been posited include objects that are placed on the Moon or another place in our solar system, similar to the monolith in the film *2001: A Space Odyssey*, that will be activated when we reach a certain level of technology. The Earth's Lagrange points with the Sun or Moon could also be places that an alien artifact would be placed. Lagrange points are places of gravitational stability, so an object placed there would not drift away (this is why we place some of Earth's space mission crafts at these points). Perhaps a Bracewell probe could be parked at a Lagrange point and observing Earth.

A piece of alien technology could also drift into our solar system, intentionally or perhaps randomly (as a fanciful example, imagine a piece of the blown-apart Death Star in Star Wars hurtling away from its location in space). Interstellar objects can enter our solar system and two have been detected to date. One of them, 1I Oumuamua (the designation "I" stands for "interstellar") was previously discussed as an object that was briefly considered to be an alien artifact. However, extraordinary evidence to support this claim has not emerged.

The Cosmic Haystack

SETI is sometimes described as looking for a needle in the cosmic haystack. We can search for these alien needles in different ways but SETI is largely focused on identifying needles that are hidden in a star's light.

Having made the decision that radio is the most likely means of communication among intelligent civilizations, we still have many questions and a daunting task ahead of us. Shall we *send* a message, or try to *receive* one? Obviously, if every civilization decides to receive only, then no one will be sending, and everyone will be disappointed. On the other hand, it may be appropriate for us to *begin* by listening, since we are likely to be among the most primitive civilizations in the Galaxy who are interested in exchanging messages.

We do not make this statement to insult the human species (which, with certain exceptions, we are rather fond of). Instead, we base it on the fact that humans have had the ability to receive (or send) a radio message across interstellar distances for only a few decades. Compared to the ages of the stars and the Galaxy, this is a mere instant. If there are civilizations out there that are ahead of us in development by even a short time (in the cosmic sense), they are likely to have a technology head start of many, many years.

In other words, we, who have just started, may well be the "youngest" species in the Galaxy with this capability. Just as the youngest members of a community are often told to be quiet and listen to their elders for a while before they say something foolish, so may we want to begin our exercise in extraterrestrial communication by listening.

Even restricting our activities to listening, however, leaves us with an array of challenging questions. For example, if an extraterrestrial civilization's signal is too weak to be detected by our present-day radio telescopes, we will not detect them. In addition, it would be very expensive for an extraterrestrial civilization to broadcast on a huge number of channels. Most likely, they select one or a few channels for their particular message. Communicating on a narrow band of channels also helps distinguish an artificial message from the radio static that comes from natural cosmic processes. But the radio band contains an astronomically large number of possible channels. How can we know in advance which one they have selected, and how they have coded their message into the signal?

Table 1 summarizes some of these factors that scientists must grapple with when trying to tune in to radio messages from distant civilizations. Because their success depends on either guessing right about so many factors or searching through all the possibilities for each factor, some scientists have compared their quest to looking for a needle in a haystack. Thus, they like to say that factors such as those in Table 1 defines the *cosmic haystack problem*.

Table 1 – The Cosmic Haystack Problem: Some Questions about an Extraterrestrial Message

Factors
From which direction (which star) is the message coming?

Factors
On what channels (or frequencies) is the message being broadcast?
How strong is the signal (can our radio telescopes detect it)?
Can we even recognize a message from a completely alien species? Might it take a form we don't at all expect?

Although the cosmic haystack problem seems daunting, many other research problems in astronomy also require a large investment of time, equipment, and patient effort. And, of course, if we don't search, we're sure not to find anything.

How much of the cosmic haystack have we searched?

There is a lot of hay (stars) in the cosmic haystack so to estimate how much of it has been searched for alien needles we need to identify some parameters that we can use to quantify this immense space in a robust way. It is reasonable to confine the search to our home galaxy, the Milky Way, for two reasons. First, the finite speed of light prohibits any meaningful intergalactic communication. Also, if we accept the Copernican principle, our galaxy is representative of any other galaxy in the universe, so life in the Milky Way would mean life in other galaxies. We can further confine the haystack to just focus on searches in the radio part of the electromagnetic spectrum. This is where most previous searches have looked and Breakthrough Listen also focuses on the region between 1-10 GHz. In 2010, Jill Tarter estimated what fraction of the haystack has been searched and she reported her result with this memorable analogy:

"So our search of the 9-dimensional haystack is equivalent to sampling about 1.6 cups of water from the Earth's oceans. If you were looking for fish instead of extraterrestrial intelligence, I don't think that you would conclude that there are no fish in the ocean after this meager sampling!"

Jason Wright at Penn State published an updated estimate in 2018 and he found that the 1.6 cups could be increased to a large hot tub or small swimming pool. On the one hand, this is progress! On the other, we still have a lot of space to search.

Key Concepts and Summary

The goal of SETI searches is the detection of technological signals from other civilizations. Recognizing that simple organisms were altering the atmosphere of Earth for billions of years before our radio telescopes appeared, astronomers can cast a wider net in the search for life by looking for the impact that life might have on the spectra of atmospheres of other planets – so called biosignatures. For example, the combination of strong oxygen markers together with methane is a signature of chemical disequilibrium on Earth that requires ongoing biological respiration. There could also be technosignatures that are being transmitted by pulsed lasers or transiting megastructures that harvest more light from their host stars. We have only sampled a tiny bit of the universe – the equivalent of sampling 1.6 cups of water from all of Earth's oceans.

Review Questions

Summary Questions

1. What is a biosignature? Give some examples of Earth's biosignatures.
2. How do SETI scientists search for biosignatures in planetary atmospheres?
3. What is a technosignature? Give some examples of Earth's technosignatures.
4. What types of technosignatures do scientists search for on exoplanets? Give at least three examples and explain each one briefly.
5. What is the "Cosmic Haystack"?
6. How much of the cosmic haystack has been searched so far? Explain your answer.

Exercises

1. Think of our Milky Way Galaxy as a flat disk of diameter 100,000 light-years. Suppose we are one of 1000 civilizations,

randomly distributed through the disk, interested in communicating via radio waves. How far away would the nearest such civilization be from us (on average)?

Activities

1. **Group activity.** Read the box on Schelling points. Imagine that as a class, you are asked to come up with a time and place to meet on your college's campus. You cannot communicate with each other at all. Individually, come up with a time and place for the meeting. Now compare all of the suggested times and places. Do they all match?
2. **The Cosmic Haystack.** Table 1 lists some factors that must be considered before carrying out a SETI search in the radio part of the electromagnetic spectrum. Can you think of any additional questions that need to be considered, in order to maximize the search strategy?

6.4: The Drake Equation and the Fermi Paradox

Is anybody out there?

Learning Objectives

By the end of this chapter, you will be able to:

- Explain why the Drake Equation was formulated and what it seeks to estimate
- Describe the main factors that go into the Drake Equation
- Discuss the “Fermi Paradox” and some reasons why no other life has been detected
- Come up with and justify your own opinion on whether or not there is other life in the cosmos

The Drake Equation

In 1961, just one year after Frank Drake’s Project Ozma was completed, a conference was held at the Green Bank Observatory in West Virginia on the topic of “Extraterrestrial Intelligent Life,” specifically if the ongoing searches for extraterrestrial intelligence had any chance of actually detecting a real ET signal. A small number of experts were brought together for the meeting, including Carl Sagan, who was just 27 at the time, and John Lilly, who studied dolphin language and communication. Because of Lilly’s expertise, the group called themselves “The Order of the Dolphin.” To help organize an agenda for the meeting, Drake put together an equation to estimate how many intelligent civilizations should exist in the Milky Way galaxy; the equation contained several terms that Drake felt were key to considering this question.

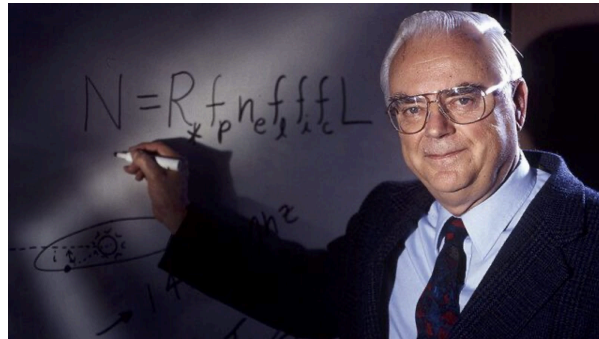


Figure 1 – Frank Drake and the Drake Equation. Credit: SETI Institute.

