Life in the Universe (PHY-30025)

Spring 2020

Jacco van Loon

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.1.1 Planet discoveries . . . 1
1.1.2 Space age . . . 2
1.2 Setting the stage . . . 2
1.2.1 What is life? . . . 2
1.2.2 Extra-terrestrial intelligence . . . 3
1.2.3 Drake's equation . . . 4

2 Life as we know it – basics **5**
2.1 Lifeforms . . . 5
2.1.1 Simple organisms . . . 5
2.1.2 Complex organisms . . . 5
2.2 Essentials . . . 6
2.2.1 Carbon . . . 6
2.2.2 Liquid water . . . 7
2.3 Chemical components . . . 7
2.3.1 Lipids . . . 7
2.3.2 Carbo-hydrates . . . 8
2.3.3 Amino acids and proteins . . . 8
2.3.4 Coding genetic information: DNA . . . 8
2.4 Energy . . . 10

3 Life as we know it – functioning **13**
3.1 Physiology . . . 13
3.1.1 The cardio–vascular system and lungs . . . 13
3.1.2 Sensors . . . 14
3.1.3 The nervous system and brain . . . 16
3.1.4 Motoric attributes . . . 17
3.2 Growth and decline . . . 18
3.3 Extremophiles . . . 19

4 Mentality **21**
4.1 The brain . . . 21
4.2 Mental states . . . 22
4.2.1 Consciousness . . . 23
4.2.2 Sleep . . . 24

4.2.3 The rôle of neurons in memory and consciousness . . . 25
4.3 Automata . . . 25

5 The emergence of life on Earth 27
5.1 How did it start? . . . 27
5.1.1 Tree of Life . . . 27
5.1.2 The Miller–Urey experiment . . . 28
5.1.3 One of life's mysteries: chirality . . . 28
5.2 When and where did it start? . . . 28
5.2.1 The changing conditions on Earth . . . 28
5.2.2 The significance of the other planetary bodies . . . 30
5.3 The origin of life on Earth in a nutshell? . . . 31

6 Habitable Zones and Earth's atmosphere 33
6.1 Conditions for a planet to be habitable . . . 33
6.2 The temperature of a planet . . . 34
6.2.1 Radiative equilibrium . . . 34
6.2.2 Albedo . . . 35
6.2.3 Temperature variations . . . 35
6.2.4 The Greenhouse effect . . . 36
6.3 The Habitable Zone . . . 37
6.4 The planet's atmosphere . . . 38
6.4.1 Hydrostatic equilibrium . . . 38
6.4.2 Evaporation . . . 40

7 Life elsewhere in the Solar System 41
7.1 Where it's definitely not . . . 41
7.1.1 Mercury . . . 41
7.1.2 The Moon . . . 41
7.2 Venus – our close twin . . . 41
7.3 Mars – our little brother . . . 42
7.3.1 Canals? . . . 42
7.3.2 Water? . . . 43
7.3.3 Maintaining the CO_2 atmosphere . . . 44
7.3.4 Evidence for life . . . 44
7.4 The Giants: Jupiter and Saturn . . . 45
7.4.1 Europa . . . 45
7.4.2 Ganymede & Callisto . . . 46
7.4.3 Titan – the moon that ought to be a planet . . . 46
7.4.4 Enceladus . . . 48
7.5 Minor bodies . . . 48
7.5.1 Dwarf planets, asteroids and meteorites . . . 48
7.5.2 Meteorites from Mars . . . 49
7.5.3 Asteroids from Outer Space . . . 49
7.5.4 Comets . . . 49
7.5.5 Outer solar system dwarf planets . . . 50

8 Searching for exoplanets **51**
8.1 Direct detection . . . 51
8.1.1 Reflected starlight . . . 51
8.1.2 Thermal emission . . . 52
8.1.3 Angular resolution . . . 53
8.2 Detection through occultation . . . 53
8.2.1 Planetary transits . . . 53
8.2.2 Planetary eclipses . . . 54
8.2.3 Transit Timing Variations . . . 54
8.3 Detection through the star's reflex motion . . . 54
8.3.1 Astrometry . . . 55
8.3.2 Doppler shifts . . . 55
8.4 Gravitational microlensing . . . 56
8.4.1 Einstein radius . . . 57
8.4.2 Amplification . . . 58
8.5 Pulsar timing . . . 58

9 Properties of discovered exoplanets **59**
9.1 Planet masses . . . 59
9.2 Orbital radii . . . 60
9.3 Planetary systems . . . 60
9.4 Planet host stars . . . 61
9.5 How complete is our picture? . . . 62

10 Exoplanet atmospheres and exoplanets in Habitable Zones **63**
10.1 Eroding atmospheres . . . 63
10.1.1 Planet densities . . . 63
10.1.2 Hot Jupiters . . . 64
10.1.3 Radiative heating and evaporation . . . 64
10.1.4 Radiation pressure . . . 65
10.1.5 Stellar wind pressure . . . 66
10.2 Have we found planets in their Habitable Zone? . . . 66
10.2.1 Moons around massive gas planets in the Habitable Zone . . . 67
10.2.2 Planets around giant stars . . . 67
10.2.3 Planets around pulsars . . . 67
10.3 Orbits and gravitational perturbations . . . 68
10.3.1 The importance of eccentricity . . . 68
10.3.2 The importance of multiplicity . . . 68

11 The formation of stars **69**
11.1 The interstellar medium . . . 69
11.2 Gravitational instability . . . 69
11.3 Hydrostatic equilibrium and the need to cool . . . 72
11.3.1 Interstellar chemistry and the building blocks for life . . . 74
11.4 The next hurdle: magnetic fields . . . 74

12 Proto-planetary discs **77**
12.1 The angular momentum catastrophy . . . 77
12.2 Star–disc interaction . . . 77
12.2.1 Accretion and feedback . . . 78
12.3 Proto-planetary discs . . . 79
12.4 Observed disc properties . . . 80
12.4.1 The showcase disc of β Pictoris . . . 80
12.4.2 An atlas of proto-planetary discs . . . 80

13 The formation of planetary systems **83**
13.1 The formation of planets . . . 83
13.1.1 Disc instability . . . 83
13.1.2 Planet formation "bottom up" . . . 83
13.1.3 Planet migration . . . 84
13.2 Gravitational perturbations . . . 85
13.2.1 The Hill radius . . . 85
13.2.2 The Roche limit . . . 85
13.2.3 Lidov–Kozai cycles . . . 86
13.2.4 The Rossiter–McLaughlin effect . . . 86

14 Finding signatures of life **89**
14.1 Planetary conditions for life . . . 89
14.2 Atmospheric signatures of a biosphere . . . 90
14.2.1 Detecting and probing an atmosphere . . . 90
14.2.2 Oxygen . . . 92
14.2.3 Out-of-equilibrium chemistry as a biomarker . . . 92
14.2.4 Methane and ammonia . . . 93
14.3 Detecting vegetation . . . 93
14.4 Direct imaging of an exoplanet's surface . . . 94

15 The Sun–Earth connection **95**
15.1 The Earth's orbit and rotation . . . 95
15.2 The Sun's activity . . . 96
15.3 Space weather . . . 97
15.3.1 The solar wind . . . 97
15.3.2 Solar flares and Coronal Mass Ejections . . . 98
15.3.3 Geo-magnetic phenomena . . . 98
15.3.4 Cosmic rays . . . 99
15.4 The Sun's evolution affects the Habitable Zone . . . 100

16 Natural and human hazards to life **101**
16.1 Are humans causing the next mass extinction? . . . 102
16.1.1 The ozone hole . . . 102
16.1.2 Carbon–dioxide and the Greenhouse effect . . . 102
16.2 Circulation patterns and weather . . . 103
16.3 Ecosystems . . . 105
16.3.1 Feedback . . . 105
16.3.2 Carrying capacity . . . 106

17 Threats from outer space **107**
17.1 Impacts . . . 107
17.2 Supernovæ . . . 109
17.2.1 The local interstellar environment . . . 110
17.2.2 A Galactic Habitable Zone? . . . 111

18 Travel into space **113**
18.1 Basics of rocket propulsion . . . 113
18.1.1 Improving efficiency of rocket propulsion . . . 114
18.2 Gravity assist manœuvres . . . 115
18.3 Entering orbit and descent . . . 116

19 Interstellar travel and settlement **119**
19.1 Limits set by relativity theory . . . 120
19.2 Alternative propulsion methods . . . 120
19.2.1 Beyond initial speed . . . 121
19.3 Challenges for interstellar travel . . . 122
19.4 Challenges for extra-terrestrial settlement . . . 123
19.5 Ethical considerations . . . 124

20 Extra-Terrestrial Intelligence **125**
20.1 SETI . . . 125
20.1.1 What would one be looking for? . . . 126
20.2 CETI . . . 127
20.2.1 What would a message be like? . . . 127
20.3 Dyson spheres . . . 128
20.4 Colonization by intelligent civilizations . . . 128
20.5 Closing remarks . . . 129

Chapter 1

Introduction

1.1 Motivation

Are we alone?

This is our prime motivation when we investigate whether life can exist elsewhere in the Universe, and when we try to find it. As part of this quest, we shall explore many facets of life with important bearings on our health, and indeed survival as a species. Apart from our growing understanding of what is life, and how it has evolved on Earth, there are two important recent developments that facilitate the study of life in the cosmos.

1.1.1 Planet discoveries

Since the first planets outside our own Solar System were discovered in the 1990s, over 4000 such 'exoplanets' are now known. This has revolutionized our knowledge of the potential for life to exist in the Universe besides that found on Earth. We are still at the dawn of exploration of these exoplanets, and have very exciting times ahead of us.

As recently as 2008, the first direct images were announced of what most people agree are planet-sized bodies in orbit around another star: one very faint planet was discovered around one of the brightest stars in the sky, Fomalhaut, and three (*sic*!) planets were found around another bright star, HR 8799. Two arguments led people to believe these are planets, and not binary stellar companions or stars in the background:

- ⋆ The objects are very cold ($\ll 2000$ K) and faint; a stellar companion would produce a lot of energy through nuclear fusion in its core, and as a result be much brighter and have a hotter surface: a stable star will need to be able to radiate away the energy at the same rate at which it is produced inside, otherwise it will inflate the star. If this rate is very high the star will need to have a very large and/or hot radiating surface in order to acquire the required luminosity. Planets are not massive enough to reach the extremely high pressures in the core that are required for atoms to overcome the Coulomb barrier and fuse.