The equation, which is presented below, provided enough discussion points for the entire three-day conference and remains a useful heuristic tool in discussions on the possibility of other life in the universe. Seth Shostak at the SETI institute characterizes the Drake equation as a map, not a destination. The equation is made up of seven different terms that can be multiplied together to provide an estimate of the number of civilizations that we should expect to hear from. The terms f are fractions and their values can range from 0 to 1, where 1 is 100%.

$$N = R_* \times f_p \times \eta_e \times f_l \times f_i \times f_c \times L \quad (6.4.1)$$

where

- N is the number of communicative civilizations that we could detect
- R_* is the rate of formation of stars like the Sun in our Galaxy
- f_p is the fraction of stars with planets
- η_e is the number of these planets that have the right conditions for life
- f_l is the fraction of these planets where life develops
- f_i is the fraction of planets with life where intelligence emerges
- f_c is the fraction of intelligent life that develops communicating technology
- L is the lifetime of communicating civilizations.

The video below by SETI pioneer Jill Tarter summarizes the terms in the Drake Equation and makes some points worth keeping in mind as we consider the various terms. As she points out, the term L is often considered the most important one to consider in making this estimation. She also reiterates that there is no “right answer” to the Drake equation, unlike other scientific equations that you have used.

Note

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



After watching the video above, consider these questions:

- How could we justify approximating the Drake Equation to be $N = L$?
- Can we ever have a definitive answer to the Drake Equation? Explain why or why not.

Unless SETI succeeds or we can show that the Earth is the only life out there, N remains uncertain.

As the attendees of the Green Bank conference did in 1961, let's consider each term separately and then attempt to come up with an estimation. We will provide some provisional numbers, but many of them are merely educated guesses (some more educated than others).

Stars

We can make a rough estimation of the star formation rate of the galaxy, R_* , by considering what we know about the galaxy today. Currently, the galaxy has on the order of 100 billion stars. We also know the galaxy to be approximately 10 billion years old. From this, we find that on average stars have been forming at a rate of 10 stars per year.

$$R_* \tag{6.4.2}$$

This is a very simplistic estimate. It does not take into account how many of these stars may be suitable for life, such as short-lived O or B type stars, or whether the rate of star formation has been constant over the lifetime of the galaxy. However, this value gives a well-educated estimate. A more conservative estimate that takes into account the different types of stars and how often they form leads to a lower number of $R_* = 1$ star per year.

Planets

Thank to discoveries over the past decade, we now have a good estimate for the fraction of stars with planets. This was completely unknown at the time Frank Drake first wrote down his equation — recall that the first exoplanet was not detected until the 1990s. The discoveries from the *Kepler* mission provide statistical evidence that essentially every star has one or more planets. When you look up at the night sky, there are actually more unseen exoplanets there than there are stars. We can therefore assign a value of 1 for f_p .

Earth-like planets

The next parameter, η_e , is the number of planets that have the conditions for life to develop. This brings to mind the concept of the habitable zone (HZ) around a star — if we can quantify how many rocky planets typically fall into the HZ around their star, this would be a good starting point. However, we need to keep in mind that some of the most promising places to find life in our solar system are the moons Europa, Enceladus, and Titan, all of which are not in the Sun's HZ. Also, just because a planet falls into the HZ around its star does not mean it can support life; factors such as flares and tidal locking can impact the potential habitability of a planet. Astronomers have nonetheless used Kepler data to extrapolate out from the parameter space where Earth-sized planets

were detected and estimate that the number of habitable Earths ranges from 0.5 to 3 per system.

Life

The term f_l asks us to consider the fraction of Earth-like planets in the habitable zone where life of *any* kind (including single-celled microbes) emerges. This term must be non-zero, because there exists life on Earth but we have yet to find life anywhere else. What we do know is that the Earth formed about 4.56 billion years ago. We have firm evidence of life from stromatolites at 3.5 billion years ago, which must have been preceded by less complex organisms. These numbers indicate that life formed relatively quickly on Earth, even though it is not yet clear to us how. If we invoke the Copernican Principle here, then Earth should not be special in this regard, but we must contend that we just don't know. We can also take the results of Miller-Urey-type experiments as evidence in favor of life developing naturally given the right conditions, but those results did not find life itself, just the building blocks (amino acids).

We can infer that perhaps life appears between 10% and 100% of the time on habitable planets, but you should make your own guess for this fraction. What is your justification?

How do we know: what will intelligent signals look like?

Given the [difficulties we have had defining "life"](#), it is perhaps no surprise that there are even more difficulties defining "intelligence". Many assumptions have been made about what intelligence implies and how it can be measured which are increasingly criticized by social scientists and philosophers. One of the first steps that we take to try to define intelligence is to distinguish between lifeforms that are not intelligent and those that are, but attempts to make these clean distinctions have run into problems. For example, some attempts at defining human intelligence as distinct from other lifeforms invoked consciousness or sentience (roughly knowledge of one's self identity) as a criteria. This requires, however, a theory of mind that assumes these aspects of our existence are generalizable and identifiable in other life forms in ways that are difficult to measure. While neuroscientists are actively investigating consciousness in new and clever ways, there is no foolproof way to determine whether and how any given lifeform is conscious at this time. Using that as a criteria for intelligence, then, will possibly cause more problems than it will solve.

At one time it was assumed that what distinguished human intelligence from that of other animals was an ability to communicate through language. Nevertheless, anyone who has spent significant time with a dog knows that animals are capable of communication in very sophisticated ways. The late Koko the Gorilla had a sign-language vocabulary in excess of 1000 signs by some measures, although whether that constituted learning a language and to what extent her human handlers "projected" communication onto Koko's interactions is of enough concern to call into serious question the claims that she "learned" sign language. Controversies over intelligence and, in particular, psychometric testing of human beings, have also highlighted the problem of under what research contexts different definitions for intelligence in individuals are useful. IQ tests, for example, may highlight more about what is valued in a particular societal contexts instead of probing some manner of objective measure of whom is more intelligent than whom.

What those interested in the Search for Extra-Terrestrial Intelligence (SETI) use as a definition for intelligence is different from that which human or biological researchers in the subject use. Astronomers assume that the ability to produce similar technology to our own to enable interstellar communication is the mark of a group of intelligent lifeforms. As far as we know, the only lifeforms capable of such on Earth are humans. Even if a particular individual human does not know how to construct interstellar communication themselves, the knowledge and the skills were possible for us as a group to develop and create as demonstrated by the existence of the technology itself. Evolutionary biologists still debate when humanity first had the intellectual capacity to develop the technology of our modern world, but most would agree that early humans from at least 100,000 years ago exhibited the same brain-size, tool-making abilities, and language processing that we have today and that seem to be necessary to accomplish this. Research is ongoing in attempts to answer the question of how smart our ancestors were in comparison to us which may have implications for how we ask these questions in the first place.

But why should interstellar communication be the hallmark of intelligence?

If we want to answer the question "Are we alone?" we need to find a way to discover the other intelligent lifeforms. If any intelligent, technical societies have arisen, as has happened on Earth in the most recent blink of [cosmic time](#), how could we make contact with them?

This problem is similar to making contact with people who live in a remote part of Earth. If students in the United States want to converse with students in Australia, for example, they have two choices. Either one group gets on an airplane and travels to meet the other, or they communicate by sending a message remotely. Given how expensive airline tickets are, most students would probably select the message route.

In the same way, if we want to get in touch with intelligent life around other stars, we can travel, or we can try to exchange messages. Because of the great distances involved, [interstellar space travel would be very slow and prohibitively expensive](#). The fastest spacecraft the human species has built so far would take almost 80,000 years to get to the nearest star. While we could certainly design a faster craft, the more quickly we require it to travel, the greater the energy cost involved. To reach neighboring stars in less than a human life span, we would have to travel close to the speed of light. In that case, however, the expense would become truly astronomical.