- ⋆ By observing the system again a year or so later, the star and candidate planets were found to move together through space, making it very likely that they are really near to each other and not simply happen to be seen in the same direction on

the sky. Better still, the planet candidates did show a tiny difference in movement compared to the star. This could be explained by motion in orbit around the star. Using Newton's law of gravitation, assuming fairly circular orbits, and making an educated guess for the mass of the star (from its luminosity and temperature; the distances to these nearby stars can be measured using the reflex motion in the sky as a result of the Earth's motion around the Sun – called the parallax), the masses of the planets could be estimated. The masses of these planets are a few times as much as that of Jupiter, which is the most massive planet of the Solar System.

In 2012, a team of astronomers announced the discovery of an Earth-mass planet in the next-nearest stellar system, around α Centauri B. This was followed in 2016 by the discovery of a potentially habitable Earth-mass planet around the cooler member of this triple star system, Proxima Centauri.

1.1.2 Space age

Since half a century, humans have ventured into space. They have walked on the Moon, lived in space stations for many months, and shot robotic probes to all the planets in the Solar System and landed on some of them (as well as on the moon Titan and comet Churyumov–Gerasimenko, in 2014). Four of these spacecraft have now left the confines of the planetary system and they are on their ways into the depth of interstellar space.

It has thus become possible to search for life elsewhere in the Solar System. We can investigate the effects on humans and other organisms when they are weightless, and no longer protected by the Earth's atmosphere and its magnetic field. We can even start thinking of the possibilities for human settlement, either in space colonies or on the surfaces of other planets or moons.

Establishing contact with other "intelligent" life in the Universe has now become a real possibility, because of our ability to travel through space, but also because of our telecommunication technology (e.g., radio broadcasts, and radio astronomy).

1.2 Setting the stage

1.2.1 What is life?

It is not trivial to come up with a definition for life that is both useful and encompasses all life as we know it – as well as life we do not (yet) know. A widely adopted definition for life is that:

- ⋆ Life is what has the capacity to reproduce.
- ⋆ Life is what has the capacity for Darwinian evolution.

By reproduction we mean the generation of like offspring. By Darwinian evolution we mean that the reproduction should not make identical copies, but allow for a degree of diversity in offspring. The success with which the offspring thrives in the environment is measured by its ability to survive and reproduce. Over time, the more successful versions

establish an optimized functionality of the species. The lack of control of evolution is perhaps odd, and may suggest that a higher lifeform may be able to influence it – indeed, humans can genetically manipulate this process, and artificial intelligence may one day manufacture superior offspring.

There are problems with the above criteria. For instance, mules cannot reproduce. So while these criteria may be sufficient (if satisfied, it is life), they may not be necessary (examples of life exist that do not satisfy these criteria). Other definitions are based around growth and the generation of energy in an organised system of chemical reactions (e.g., metabolism, respiration); a thermodynamical definition is based around systems that convert low-entropy mass (nutrients) into high-entropy energy (waste, exhaust). A spreading fire also conforms to these definitions. Are stars alive? What it does suggest, is that life may just be one form among the many dynamical phenomena we encounter in the Universe.

A common aspect of what we would all agree on are lifeforms, is that they work based on bringing systems out of equilibrium, and benefitting from their subsequent drive to return to equilibrium. This is potentially helpful in inferring the presence of a biosphere on the basis of the detection of non-equilibrium conditions. The emergence of life itself may have been the product of the same tendency of systems to aspire to achieve equilibrium. Once they became self-organised, these systems would be called organisms. This would imply that the initial conditions in which life formed were highly out of equilibrium. The identification of these non-equilibrium environments, and their causes, could be an important step in explaining the origin of life.

"You know it when you see it" is a rather unsatisfactory definition, but all organisms on Earth that we would call "alive" carry a piece of information that is coded in the form of a complex molecule, a 'genome'. It appears this is an essential requirement for organisms to function, grow and reproduce as well as to adapt to a changing environment. It would include viri, although these are reliant upon other organisms for their reproduction.

Lifeforms often do not work in isolation. Contributing to a greater good could make a population of lifeforms more sophisticated than the individual lifeforms themselves – think of ant colonies, for example. Symbiosis benefits lifeforms of different species altogether. Should the system of lifeforms be considered a lifeform in itself? Is the entire biosphere of Earth a lifeform?

1.2.2 Extra-terrestrial intelligence

What we are really keen on finding, is extra-terrestrial *intelligent* life. But here, too, it is unclear what we mean by this, and how we could define it. Do we mean species with whom we might communicate (language barriers notwithstanding)? Who are technologically advanced? Or who are conscious of their own existence and thoughts? Do these go hand in hand – is language a necessity for thoughts and learning and the development of technology? What about machine learning, the ability of artificial systems to train and develop? If a machine passes the 'Turing Test', outwitting a human, can it be said to have mentality? Would artificial intelligence ever be truly intelligent, if it possessed no sense of awareness? Could we infer consciousness (awareness of one's mental self) from observed behaviour?

Where do we draw the line? Are dolphins intelligent? Were the Neanderthal intelligent? We would never have detected their presence from outer space. On the other end of the scale, hyper-intelligent species may be so different from ourselves that we have no chance of communicating with them, and who says they even need technology... or a body?

Most philosophers and neuro-scientists no longer support Descartes' substance dualism that separates the mind from the body. But does the fact that the mental supervenes upon the physical imply a lack of Free Will, as it would be governed purely by physical processes and thus obey the laws of Physics? Life is inseparably linked to the passing of time – it is a dynamical phenomenon, and not defined instantaneously. Without an understanding of the meta-physical nature of time (is it a dimension or a parameter?), how can we ever understand the mind, and explain our motivation to search for extra-terrestrial life?

In subsequent chapters we will be investigating what organisms require in order to live and reproduce. This will be heavily biased by life as we know it. But it enables us to start searching the Universe for places where similar requirements are met. And all the while, we will be learning more about ourselves and our world.

1.2.3 Drake's equation

How common is life in the Universe?

The American radio astronomer Frank Drake was the first to broadcast a message into space (in the early 1960s) with the deliberate intention that it be received by an extra-terrestrial intelligent civilization. Perhaps for the better, it was sent in the direction of a dense cluster of stars, taking many thousands of years to arrive at what is now regarded as a rather inhospitable place. To estimate the chances that we ourselves would be able to receive messages sent by such civilizations, he devised a simple mathematical expression for the number of broadcasting civilizations that exist at any given time:

$$\begin{aligned} N_{\text{planets with (intelligent) life}} &= R_{\text{at which suitable stars form}} \\ &\times \langle n_{\text{planets per star}} \rangle \\ &\times P_{\text{planet to be habitable}} \\ &\times P_{\text{life to emerge on such planet}} \\ &\times T_{\text{(intelligent) life to exist}} \end{aligned} \tag{1.1}$$

Here, R is a rate ($\mathrm{d}N/\mathrm{d}t$), $\langle n \rangle$ is an average number, P is a probability ($0 < P < 1$), and T is a duration. This presumes that life requires a suitable planet orbiting a suitable star.

We have made great progress in developing an idea of the magnitude of some of the terms in Drake's equation. For instance, in our Milky Way, we observe $R \sim 1$ yr^{-1}. Exoplanet searches suggest $\langle n_{\text{planets per star}} \rangle \sim 1$ (perhaps several), and even that $P_{\text{planet to be habitable}}$ may be as much as ~ 0.1. If life is a natural outcome of the chemical and physical processes on such planets then $P_{\text{life to emerge on such planet}}$ may also be close to unity. Given the short time humans have been around, especially in the technological age, $T \sim 1000$ yr is a fair educated guess. This would yield $N_{\text{planets with intelligent life}} \sim 100$. If space is more inhospitable than that, and/or the formation of life is dependent on a critical condition or event that is rare, then $N_{\text{planets with intelligent life}} \ll 100$ and easily less than unity and we are the only intelligent civilization in the Milky Way and among the galaxies around us.

Chapter 2

Life as we know it – basics

2.1 Lifeforms

2.1.1 Simple organisms

For life to be able to reproduce (in a way that allows Darwinian evolution), it requires:

- ⋆ A cell, or a body of cells, that make up a living entity (the 'organism').
- ⋆ Genetic information, to govern the organism itself and to allow it to replicate itself.
- ⋆ Nutrients, as building blocks for growth and functioning.
- ⋆ Energy, in order to function, grow and reproduce.

The cell, in its simplest form, is an enclosure consisting of a cell membrane (and sometimes a cell wall, for added rigidity) that regulates the exchange of molecules between the cell and the environment. The cell is filled with a liquid plasma called cytosol. Within the cytosol, genetic information is stored (often in a separate cell nucleus), and chemical factories drive the metabolism that generates and stores energy and manufactures complex molecules such as proteins (in ribosomes and, in complex organisms, in mitochondria and the endoplasmic reticulum or in chloroplasts).

2.1.2 Complex organisms

Complex life differs from simple life in that:

- ⋆ Complex organisms are multi-cellular.
- ⋆ Cells in complex organisms have specialized functions.

The human body contains roughly 10^{13} cells. That is a thousand times as many cells in one human body as there are human individuals on Earth. It is a similar number to the number of stars in the entire Milky Way galaxy. Just as different individual human beings perform different rôles in a society, the cells in a human body have specialized to perform certain functions. Because they all carry the same genetic information (see below), this differentiation must have been instantiated by external factors, as part of the

higher-order self-organisation of the body. And just as humans have domesticated animals and cultivate crops for their benefit, the human body harbours a plethora of microscopic organisms that benefit the human body, and that benefit from the human body[1]. Should one consider that internal ecosystem to be part of the body, or external to its surface?

In fact, the cells in the ancestors of complex organisms probably captured and enslaved bacteria, which later became the mitochondria and chloroplasts (which are essentially cyanobacteria), and the cell nucleus may originally have been a virus (cf. Chapter 5).

2.2 Essentials

All lifeforms on Earth depend on hydrogen (H, which is the most abundant element in the Universe), oxygen (O), carbon (C), nitrogen (N), phosphorus (P) and sulphur (S). The human body is composed, by mass, mostly of oxygen (mostly in its hydrated form of water), carbon, and hydrogen[2]. Complex organisms often rely on other elements as well, for instance the human body requires iron (Fe) to make the hæmoglobin in the blood that binds and transports oxygen, and zinc (Zn) for enzymes that act as catalysts (facilitators) of metabolic reactions; likewise, molybdenum (Mo) is an essential component of enzymes that break nitrogen bonds. Sodium (Na) from salt (when it is bound with chlorine, Cl) is used to maintain the right osmotic balance within and in between cells.

2.2.1 Carbon

All life as we know it, is based on carbon as a crucial element. Why carbon?

- ⋆ Carbon is chemically versatile: it has 4 valence electrons (electrons in the outer shell furthest from the nucleus, and hence most weakly bonded to the atom) and thus four options for other elements to "dock" and form molecules.
- ⋆ Carbon can bond with many other, important types of elements.
- ⋆ Carbon also easily bonds with itself.
- ⋆ This enables carbon to form the basis for building large, complex molecules.
- ⋆ Some of these molecules are soluble in liquid water.
- ⋆ Carbon is electrically conductive, facilitating transport of information and agents.
- ⋆ Carbon is relatively light-weight, and so will be carbon-based organisms.
- ⋆ Carbon is one of the most common elements in the Universe.