Intelligence

As we saw above when considering the fraction of habitable planets that go on to develop life, the estimates get increasingly difficult as we move into the biological factors (f_l , f_i and f_c) in the Drake Equation. As emphasized in the box above on “How do we know what intelligence is?”, it is hard to describe, hard to quantify, and hard to detect intelligence. Besides the example of intelligent life on Earth, we have no other concrete evidence towards a true value for the fraction of life that becomes intelligent. We could again use what happened on Earth as a general guiding example – a sort of Copernican Principle view for biology. From this viewpoint, Earth is not special and any life that forms should eventually develop intelligence.

One way to approach the question of intelligence developing is to ask when intelligent life first appeared on Earth. We know that complex life on Earth requires energy, which requires efficient metabolism. On Earth this takes the form of aerobic respiration, and the Cambrian explosion, about 540 million years ago, marks the time when complex life emerged. If we take 4 billion years ago as the time when the first life appeared on Earth, it then took 3.5 billion years for that life to evolve to be complex enough for intelligence. We can then estimate f_i as

$$f_i = \frac{0.5 \text{ billion years}}{4 \text{ billion years}} = 0.125$$

Inherently, right or wrong, we are guessing that life everywhere will evolve and that statistically perhaps 13% of the habitable planets that we find with life will have life forms that evolved to become intelligent.

However, we have no true understanding of how typical our planet is. We do not have a sense of whether the evolution of complex life is rare or inevitable for all Earth analogs. An upper bound might be 1, meaning all planets with life will eventually evolve the complexity that leads to intelligence or it might happen on one of a million inhabited worlds. That’s a big range!

Technology

The next term, f_c , tries to capture another attribute that is hard to estimate. On inhabited planets with intelligent life, how often will that life develop technology so that it can communicate across the galaxy? We are technological adolescents — intelligent life on Earth has had communication technology for only about 100 years. That is a tiny fraction of time for intelligent life on Earth: 100 years out of 200,000 years that our species, *Homo sapiens*, have existed.

While we can easily identify other species on Earth that are intelligent, such as dolphins, bats and cats, it is a stretch to imagine any of these other intelligent species developing the capability for interstellar communication, suggesting a low probability for f_c . A familiar argument by now in favor of a higher probability is the Copernican Principle, but this argument must be tempered with the reality that we simply do not have enough data to make any meaningful estimates of f_c .

Lifetime of an Advanced Civilization

Once a civilization has become advanced enough to develop science and the technology to build radio telescopes and transmitters, how long will it last? In the discussion of f_c above, we noted that Earth has only been in this category for 100 years. Technology has rapidly advanced in that century, sometimes in ways to extend the favorable conditions for life on Earth and other times in ways that put the longevity of our planet at great risk. Biotechnology and genetic engineering are at a stage where we can combat and eradicate many diseases. On the other hand, the average temperature of the Earth is increasing exponentially

and habitats are being destroyed as a result and some species need to relocate to avoid extinction. AI is now commonplace and this is also seen by some as positive and others as detrimental to the survivability of humanity.

The global political situation in 1961, when the Drake Equation was first presented, was tense as the Cold War put the Earth at increased risk for nuclear destruction. The risk of catastrophic warfare still looms more than 50 years later.

Considering these factors, a pessimistic take on L would be maybe 100 more years or less.

Estimating N

Finally, the finale. We can collect our musings from above to calculate a value for N , the number of communicating civilizations that we could detect. Gather all of your values for each term in the Drake Equation and then use the calculator in the box below (or simply carry out the math on your own — it is just multiplication) to get your value of N .

Once you have all of your values ready, you can use the Drake Equation calculator provided by PBS: <https://www.pbs.org/lifebeyondearth/listening/drake.html>

The Fermi Paradox

“Where is everybody?”

Physicist Enrico Fermi is credited with posing this question in 1950 during lunch with some colleagues — all having worked on the Manhattan Project with Fermi — at Los Alamos National Lab in New Mexico. There had been a recent increase in reports of UFO sightings and this led Fermi to pose the question and ponder why we haven't heard from any other extraterrestrial civilizations. The question is not so much a paradox per se but a prompt for discussions on the challenges of detecting other life in the cosmos. The question becomes a paradox if we conclude that the universe is filled with life...and we should see it everywhere but don't.

After considering the various terms in the Drake Equation, we have some context to think about reasons why we have not made contact. Of course, one easy solution is that we truly are alone and the only instance of life in the universe. While a generally sobering thought to contemplate, we must acknowledge that this is a possibility.

They're Out There

Let's consider the Fermi Paradox if we assume that other intelligent extraterrestrial civilizations exist in our galaxy. You can keep in mind your estimate of the number of such civilizations from the Drake Equation above as we make this assumption (for example, if you estimated a very low value for N , then you may consider this a poor assumption). If other advanced life is out there and this life developed much earlier than our technological civilization on Earth, we might expect that the galaxy would be completely filled with evidence of this more advanced civilization. For example, they could have sent out self-replicating probes to all of the planets in the galaxy, either to see what's there or to mine resources.

Before moving on and reading a few possible resolutions, take some time to think of your own resolutions to the Fermi Paradox in this case, as suggested in the box below.

Assume that there are a large enough number of other technologically advanced civilizations in the galaxy that we should expect to have heard from at least one of them by now.

Make a list of reasons as to why we have not heard from these civilizations yet (or may never hear from them).

We Don't Recognize Them

One possibility contends that other advanced life may not communicate in the same ways that we do on Earth. Technology on

Earth has exploded in the century since the first radio transmissions were sent out, and it seems possible (as AI flourishes) that our communication tools will continue to advance exponentially. If another civilization has a few billion years jump start on technology, they may be communicating in a way that we simply don't recognize. Consider a less advanced but still intelligent species on Earth, such as bats. We simply cannot communicate in any meaningful way with bats about human society, and the world as experienced by a bat is extremely limited when compared to a human's understanding of the cosmos. A similar argument holds that life outside of the Earth may develop and evolve completely differently than it did on Earth, perhaps based on an element other than carbon, and we simply are unequipped to perceive it. For example, life on Saturn's moon Titan could use methane or ethane instead of water as the liquid to facilitate chemical reactions inside of cells.

Interstellar Travel Is Too Difficult

It is easy to assume that a civilization that is much more advanced than our own would have figured out how travel to other civilizations or, at the least, to be able to communicate their presence effectively. But what if this assumption simply does not hold up? While we currently have the technology on Earth to send a probe to the nearest star in our life time (Breakthrough Starshot's design for a light sail to be sent to Proxima Centauri), this still does not get us very far. Perhaps this is a barrier that few or no civilizations will be able to overcome. This idea that there is a barrier that any life must overcome before reaching an advanced stage, such as a Type II or Type III civilization on the Kardashev scale, falls into a general category of ideas that imagine there is a "great filter" (or filters) that life must overcome before it can continue to evolve. If this evolutionary bottleneck still lies in the future for humans, we can imagine a very pessimistic outlook for our species evolving further and developing more advanced technology.

Even without traveling to another planet, two-way communication is always restricted by the speed of light. If a civilization just 100 light years away from Earth sends us a message that we receive and can interpret, we would need to wait at least 200 years for a response. A wormhole — an interstellar or even intergalactic portal — could allay these time scales, but there is no evidence that wormholes exist.

Some additional ideas on resolving the Fermi Paradox are discussed in the video below.

The Fermi Paradox

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Video Credit: @TEDEd, CC BY-NC-ND 4.0

Are We Alone?

The Drake Equation and the Fermi Paradox both provide ways to narrow the answer to this question based on the astronomical and biological data in hand. In 1966, when Sagan and Shlovskii published their book *Intelligent Life in the Universe*, no exoplanets or protoplanetary disks had been detected and f_p was unconstrained. Today, we can say with confidence based on evidence from the Kepler and TESS missions that f_p is close to 1. We have more data today to support the first three terms of the Drake Equation — R_* , f_p , η_e and — but the biological terms remain just as uncertain today as in 1966. Sagan estimated N to be 10^6 , or 1 million, in the 1960s and he acknowledged that the most sensitive term is L , the lifetime of a technologically advanced civilization. He looked at L both optimistically and pessimistically and then took the average to arrive at the value of $N = 10^6$ other civilizations out there. This same guesstimate holds today.