In contrast, while silicon (Si) is overall more abundant in the Universe it binds more easily with oxygen than with itself, and under typical conditions silicon–monoxide (SiO) is an insolubable solid whereas carbon–monoxide (CO) is a gas. Silicon may therefore be more useful for artificial intelligence than for biological organisms.

[1] A typical human carries as many bacteria in the gut, as cells in the body.
[2] Hydrogen is called this way because it makes (oxygen turn into) water.

2.2.2 Liquid water

All life as we know it requires a liquid, to transport molecules to where they are needed or where they can be disposed of, and to bring molecules into contact so that chemical reactions may happen. Invariably this liquid is water (H_2O). Why water?

- ⋆ Water is liquid at the right temperatures: neither too cold (chemical reaction rates increase with temperature, and low temperatures would make for very sluggish life at best), nor too hot (too much kinetic energy on an atomic level leads to the irreversible destruction of complex molecules and hence ends their functioning).
- ⋆ Water easily forms weak bonds between molecules: complex molecules are thus easily made as well as easily decomposed.
- ⋆ Water is a good solvent for many important complex molecules: as the hydrogen and oxygen atoms are misaligned, the water molecule has a net electric dipole (it is said to be 'polar'). Polar molecules easily dissolve in polar liquids, and apolar molecules easily dissolve in apolar liquids. But polar and apolar do not mix well. Few complex molecules are completely apolar.
- ⋆ Water is relatively light-weight (it consists of oxygen and two hydrogen atoms; hydrogen is the lightest element in the Universe and oxygen is relatively light-weight too). As liquid water is a major constitutent of most organisms it keeps the weight of the entire organism reasonable.
- ⋆ Oxygen is one of the most common elements in the Universe (and hydrogen *the* most common).

Water plays many more rôles, e.g., in the fusion and fission of molecules. Also, water has the unusual property that its liquid form is denser than its solid form (ice). Thus, bodies of water start freezing on the surface (because ice is more buoyant than liquid water), not throughout. This is important when considering life in the oceans of frozen worlds.

2.3 Chemical components

Chemistry is all about electro–magnetic force. This is exploited by lifeforms foremostly in the generation of energy. But to fabricate useful compounds requires a certain degree of ingenuity. Besides, the rate of a chemical reaction depends on the concentration of the reactants and on the chemical and physical environment in which it happens (competing reactions, temperature), and can sometimes be sped up using catalysts and/or increased concentration or surface area.

2.3.1 Lipids

Lipids – oils (liquid) and fats (solid) – are carbon chains, laced with hydrogen atoms.

Lipids are weakly bonded, and it is thus easy to break them up. Because the binding energy is released by doing so, lipids are ideal to store energy which can be readily accessed.

One end of the lipid is apolar, and will not dissolve in water (it is said to be 'hydrophobic'), but the other end generally is less symmetrical, for instance due to the presence of a hydroxyl (OH) instead of a hydrogen atom. That polar end can dissolve in water (and is said to be 'hydrophilic'). This unique property of lipids makes for an extremely simple but important behaviour: a bunch of lipids, all sticking their hydrophilic heads into the water and their hydrophobic tails into the air, naturally form a membrane separating the water from the air. These membranes can turn into themselves, and thereby form little balls called 'micells', within which other molecules can be contained. These micells can thus serve as storages, or as transport vehicles. Back-to-back membranes can separate bodies of watery liquid from surrounding watery liquid. If it forms a structure closed in itself, it is called a 'vesicle', which can serve as the basis for a cell.

2.3.2 Carbo-hydrates

Carbo-hydrates are carbon chains that contain many hydroxyl groups (besides hydrogen atoms; they are really hydrated carbon: C + H_2O). This makes them easily soluble in water and thus easy to transport through a cell or through an organism.

Carbo-hydrate molecules are easily stuck together to form larger carbo-hydrate molecules (called 'polysaccharides') by dehydration: a hydroxyl group on one carbo-hydrate binds with a hydrogen atom on another carbo-hydrate to form a water molecule; this water molecule detaches and the two carbon atoms that previously hosted the hydroxyl group and hydrogen atom now cling to each other. This process is called 'polimerization'.

The opposite may also happen: when a polysaccharide encounters water, the hydroxyl group and hydrogen atom in the water molecule may replace the carbon–carbon bond and split the molecule up. This process is called 'hydrolysis'.

The ease with which these molecules can be built and broken makes them useful as energy storages, as well as building material for structural support (e.g., cellulose in plants).

2.3.3 Amino acids and proteins

Amino acids are often associated with living matter, but they are in fact very simple molecules that are also found outside of organisms. Carbon-based, they contain the common element nitrogen, as well as a carboxyl group (COOH, which makes them acid).

Proteins are large, complex molecules that form by polimerization of amino acids by dehydration. Proteins provide structural support, but they are most crucial in their rôle as catalyst (proteins fulfilling this rôle are called 'enzymes') to facilitate and speed up important chemical reactions.

2.3.4 Coding genetic information: DNA

The Cambridge (UK) scientists James Watson and Francis Crick in 1953 proposed the double-helical structure of the molecules that carry the genetic information ("memory") so vital for a cell to function and replicate. This structure was proven to be correct in the same year by Cambridge colleagues Rosalind Franklin and Maurice Wilkins. They subjected the molecules to X-ray radiation. The X-ray wavelengths being similar to the

atomic scales within the molecule, these rays are sensitively diffracted according to the atomic arrangements within the molecule. The diffraction pattern as recorded on photographic film is a direct reflection of the structure of the molecule. Tragically, Rosalind who had done the crucial experiment, could not share the Nobel prize that was awarded to the other three, because she succombed to cancer at a young age.

The double-helix of this molecule provides the overall structure, which is made up of a particular sugar (a ring-shaped carbo-hydrate), namely deoxyribose, and phosphate groups (PO_4; phosphorus is also a fairly common element in the Universe). In between, and joining, the two helical backbones are pairs of nitrogenous 'bases' (nucleic acids). Hence the name of this gigantic molecule: Deoxyribose Nucleic Acid, or DNA for short. Proteins package the otherwise unwieldy molecule in the form of chromosomes.

There are five bases, four of which take part in DNA (the other one takes part in a similar molecule called RNA, in which deoxyribose is replaced by the more oxygenous sugar ribose). The sequence of combinations of pairs of bases along the DNA molecule represents the record of genetic information (the genome): the bases are like the alphabet, with the pairs the words and the sequence the story. Species are individuated on the basis of their genome, but one species does not have a unique genome: humans have genomes that differ from almost all other people, and a variety of prominent features may result from only a small amount of genetic diversity – in fact, genetic diversity among black Africans is larger than the genetic difference between the average black and white person. The genome also does not dictate the appearance or functioning of an individual, with genes being expressed depending on environmental cues. Likewise, while human physiological sex is largely determined by the specific combination of X- and Y-type chromosomes, gender identification transcends this natural definition.

The base pairs can be unbound by hydrolysis, causing the two helices of the DNA to detach like the opening of a zipper. If the DNA is immersed in a liquid containing free bases, these free bases could attach themselves to the bases stuck to each of the helical strands. Because a particular base only bonds with one particular other base, the resulting sequence of combinations of pairs will be identical to when the DNA was still in one piece. With all bases in place and the new ones being glued together with the sugar/phosphate substrate, a DNA molecule identical to the original one is formed. Likewise, the other half of the original DNA molecule will form a second, identical DNA molecule. In practice the process of replication proceeds via the use of RNA as an intermediary. Somehow the cell reacts by splitting in two, each carrying their portion of DNA: growth.

Copying genomes can – and does – go awry, and hits from energetic particles can alter existing genomes. Usually this will cause the cell to malfunction, or for offspring to be less viable. If it leads to rapid growth then it may develop into a tumor. But now and then it results in a superior version, and evolution proceeds. Occasionally, "horizontal mutations" occur, whereby plasmoid genes cross between species. This can cause antibiotic resistance of bacteria, one of the main health risks humankind currently faces.

DNA sequencing ("reading" the genome) has become a massive, commercial industry and a plethora of techniques have been devised since the early, time consuming and expensive methods. The development of novel DNA sequencing methods goes hand-in-hand with developments in nanotechnology, from material science to microscopy. First

generation techniques such as the well-known Sanger method (1970s) were based on a first step of chemical and radioactive labelling of DNA fragments, a second step of size separation through electrophoresis (using an electric field to separate charged fragments moving through a gel) and a final step of visualization with X-rays or through fluorescence or bioluminescence (with enzymes called luciferase in a process called pyrosequencing). Fragmenting allows parallellization, exploited by second generation techniques (in the 2000s): polymerase enzymes are used to clone DNA fragments and thereby amplify the signal; the information is reassembled on the basis of overlapping portions. This requires computer software for the analysis and interpretation. New generation techniques include nanopore sequencing: the DNA molecule travels through a tiny hole, affecting the ion current through the pore for varying duration.

2.4 Energy

Nothing comes for free. Energy is needed for most chemical reactions to occur, even the ones that are used to *generate* energy. Energy is needed for building large molecules, for transportation of these molecules to where they are needed, for growth and reproduction. The entire organism may need to move. It is therefore vital that the organism has a means of producing energy. The organism will need to take in nutrients (fuel) and in some cases use external energy sources (e.g., light).

Energy can be stored and transported in molecular packages of adenosine triphosphate (ATP) – which is adenosine diphosphate (ADP) with an additional phosphate group (H_3PO_4). This is typically done within mitochondria or chloroplasts, by means of a process called 'chemi-osmosis': an electrical potential difference and/or a concentration difference (pH gradient) is set up across the inner membrane, by initial transfer of electrons which then "pump" protons out against the potential/pH gradient. This generates a 'battery', which then drives protons back across the membrane where they convert ADP into ATP (and water). A larger membrane surface area, in principle, facilitates a larger production rate of ATP – hence these organelles often have a folded, or layered structure.

One can distinguish the following mechanisms for energy generation (and manufacture):

- ⋆ Respiration: oxygen gas is inhaled to be used in (ærobic) burning of carbo-hydrates. It produces water, which may be used by the body in various ways including the disposal of toxic waste products. The main waste product of respiration, carbon–dioxide gas is released by exhalation.

- ⋆ Fermentation: this is an anærobic way of burning sugars (not using oxygen), leading to waste products such as ethanol, lactic acid, or methane gas.