We have not detected any evidence of other life in the universe. However, we are currently poised to be able to make a meaningful statement of whether or not other life is out there in the next decade. This is for two main reasons. First, SETI is now exploring more stars at the widest range of frequencies ever thanks to the Breakthrough Listen initiative. The sheer number and volume

of stars surveyed by Breakthrough Listen should yield a statistically significant sample of results to draw meaningful conclusions from. Second, the James Webb Space Telescope allows us to point directly at a potentially inhabited planetary and study its atmosphere for biosignatures and technosignatures. Next-generation telescopes like the Habitable Worlds Observatory will allow Earth-sized exoplanets to be imaged.

Key Concepts and Summary

The Drake equation is a way of estimating the probability that life exists elsewhere. First written down by Frank Drake in 1961, the Drake equation includes factors like the rate of star formation, the number of planets per stars, the fraction of those planets with the right conditions for life to emerge and evolve into intelligent, technological civilizations. The final term is the lifetime of those communicating civilizations – a recognition that species and technology are ever changing and not permanent. With all of the searches that have been carried out, the Fermi paradox asks why we have not found anything yet: Where are they?

Review Questions

Summary Questions

1. Provide your own estimate of the Drake Equation giving an argument for why you think each value should be what it is.
2. In examining the Drake Equation, some enterprising commentators have said that optimists need to hope for big values of either η_e or L or both. Explain why, mathematically, this is so.
3. What is the “Fermi Paradox”? Is it really a paradox or would a word other than paradox be more appropriate? Explain your reasoning.
4. Do you think life will be discovered outside the Earth within your lifetime? What about intelligent life? Support your argument with explanations from this textbook or other reliable resources you have found on your own.

Exercises

1. Some SETI theorists argue that $N = L$. What does this imply about the other values in the Drake Equation?
2. Frank Drake predicted that $N = 10^9$. Think of our Milky Way Galaxy as a flat disk of diameter 100,000 light-years with the intelligent groups spread evenly throughout the disk. If Drake is correct, how long would it take for us to send a signal and receive a reply from the nearest intelligent group?

6.5: Speculations and Skepticism

intro para

Learning Objectives

By the end of this chapter, you will be able to:

- Critically assess claims, especially claims in news media
- (Discuss the psychology of delusional beliefs)

Since we have yet to find scientific evidence for life outside of the Earth, much of the study of astrobiology is necessarily speculative. Perhaps because of this, there has been and continues to be a tendency for people to believe that evidence does exist for life from beyond Earth interacting with us.

This tendency can be seen in the mythologies associated with the skies across many human cultures. Intelligence and life was assumed to be part of the celestial world, with gods, monsters, and the spirits of once-living terrestrial creatures complete with agency and personalities that could interact with humanity. Constellations, planets, and stars were the first “extraterrestrial life forms” that human beings assumed existed.

With an adoption of scientific perspectives on how the natural world operates, such enchantment of the natural world was eschewed. The planets, stars, and other phenomena we observe with our telescopes has up to now exhibited no signs of magic, intelligence, or life. However, a somewhat large contingent of humanity takes issue with this lack of evidence and proposes, contrariwise, that there is in fact ample evidence for such. This chapter aims to examine some of these claims in some detail.

Before proceeding much further, however, the evidence that will be considered in this chapter will be, in most cases, rejected as problematic, incomplete, or even completely incorrect. How does anyone decide this is the case? Ultimately, you have to use your own judgment to weigh evidence, and this includes the scientific evidence that has thus far been presented in this book.

Skepticism

One possible approach to evaluating claims is that of skepticism. Skeptics, as those who adhere to this position are called, adopt a set of standards for evidence used to hopefully distinguish fact from fantasy. The nature of this position is, however, that there will be people who disagree with the skeptical perspective. If one person argues that a phenomenon exists, the skeptical position is to propose that unless there is convincing and verifiable evidence that can be measured, the phenomenon is assumed not to exist. This type of assumption is called the null hypothesis.

What standards of evidence would be enough to have a skeptic reject the null hypothesis? There is no firm demarcation, and this is likely why there is often controversy surrounding these subjects especially within the popular, philosophical, and sociological discourse. Nevertheless, surveys of experts and scientists tend to show remarkable agreement as to the lack of evidence for a lot of the ideas we will be discussing, according to some general principles of “skepticism” including a famous dictum often attributed to the late Carl Sagan, “Extraordinary claims require extraordinary evidence”. What follows is one possible set of criteria that attempt to offer a means to determine which claims are supported by evidence provided by the late Carl Sagan.

Baloney Detector Kit

In his 1991 book, *Demon-Haunted World*, Sagan attempts to provide his readers with a set of tools that can be used to indicate concern with certain arguments. These tools are not meant to be foolproof methods to determine with absolute certainty whether a claim is true or false, but the hope is as one starts to evaluate more and more claims, the “Baloney Detector Kit” will work to help steer the user in the direction of being able to identify which claims are worthy of more careful consideration and which can be dismissed more quickly. Entertaining every claim is an exhausting endeavor, and with limited time such techniques can be used to move on to more productive avenues of investigation. Sagan writes that the tools in this kit are the following principles, quoted here in full:

1. Wherever possible there must be independent confirmation of the “facts.”
2. Encourage substantive debate on the evidence by knowledgeable proponents of all points of view.
3. Arguments from authority carry little weight — “authorities” have made mistakes in the past. They will do so again in the future. Perhaps a better way to say it is that in science there are no authorities; at most, there are experts.
4. Spin more than one hypothesis. If there’s something to be explained, think of all the different ways in which it could be explained. Then think of tests by which you might systematically disprove each of the alternatives. What survives, the hypothesis that resists disproof in this Darwinian selection among “multiple working hypotheses,” has a much better chance of being the right answer than if you had simply run with the first idea that caught your fancy.
5. Try not to get overly attached to a hypothesis just because it’s yours. It’s only a way station in the pursuit of knowledge. Ask yourself why you like the idea. Compare it fairly with the alternatives. See if you can find reasons for rejecting it. If you don’t, others will.
6. Quantify. If whatever it is you’re explaining has some measure, some numerical quantity attached to it, you’ll be much better able to discriminate among competing hypotheses. What is vague and qualitative is open to many explanations. Of course there are truths to be sought in the many qualitative issues we are obliged to confront, but finding them is more challenging.
7. If there’s a chain of argument, every link in the chain must work (including the premise) — not just most of them.
8. Occam’s Razor. This convenient rule-of-thumb urges us when faced with two hypotheses that explain the data equally well to choose the simpler.
9. Always ask whether the hypothesis can be, at least in principle, falsified. Propositions that are untestable, unfalsifiable are not worth much. Consider the grand idea that our Universe and everything in it is just an elementary particle — an electron, say — in a much bigger Cosmos. But if we can never acquire information from outside our Universe, is not the idea incapable of disproof? You must be able to check assertions out. Inveterate skeptics must be given the chance to follow your reasoning, to duplicate your experiments and see if they get the same result.

In the discussions that follow, try to see where these tools can be applied to the evidence that is provided. It is important to understand that these tools do not say “yes” or “no” in answer to the question, “Is this claimed phenomenon real or not?” The hope is rather to provide enough tools so that the merits of the claims can be evaluated and the reader can make up her or his own mind.

Unidentified Flying Objects

One of the first subjects that students are interested in exploring is the question of whether spacecraft from other intelligent groups are visiting Earth. An extensive catalog of stories attempting to demonstrate evidence for this has been assembled by various investigators often termed “ufologists” as a way of identifying that they study “UFOs”. The term “UFO” is an acronym that was coined by Air Force Captain Edward J. Ruppelt in the 1950s to stand for the words “unidentified flying object”. At the time, this was intended to be a neutral indicator of the reports: most people who believe they have seen a UFO report that it is flying in the sky, that it is a physical object that can be seen or tracked by, for example, radar, and that the observer does not know what the object is making it “unidentified”. Over the decades, the term has become so closely associated with the claim that there are extraterrestrial intelligent beings visiting the Earth, that now many associate UFO with “alien spacecraft”.

The official investigations of UFOs first happened in a coordinated fashion after World War II. At the time, there was heightened interest in threats from flying objects of human origin. During the Second World War, aerial technology advanced quickly to jet-propelled aircraft and rocket artillery and in the United States a Civil Air Defense Corps was inaugurated to monitor the skies for incoming enemy threats. The culture of reporting unusual aerial phenomena to the authorities was well-established by the time UFO investigations sponsored by the United States Air Force first began.

Under a variety of monickers, after assembling a team of investigators, the US Air Force undertook a number of projects to investigate the reports and determine whether they could be explained or whether they constituted a threat to the national security of the United States. Over the course of nearly a decade, a number of reports came back indicating consistently that some sightings could be identified as being caused by known phenomena and the ones that could not be explained did not rise to a level such that they could conclusively point to any new phenomena beyond the prosaic explanations that identified the other sightings. Since there was nothing conclusive that could be ascertained from the reports the investigations were ended so as to avoid prolonging a wild goose chase.

Since that time, formal studies of UFOs have been only intermittently conducted by governments and the scientific community has all but ignored such ideas. Ongoing claims that evidence for UFOs exist continue to be made, but so far there has not been the acceptance that these extraordinary claims have generated any extraordinary evidence.



Figure 1 – Scene from H. G. Wells' *War of the Worlds* depicts the invasion of the Martians with spacecraft not dissimilar to UFO shapes later claimed in various sightings.

Popular interest in UFOs continues in part due to such stories featuring in entertainment and in journalism. H.G. Wells in 1897 wrote *War of the Worlds* which presented a scenario of intelligent lifeforms from Mars invading the Earth. Likely, Wells was influenced by the recently published book by Percival Lowell entitled, *Mars* where, erroneously, Lowell reported on the observations of Giovanni Schiaparelli and others who argued that there was a network of canals that could be seen on the planet. This inspired considerable speculation that there could be flowing water and life on Mars, but later work revealed that these observations were entirely spurious.

At around the same time, humanity began experimenting with flight starting with airships that used hydrogen gas to achieve the necessary lift. A number of mystery airships were reported in the media with a few newspaper reporters uncritically claiming that they were made by lifeforms from other worlds who had used them to travel to Earth.

Reports of UFOs began to proliferate after a media sensation surrounding the claims of a pilot named Kenneth Arnold who in 1947 who reported seeing nine aircraft that appeared to him to move in a way that was far different than any aircraft he had seen until that point. He argued that such objects appeared to be moving like “saucers skipping across a pond” and the term “flying saucer” was soon attached to the unknown objects. On the basis of his personal observations, Arnold was convinced that aircraft were visiting Earth and were developed by an advanced extra-terrestrial intelligent beings. He consistently argued this for the rest of his life.

Not long after this, the now famous Roswell Incident occurred where a New Mexico rancher saw a balloon crash into a field and wreckage was recovered. The local news story complete with photographs and a visit by the officials from the Army Air Force who removed the physical evidence and eventually explained the event as a weather balloon crash although it was actually a nuclear testing monitoring balloon. The event might have disappeared into obscurity, but interest was revived in the late 1970s through the 1990s when a number of claimants began to argue that there had been recovery of extraterrestrial artifacts including lifeforms. The culmination of this sensation was the production of a so-called “Alien Autopsy” video which was broadcast nationally in 1995 and claimed at the time to be authentic footage, though the filmmaker years later would clarify that it was a reproduction of what he claimed to remember.



Figure 2 – This scene from a New Mexico museum dedicated to ufology is a reproduction from the 1995 “Alien Autopsy” movie. At the time, the footage was claimed to have smuggled out of Area 51, but later the filmmakers admitted to having staged a recreation of what they said were events they witnessed.

Identifying UFOs

Claims of visitations by intelligent lifeforms from beyond the Earth continue to be made in news stories, science fiction, and popular media. The very identification of a sighting of an “unidentified” object makes attribution difficult. The null hypothesis is that such observations have prosaic explanations and such explanations fall into four categories: natural phenomena, human technology, delusions, and hoaxes.

Natural Phenomena

A number of natural phenomena have confused observers into believing they were due to fantastical technology or the supernatural owing to peculiar characteristics, circumstances, or simply not knowing how the phenomena appears in nature.

A number of astronomical objects have been reported as “flying” in ways that observers thought looked like they were piloted by intelligent lifeforms, but this is not the case. Famously, the planet Venus is often puzzling for those who are unaware that it is the third brightest natural object in the sky and can rival the brightness of plane landing lights. In dark sites when there is no moon, Venus is bright enough to cast shadows. As a planet, it is also not a point source and can therefore be distinguished from stars because it exhibits less twinkling than stars exhibit as the random fluctuations in the atmosphere that cause twinkling stars are washed out over the larger apparent size of the planet. Venus is never far from the horizon, giving it the appearance of flying closer to the ground. When it is in the evening sky, Venus sets at approximately the rate of the Sun which makes it appear to be moving through the sky or perhaps landing at some distance. When Venus is very low in the sky, the same atmospheric refraction that causes sunsets can affect the color of Venus on short timescales, making it appear to be flashing different colored lights.



Figure 3 – A bolide such as the one shown here is a rare sight, but some who see such things not knowing that meteors can be extremely bright report seeing a UFO.

Some transient astronomical phenomena include meteors or bolides which appear as bright streaks and, in some instances, smoke trails in the sky. These are the visible signs that a rocky body from space has entered the Earth's atmosphere on a collision course. Shooting stars can be seen almost any night, but some points during the year there are many more of them in so-called "meteor showers" at a point in the Earth's orbit where it passes through the debris field of a comet. The largest fireballs are unusual, but can be extremely bright and happen high enough in the atmosphere to be visible for hundreds of kilometers. Occasionally, these meteors are large enough to cause damage when they hit the ground as meteorites or propagate shockwaves and sonic booms if they explode in the atmosphere.

Birds are, of course, flying objects and while most people think they can readily identify birds, some birds are large enough to confuse the perspective of the observer. One suggestion that has been made to explain Kenneth Arnold's sighting was that he was observing a flock of pelicans. These large birds can fly fairly high and if they are not identified as birds, it is possible to misjudge their size. Arnold proposed that the objects he saw were the size of aircraft and that they moved very quickly. Pelicans are approximately the size of a small human and if they are mistaken for larger objects, the person seeing them may believe that they are farther away than they actually are as a large object can have the same angular size as a small object when it is at a further distance. What is more, a closer object will appear to move across the field of view at a faster rate than one that is further away, so it has been suggested that Arnold's estimate of the speed of the objects could have been misjudged as being much faster if he assumed the objects were further away than they were.

A commemorative chair from Flatwoods, WV meant to depict the Flatwoods Monster. The figure has a red head with large yellow eyes, a silver color, and green body.

Figure 4 – A chair in Flatwoods, WV commemorates the UFO sighting and encounter with the Flatwoods monster. Note the large eyes in the depiction.



Figure 5 – Note the similarities between the face of this barn owl and the face of the Flatwoods Monster

Another instance of a possible bird sighting was the so-called “Flatwoods Monster” which was a reported contact with what the observers thought was an extraterrestrial lifeform. The night of the incident was the same night that a meteor was observed, and the people who reported seeing the Flatwoods Monster went to investigate the meteor crash in the local forest. When they happened upon the location where the meteorite hit, they reported seeing a human-like figure that scared them. A sketch artist’s impression of what they saw was mocked up, and a comparison of the drawing to what a barn owl on the top of a moderately-sized tree or stump has been suggested as an intriguing correlation.