- ⋆ Photo-synthesis: the energy carried by photons emitted by the Sun is used to bind carbon–dioxide gas from the air to water. This process takes place for instance in cyanobacteria or in the chloroplasts of plant cells. The waste product, oxygen is released as a gas to the air.

- ⋆ Chemo-synthesis: where sunlight lacks, other sources of energy and nutrients are sometimes available. For instance, in hydrothermal vents the energy in the super-

heated water instantiates chemical disequilibrium, which can be used to power chemical reactions that can involve a range of molecules including sulfides and methane.

After rapid exhaustion of any pre-existing ATP (about ten seconds in humans), anærobic burning of glycogen in the muscle is an inefficient but very prompt way of releasing ATP (2 ATP molecules per glucose molecule) for a duration of around a minute. This mode of burning may be a relic from when the oxygen levels in the Earth's atmosphere were still very low. The ærobic burning of glucose, first, and fat, later, is much slower but also much more efficient (36 ATP molecules per glucose molecule) and therefore lasts a lot longer (about two hours of sustained intensive activity).

Note that oxygen is vital for organisms relying on respiration, but that oxygen is a very reactive element. In large concentrations it is a toxic, leading to oxidization ("rusting") of many important molecules and forming radicals that can cause chemical havoc. This is a major agent of the aging process and one of the limiting factors to longevity of life.

Co-existence of organisms relying on respiration and organisms relying on photo-synthesis leads to a natural balance in the oxygen and carbon–dioxide content of the air, as a result of feedback between the rate of production and the rate of consumption of these two gases.

Chapter 3

Life as we know it – functioning

3.1 Physiology

Complex organisms (e.g., humans) have evolved to develop multi-cellular components ('organs') designed to perform tasks such as the processing of nutrients (stomach), the storage of energy (liver), or cleansing of body fluids (kidney). A circulation system transports fluids (blood, pumped by the heart) and in warm-blooded animals helps maintain a constant body temperature. Sensors (such as eyes and ears) capture information about the environment and the relation of the body to that environment, and information and instructions (to muscles) are transmitted by means of a nervous system, and stored and controlled in a brain. The design of the body is adapted to the manner of movement (limbs, wings), the medium in which it operates, and the body mass distribution. As an example, we here focus mainly on the human body.

3.1.1 The cardio–vascular system and lungs

The heart pumps around warm blood. The contraction of the heart increases pressure which pushes the blood through vessels. As the width (or branching) of the vessels varies across the body, the speed of the blood stream varies according to the continuity equation ("what goes in must come out", at the same rate):

$$\mathrm{d}\left(\frac{\mathrm{d}V}{\mathrm{d}t}\right) \;=\; \mathrm{d}(Av) \;=\; 0, \tag{3.1}$$

where V is the volume, v is the speed, A is the cross section of the vessel and t is time. The pressure within the vessel is related to the speed according to Bernoulli's Law:

$$\mathrm{d}\left(P + \frac{1}{2}\rho v^2\right) \;= 0, \tag{3.2}$$

where P is the pressure ($= F/A$ for a force F) and ρ is the density of the blood. We see that when the vessel narrows the blood speeds up but the pressure drops.

While the blood warms surrounding tissue, the body loses heat (Q) at a rate $\mathrm{d}Q/\mathrm{d}t$ which, for a specific heat capacity C, results in the temperature dropping at a rate $\mathrm{d}T/\mathrm{d}t$:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} \;=\; C\,\frac{\mathrm{d}T}{\mathrm{d}t}. \tag{3.3}$$

Conduction of heat from the skin into the air is exacerbated by air flow, e.g., convection or wind. Heat is also lost through radiation, at infrared wavelengths according to Wien's Law and at a rate prescribed by Stefan's Law for a blackbody (or some fraction thereof):

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\sigma A T^4, \tag{3.4}$$

where σ is the constant of Stefan–Boltzmann. On the other hand, the body can get rid of excess heat through perspiration, in which latent heat is extracted by vapourization of liquids pressed out through pores in the skin.

Note that muscle action also produces excess heat, as in accordance with the Second Law of Thermodynamics energy cannot be converted into work at 100% efficiency – there must always be waste heat.

The lungs are intrinsically related to the cardio–vascular system as this is where blood captures oxygen. The contact surface between the blood vessels and the air is greatly increased by means of a branched system of 'alveoli'. The inhalation (of oxygen-rich air) and exhalation (of air rich in CO_2) is governed by the equation of state for an ideal gas:

$$PV = nRT, \tag{3.5}$$

where n is the number of moles and R is the gas constant. The work (W) done in the process – and thus the energy spent – is given by:

$$\mathrm{d}W = P\,\mathrm{d}V. \tag{3.6}$$

Under normal conditions the blood is saturated with oxygen, and it is not the breathing capacity but blood circulation speed that determines the oxygen uptake. However, breathing needs to be able to remove CO_2 at a sufficient rate, otherwise the acidity level of the blood rises. If the heart beats faster then, in principle, the blood pressure would go up. However, the muscle capillaries open, thus equilibrating the blood pressure following Bernoulli's law.

At reduced air pressure (for instance at high altitude), breathing starts to matter more as the partial pressure of oxygen in the inhaled air drops. Because the boiling temperature also drops, at low enough pressure the blood starts boiling spontaneously. On the other hand, at high pressure (for instance when diving) the high partial pressures of the inhaled nitrogen, oxygen and CO_2 render them toxic. Also, high pressure forces gases to dissolve into tissue. When the high pressure is reduced quickly (for instance during rapid ascent after a dive), nitrogen can get out of solution in the blood and cause bubbles.