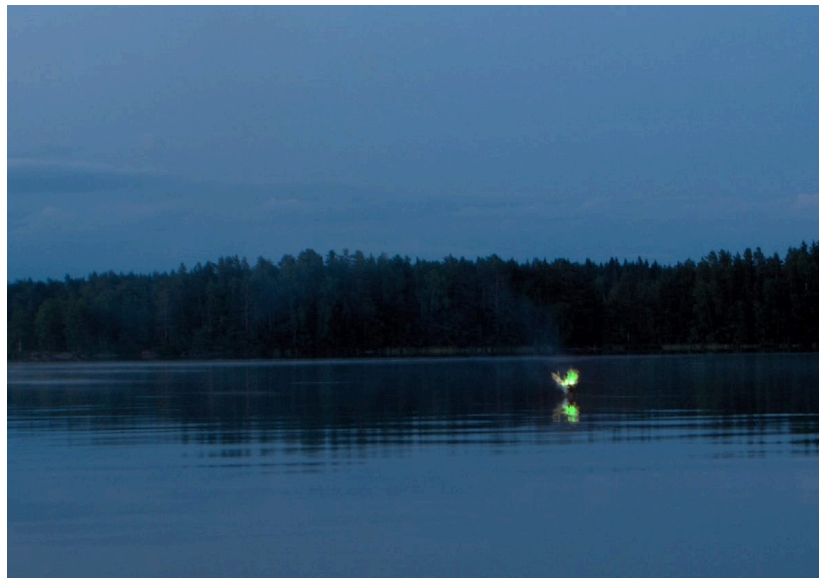


Figure 6 – The green light in this photograph is a will-o-the-wisp. Gases emitted by decaying material in wet areas can occasionally ignite to produce colorful but fleeting lights.

A few other usual natural phenomena have been identified in connection to UFO reports. Will-o-the-wisp are claims of glowing lights typically seen in marshy or swampy areas that have identified as sometimes being due to the igniting of combustible methane gas. St. Elmo’s Fire has been described since antiquity but has only recently been identified as electrical discharge that can occur in certain situations where there is substantial charging in the atmosphere as what might happen before a thunderstorm.



Figure 7 – This lenticular cloud was seen over Dublin, Ireland. The resemblance to the shape of a UFO is uncanny.

Finally, cloud formations have been misidentified, especially lenticular clouds. These clouds are formed under very particular atmospheric conditions and look superficially like the stereotypical “Flying Saucer” spacecraft often claimed in UFO reports. When such clouds appear in the evening or morning, they can reflect and refract light from the sun on the horizon taking on unusual coloring and mimicking flashing lights.

Human Technology



Figure 8 – This “space jellyfish” was produced after a night launch in California.

Many comparisons to UFOs are directly to human technology either as the airships of the late nineteenth and early twentieth centuries through to jet aircraft and rocket ships today. Even so, rocket launches are unusual enough that when they happen and an unsuspecting observer happens to be looking in the direction of a launch, they can be led to believe that they are witnessing something otherworldly. An example of this is the 2009 Norwegian Spiral Anomaly which formed a dramatic blue spiral visible from many parts of Norway one evening. It was identified as a failed rocket launch from Russia, but the unusualness of contrails and exhaust that come from the launch are such some classify it as a member of a class of “Space Jellyfish”.

Once in orbit, satellites can appear in ways that are very unusual in comparison to planes. Low Earth orbit implies a satellite that passes directly overhead will cross from horizon to horizon in 45 minutes, but most of that time will be spent near the horizon. As

it passes overhead, it moves at a considerable clip crossing the distance of five full moons in one second. Such satellites reflect the light of the Sun and so can mostly be seen near sunrise or sunset. Low-Earth orbiting satellites pass either into or out of the shadow of the Earth appearing to appear or reappear suddenly, and if the architecture of the satellite is favorable, this can lead to satellites appearing to “flare” as they point a flat reflective surface towards the unsuspecting observer.

Perhaps one of the most famous UFO-related human technology observed are balloons. While most people have experience with party balloons, there are extremely large balloons used for scientific and military purposes which look nothing like those at a party. These high-altitude balloons do not fully inflate until they reach the height of the stratosphere past 15 km or so. Once at these heights, the balloons inflate to the size of a small building and in the early evening when they are typically launched will reflect the setting sun as they travel quickly through the action of the jet stream. West Texas and New Mexico are the launching point for much of the scientific ballooning done in the US and often reports of UFOs to the local media increase during the launch season.

In the last few decades, robotic drone flyers have become another source for confusion by people seeing unexplained things in the sky. Drones used for classified missions have confounded pilots as the size of the drones and therefore their distances are difficult to ascertain. Drones may also be subject to radar detection, but in the case of military uses, stealth technology can be employed to hide or confuse the radar signal so that distances and speeds cannot be accurately ascertained. A number of the latest UFO sightings breathlessly reported in even reputable media have been attributed to drones by skeptics. Commercially available drones are also one of the most common ways to create a UFO hoax.

Delusions

A certain number of UFO sightings are due to delusions, but it is important to realize that anyone can suffer from delusions. The accusation that someone is delusional is often a tool of derision, but intelligent, capable and otherwise competent human beings are suspect to any number of effects that can cause them to believe they have perceived something that they have not perceived. A common example of this is the so-called “Misinformation Effect” where memories can be distorted and inaccurately recalled due to the introduction of misleading information or suggestions. Indeed, suggestibility is a psychological phenomenon that is offered to explain how hypnosis functions. Studies of such effects have cast doubt on the reliability of eyewitness testimony in courts, and in instances where testimony is the only evidence for an observation of a UFO, such effects may be cause for concern. Unfortunately, those who offer eyewitness testimony are often completely unaware that their memory has been distorted and instead take umbrage at the suggestion that they are mistaken, but such concerns are not intended to cast those offering their testimony as liars. Instead, it is an effect to keep in mind as a possible explanation for a testimony that was intended to be offered truthfully.

In some instances, these sorts of misremembered events can be the result of trauma. An instance of this seems to be indicated in the community of abductees who claim with varying degrees of detail and intensity, that they have not only been visited by lifeforms from beyond Earth, but that these lifeforms have captured and, in some cases, tortured them. Some accounts follow patterns that have been identified by psychologists as similar to a condition known as sleep paralysis where a certain somnambulatory state between dreaming and waking happens when a person is aware of what is happening but is unable to physically move.

Image of the depiction of the head of
a gray alien with large, black, almond-
shaped eyes, a large head, and a small
nose and mouth.

Figure 9 – The iconic “gray alien” was first described by Betty Hill, although there are similar figures from science fiction television and film shows that preceded her account.

The first claimed abduction is worth recounting due to the iconic nature of its occurrence. Barney and Betty Hill were an interracial married couple living in New Hampshire in 1961 at a time when interracial marriages were uncommon and actually illegal in many parts of the United States. During a drive home, they saw a UFO, followed it in their car, left the vehicle to investigate, and then became very frightened of what danger it could pose. They reported a strange loss of time upon their arrival back home. For some days afterwards, they were haunted by the trauma of the event and Betty had reported troubling dreams. At the time, UFO stories were still being investigated by national authorities, and the stories often ended up reported in the press uncritically. Also at the time, hypnosis was being actively used in a way we now know is problematic: to help people recall so-called “repressed memories”. Unfortunately, it has been convincingly documented that encouraging people to participate in an activity where they are subject to suggestibility as in a hypnosis session is a way to plant false memories. Nevertheless, through hypnotic sessions that began some years after the incident, Barney and Betty Hill began to describe more remembered features of their encounter including an abduction by what they described as extraterrestrials who took them to their spacecraft and examined them. Betty

Hill outlived her husband by decades and continued to produce more and more claims about her contact with the extraterrestrial lifeforms including describing the aliens as the now familiar “gray alien”, and eventually determining that they came from a planet orbiting the star Zeta Reticuli. An entire community of people believe that “Zetans” are the intelligent extraterrestrials visiting Earth largely based on Betty Hill’s prolific accounts that are stored in the archives of the University of New Hampshire library.

Delusions do not only happen to individuals. There are instances of so-called “mass delusions” where multiple independent people or groups of people become convinced of events or phenomena which simply do not exist. One example of this that predates the UFO phenomenon were reports of the Edison Star in March and April of 1897. At this time, there was concern across the United States that Thomas Edison was producing an artificial star to light up the night. Sightings of this star happened in many locations, but there was no such object to be seen. This event is very reminiscent of various “UFO waves” which occur in places from time to time. As media interest intensifies, more and more reports are made of people seeing unexplained objects flying in the sky to the point where the argument is offered, sometimes even explicitly, that if it had to be interactions with extraterrestrial lifeforms because otherwise there would not be so many sightings. After an increasing flurry of interest, the reports eventually come to an end as the attention of the media and society is moved towards other things.

Hoaxes

Since the claims began to proliferate, hoaxes have been promulgated. Faked photographs, video footage, and physical evidence have been used to claim “clear proof” and discoveries of hoaxes inevitably lead skeptics to question claims that come after this. One instance where a hoax was strongly suspected was the first photograph ever taken of a UFO from a farm near McMinnville, Oregon. The pair of pictures appear to show a flying saucer and the claim was that the pictures also indicate the UFO was moving from one part of the scene to the other. Suggestions that the photographs were faked were made almost from when they were first published, most explanations proposing that either an old motor, a hubcap, or a dinner plate was suspended by fishing line. The family never admitted that it was a hoax.



Figure 10 – Crop circles are still made by artists and jokesters today.

Perhaps the most famous hoax associated with ufology was that of crop circles. For years in the late 1980s and early 1990s, crops in south England were discovered mysteriously flattened in circular patterns that became increasingly elaborate. A craze of investigation happened that captured worldwide attention. Various phenomena were invoked as plausible explanations including claims by many who believed that UFOs were extraterrestrial spacecraft that the crop circles were related to the intelligent lifeforms. In spite of the media sensation, in 1991 two men admitted to producing most of the crop circles in England using fairly rudimentary devices. Even as they were able to demonstrate their production of crop circles, there still remained believers convinced that this was not the explanation some of whom still do not believe the phenomenon was a hoax. Crop circles continue to be produced by artists and other aficionados including one that famously paid homage to the Arecibo Message.

Conspiracy Theories

Even with the large number of prosaic explanations for claimed UFOs and the number of reports issued that concluded there was no strong evidence for UFOs indicating an interaction with life from beyond the Earth, there continue to be those who believe that some UFOs are not explainable by the four categories of identifications listed above. When faced with the question as to why government agencies and experts dismiss such arguments, there is sometimes a claim that there is a conspiracy among a large number of interested parties to conceal the truth of the matter. Stories as to how this has done often include hiding evidence at secret military installations such as the famous “Area 51” in the United States Air Force’s Nevada Test and Training Range where the promoters of the “Alien Autopsy” video claimed evidence was being hidden. There are declassified documents which indicate that the United States Armed Forces in part did not spend time refuting such claims in order to hide the true classified nature of some of these institutions (some of which were involved in the production of espionage technology, for example).

Whether these stories are promoted by people who know they are incorrect or whether they were promoted by people who truly believe in “alternative scenarios” is hard to determine. Suffice to say, these expansive narratives become increasingly complicated as the interested interlocutor continues to ask additional questions about the existence or lack of existence of evidence for such conspiracy theories. Inevitably, completely unrelated arguments end up incorporated. Arguments get made such as, “If NASA is lying about the existence of extraterrestrial lifeforms interacting with human beings, perhaps it is also lying about the shape of the Earth.” Aside from the fact that it is possible to perform some simple experiments yourself to discover that NASA is not lying about the shape of the Earth, the chain of argument becomes longer and longer and weaker and weaker as the explanations continue.

One particular conspiracy theory that has gained traction in the popular media is the proposal that not only are extraterrestrial lifeforms interacting with Earth, but that there is evidence that they have been doing so for millennia. The evidence for such “Ancient Aliens” is usually the large structures that humans built before the Industrial Era. The pyramids of Egypt and Central America, the megaliths of Europe and Asia, the now-abandoned cities of the past in the Eastern and Western Hemispheres are all cited as in certain television shows and popular books that there must have been contact with intelligent lifeforms possessing advanced technology to enable humanity to construct these impressive structures.

Archeologists and anthropologists have entire journals dedicated to explaining the way in which our predecessors constructed the large and complicated features of their societies. Not everything is completely understood, but there is ample evidence of the construction techniques, design knowledge, and abilities of the craftspeople from written records, artifacts, and even oral histories. The presumption by many of those arguing that human beings were incapable of achieving the incredible feats of the past is that because *they* do not understand how, for example, a pyramid could be built therefore no human being can. The people who built the pyramids beg to differ.

The allure of such arguments is that they excite the imaginations of many. In the modern world, science fiction and fantasy literature may hold functions similar to those that folktales and mythologies used to hold. That some people confuse these stories for factual accounts is not surprising, but as with any claim to truth, it is important to keep stock of the full range of possibilities and narrow on in explanations that are more likely to be correct while avoiding those that are more likely to be incorrect. Hopefully, the baloney detector kit can work.

An Enchanted Universe

The connection of these ideas to mechanisms that are unknown to scientific investigation inspires some people to delve into open questions which have been attempted to be answered through alternative philosophical and religious means. Examples include: Is there something special about life that requires explanations that our physical models will be unable to provide? Are there ways that we and other aspects of reality are influenced beyond the four forces of physics?

While open questions of this sort are, by definition, not answered, there are auxiliary proposals and ideas for which we can examine the evidence.

Astrology

The predictability of the motions of the heavens enabled early astronomers to become accurate timekeepers and calendar makers. This skill required observation and modeling as we currently enjoy, but there were extensions to such predictions that were abandoned by the scientific community in the eighteenth century. These included so-called “astrological” claims that the planets influenced human affairs and current events through personality-like influences. These claims continue to be made today

with the influences argued as being due to personal characteristics that the planets and constellations were given. For example, Mars was considered a masculine god of war and Venus a feminine goddess of love. The position of these visible planets with respect to the Sun, Moon, and background stars was meant to influence daily life in individuals depending most strongly on where the Sun appeared to be when they were born, and less strongly on where the Moon and the rest of the planets were. As the Sun, Moon, and planets all orbit in a plane, the only constellations that they can be seen in are the twelve constellations that form the Zodiac or the Plane of the Ecliptic which indicates the orientation of the plane of our Solar System with respect to the Earth (if there were no tilt to our axis, the Plane of the Ecliptic would align with the Celestial Equator).

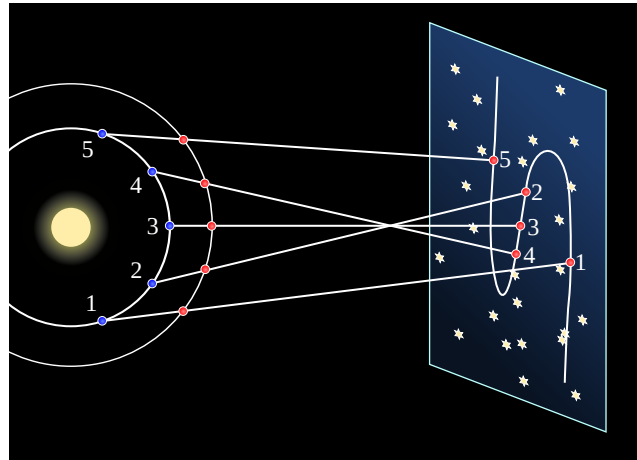


Figure 11 – Retrograde motion happens in this instance as an outer planet travels more slowly but in the same direction as an inner planet orbiting the Sun.

The idea that the position of the planets with respect to stars influences human behavior or future outcomes generally was (and, by contemporary believers in astrologers, continues to be) attributed to the same mysterious forces which kept the planets moving in their courses. However, with the advent of both the heliocentric model for the solar system and our modern theory of gravity, it became possible to explain the motion of the planets and precisely calculate the strength of the forces that planets exert. The scientific explanations for the motions of the planets often recast the stories of astrology in a different light. For example, retrograde motion of the planets still worries people, many of whom do not know what retrograde motion is. At various points along the synodic period of a planet (the time it takes for a planet to return to the same position in the sky), a planet will appear to stop moving in the normal direction it wanders across the constellations of the Zodiac. At that point it will turn apparently make an abrupt change of course and move in the opposite direction before stopping again and returning to its previous path. This retrograde motion is said by astrologers to indicate a disorienting or disordered characteristic to the planets' motion (recall that planets were once considered imbued with consciousness as deities or spirits), but in fact it is the result of the Earth moving around the Sun. The analogy can be drawn to cars passing each other on a roadway while traveling in the same direction. For most of the journey, the cars will appear to move in the same direction, but at certain points one car will pass another causing the appearance of the cars moving in different directions. Retrograde motion occurs as the Earth passes the planets that are farther away than the Sun or as the planets that are closer to the Sun swing in their orbit to the opposite side of the Sun and appear to move in a different direction for a time. If you believe the Earth is the center of the Solar System, such motion is strange and puzzling and famously needed to be explained using a complicated system of epicycles which allowed a planet to move backwards at certain points in its motion. With the heliocentric Solar System, the explanation becomes as simple as explaining relative motion.

Our advancing understanding of the motions of the celestial sphere also mean that the claims of many so-called “classical” astrologers are based on calculations that place the planets in different places in the sky than they actually are. The Earth's axis precesses every 26,000 years which means that the Celestial Poles point in slightly different directions with respect to the background stars year by year. This motion was unknown to ancient astronomers because it requires either a significant baseline of observations before the shifts are seeable without advanced technology. But it has been long enough since the invention of astrology that most people's “Star Signs” are no longer the constellation that the Sun was in when they were born.

Finally, there is a question of mechanism when it comes to astrology. Some astrologers were probably encouraged by the fact that Isaac Newton's explanation for gravity required the planets to exert gravitational pulls on human beings. However, the force exerted is minuscule compared to other forces we encounter in our day-to-day lives. The planet Jupiter exerts less of a gravitational influence on you than the gravity from a nearby human being; it is essentially too small to measure. As for the other forces known to physics (electromagnetism, the strong force, and the weak force), their influences are even smaller.

Thus, in order for astrology to work there would have to be some unknown fifth force originating from the planets, influencing humanity, and following the traditional stories that were told to explain motion of the planets that we now explain through celestial mechanics. Such a force should be measurable, but it does not appear in any of the tests that physicists make to look for such fifth forces. The conclusion that most physicists arrive at, then, is that astrological claims cannot be attributable to actual physical influences.

At this point, some believers in astrology appeal to spirituality to argue that the influences cannot be measured. There are profound philosophical implications to believing in the existence of things that cannot be measured, well beyond the scope of this text. By definition, there is no test that can be done to either substantiate or debunk the existence of immeasurable things. We have left the purview of science and have entered into a space of ideas that is more properly contained within metaphysics or religion.

Spirituality and Religion

Some religions make claims directly relevant to the questions this book is seeking to explore. For example, there are those who have a religious belief that there cannot possibly be intelligent life in the Universe other than humanity. Others not only believe in such life, but make claims to know what the life is and how it operates. Some modern religions even incorporate UFOs as part of their belief systems.

Other religious believers question fundamental results of scientific investigation. However, most mainstream religions do not see any necessary conflict between scientific investigations, the facts discovered by science, and the religious beliefs. There may be outstanding queries and disagreements as to what is possible or impossible in such discussions of religion and science, but generally there is an ongoing demarcation that is maintained between religion and science that is only breached in certain philosophical fora. Students are often intrigued by this (lack of) interaction, but because the approaches to explaining the world are very different between a religious perspective and a scientific perspective, it is typically difficult to isolate even what the topics up for discussion should be. Open questions in science are sometimes left to religions to answer. Whether they provide reasonable answers or not is largely a question of faith.

Key Concepts and Summary

Without a second example of life elsewhere, the field of astrobiology is speculative. This offers an excellent opportunity to develop critical thinking, skepticism, and logic as thinking habits. It is human nature to want to succeed. It is scientific nature to be skeptical and demand empirical evidence before jumping to conclusions and the scientists working on SETI and the search for life employ these best practices in their work. They will be the first to say there is no evidence that UFOs are alien spacecraft or to identify hoaxes met to deceive. These practices may be the most important thing you can take away from your study of astrobiology.

Review Questions

Summary Questions

Activities

INDEX

No index terms found.

This page titled [Index](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled [Glossary](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .

DETAILED LICENSING

Overview

Title: Astrobiology (Fischer, Sheffield, Tan, and Zhao)

Webpages: 43

Applicable Restrictions: Noncommercial

All licenses found:

- [CC BY-NC-SA 4.0](#): 97.7% (42 pages)
- [Undeclared](#): 2.3% (1 page)

By Page

- Astrobiology (Fischer, Sheffield, Tan, and Zhao) - [CC BY-NC-SA 4.0](#)
 - Front Matter - [CC BY-NC-SA 4.0](#)
 - TitlePage - [CC BY-NC-SA 4.0](#)
 - InfoPage - [CC BY-NC-SA 4.0](#)
 - Table of Contents - [Undeclared](#)
 - Licensing - [CC BY-NC-SA 4.0](#)
 - 1: Understanding the Cosmos - [CC BY-NC-SA 4.0](#)
 - 1.1: Sizing Up the Cosmos - [CC BY-NC-SA 4.0](#)
 - 1.2: Energy and Light - [CC BY-NC-SA 4.0](#)
 - 1.3: Spectroscopy and Stars - [CC BY-NC-SA 4.0](#)
 - 1.4: Laws of Motion and Gravity - [CC BY-NC-SA 4.0](#)
 - 2: The Chemistry of Life - [CC BY-NC-SA 4.0](#)
 - 2.1: Introduction to Atoms and Elements - [CC BY-NC-SA 4.0](#)
 - 2.2: Basic Chemistry of Life - [CC BY-NC-SA 4.0](#)
 - 2.3: Key Biochemicals- Amino Acids, Nucleic Acids, and the Genetic Code - [CC BY-NC-SA 4.0](#)
 - 2.4: Alternative Biochemistries- Speculations about Strange Possibilities - [CC BY-NC-SA 4.0](#)
 - 3: Origins of Life - [CC BY-NC-SA 4.0](#)
 - 3.1: Definition of Life - [CC BY-NC-SA 4.0](#)
 - 3.2: From Biochemistry to Biology - [CC BY-NC-SA 4.0](#)
 - 3.3: Emergence of Life on Earth - [CC BY-NC-SA 4.0](#)
 - 3.4: Development of Life on Earth - [CC BY-NC-SA 4.0](#)
 - 3.5: Extremophiles - [CC BY-NC-SA 4.0](#)
 - 4: The Search for Life in Our Solar System - [CC BY-NC-SA 4.0](#)
 - 4.1: Solar System Formation - [CC BY-NC-SA 4.0](#)
 - 4.2: Earth Characteristics - [CC BY-NC-SA 4.0](#)
 - 4.3: Searching for Life In Our Solar System - [CC BY-NC-SA 4.0](#)
 - 4.4: The Search for Life on Mars - [CC BY-NC-SA 4.0](#)
 - 4.5: The Jovian Moons - [CC BY-NC-SA 4.0](#)
 - 5: The Search for Life Beyond Our Solar System - [CC BY-NC-SA 4.0](#)
 - 5.1: The Habitable Zone - [CC BY-NC-SA 4.0](#)
 - 5.2: Exoplanets- Early History and Direct Imaging and Doppler Detection Methods - [CC BY-NC-SA 4.0](#)
 - 5.3: Transit, Astrometry and Gravitational Microlensing Techniques - [CC BY-NC-SA 4.0](#)
 - 5.4: Exoplanets- Statistics and Modern Discoveries - [CC BY-NC-SA 4.0](#)
 - 6: Searches for Technological and Intelligent Life in the Universe - [CC BY-NC-SA 4.0](#)
 - 6.1: The Challenges of Space Flight - [CC BY-NC-SA 4.0](#)
 - 6.2: The Search for Extraterrestrial Intelligence (SETI) - [CC BY-NC-SA 4.0](#)
 - 6.3: Biosignatures and Technosignatures - [CC BY-NC-SA 4.0](#)
 - 6.4: The Drake Equation and the Fermi Paradox - [CC BY-NC-SA 4.0](#)
 - 6.5: Speculations and Skepticism - [CC BY-NC-SA 4.0](#)
 - Back Matter - [CC BY-NC-SA 4.0](#)
 - Index - [CC BY-NC-SA 4.0](#)
 - Glossary - [CC BY-NC-SA 4.0](#)
 - Detailed Licensing - [CC BY-NC-SA 4.0](#)

This page titled [Astrobiology \(Fischer, Sheffield, Tan, and Zhao\)](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by Debra Fischer, Allyson Sheffield, Joshua Tan, and Lily Ling Zhao via [source content](#) that was edited to the style and standards of the LibreTexts platform.