While hæmoglobin carries the oxygen around in the blood, myoglobin keeps it available in the muscles. Underwater or deep-diving mammals have relatively large concentrations of myoglobin, as well as relying more on anærobic metabolism.

3.1.2 Sensors

Different organisms employ a range of sensory mechanisms, including touch, smell, and magnetism (for instance migratory birds and magneto-tactic bacteria that carry magnetic crystals as compass needles). Here we briefly consider the audio-visual sensors.

The eyes provide us with an image of the surroundings as they capture electro–magnetic waves (light). The optical system of the human eye involves two lenses – the 'cornea' and the 'lens' – of which the cornea does most of the refracting as the difference in refractive index between the air and the watery liquid in between the cornea and the lens is larger than that between the liquids at either side of the lens. Contrary to the cornea, the lens is flexible and muscles can deform it (by deforming the eyeball) to change the focal length, from ≈ 1.7 cm when staring at a distant object to ≈ 1.4 cm for a nearby object. Under water, however, the human eye cannot achieve focus; obviously this will be different for the eyes of animals that do rely on under-water activities.

A diaphragm, the 'pupil' can be adjusted in aperture to adapt to different light levels. Because the angular resolution depends on the aperture of the optics this would mean that finer detail can be seen in the dark. However, the light detector, the 'retina' has two types of receptors, 'rods' and 'cones', of which the cones provide the higher resolution; but the cones are not responsive under dark conditions so only the rods are used when it is dark. Both have a logarithmic response, and together they have the astonishing ability to span a range of 14 orders of magnitude in light level sensitivity.

The human eye can see light with wavelengths between $\approx$ 370–760 nm (violet to red). The significance of the cones is that they provide colour vision by way of separate receptors for blue, green and red photons. The rods are blue-sensitive and lack response to red light. This is why blue light from light-emitting diodes (LEDs) and electronic screens is both tiring (because the nervous system has to process strong signals) *and* inhibiting sleep.

Most organisms that have visual sensors have at least a pair of them, which allows them to see in stereo and thus estimate distances and speed through the parallax effect (each eye seeing the same object under a different angle). Also, the angular resolution is better in the plane of the eyes as the pair of scopes constitute an interferometre, with the resolution determined by the length of the baseline joining the two scopes. Hence our eyes are aligned parallel to the horizon (though climbing animals such as spiders might benefit from vertical alignment).

The ears enable us to sense sound, which is a longitudinal pressure wave travelling through a medium. The human ear is receptive to waves within a frequency range between $\approx$ 16 Hz – 20 kHz, i.e. a 600 times broader range than the response of the human eye to light. Like the eye, the ear can sense a huge range in intensity, over 12 orders of magnitude.

The wavelength (λ) of the audio waves is related to the frequency (ν) of the oscillation via the group velocity (sound speed, c_s): $c_\mathrm{s} = \lambda\nu$. The sound speed depends on the pressure (P) and density (ρ) of the medium:

$$c_\mathrm{s} \;=\; \left(\frac{\partial P}{\partial \rho}\right) \;=\; \sqrt{\frac{K}{T}}, \tag{3.7}$$

where K is the 'bulk modulus' ("stiffness") of the medium, which for a gas is related to the adiabatic coefficient γ (ratio of heat capacities) as $K = \gamma P$. In air, $c_\mathrm{s} \approx 330$ m s^{-1}. Hence the human ear is receptive to sound waves between $\sim$ 2 cm – 20 m in length. This is generally larger than the size of the ear, so resonances that could hurt the ear are only important at the highest pitched sounds but at lower frequencies other parts of the body (e.g., the heart and lungs) may resonate.

The ear is composed of an outer ear where the eardrum receives the sound wave, which is then transformed through the middle ear to a vibration in the inner ear where, within a narrowing coil ('cochlea'), the vibrations are detected by dendrites ("hairs")[1]. In order to accommodate variations in external pressure (for instance under water, or at altitude) the eucharian tube enables a certain degree of pressure rebalancing in the middle ear.

The ears also sense our orientation in space, by way of dendrites within the 'maculæ', and changes in our orientation (acceleration), by way of flexible appendices ('cupulæ') in the vestibular organ, both within the inner ear. The orientation of the body is determined by means of gravity, which would pull the dendrites into a different direction depending upon the orientation of the ear (and rest of the body). The acceleration is determined by means of the inertia of the cupulæ, which would also change their orientation as they lag behind the motion of the rest of the ear.

As with the eyes, most organisms that have audio sensors have at least a pair of them. The stereo-phonic effect allows them to determine the direction from which the sound waves came, and the parallactic effect yields an estimate of the distance from which they originated. Receding waves spread out (lowering the pitch of the sound) while approaching waves catch up (resulting in a higher pitched sound). This 'Doppler effect' allows us to sense changes in travel of the origin of the sound. Some organisms, like dolphins and bats, use echo-location to determine distances to objects; the objects that can be probed are of similar size to the wavelength of the sound – hence bats use ultra-sound to locate insects.

The nose remains mysterious. While some animals have far better sense of smell than humans, it is extraordinary that we can smell something without touching it – this must mean our nose can detect minuscule concentrations of volatile molecular matter. A single molecule can attach to a nerve ('olfactory cell') in the nose, and thus be detected by the brain; different nerves are good at responding to different odor molecules. It is similar to taste, however taste is much cruder and the sensation is greatly enhanced by the simultaneous sense of smell. Having two nostrils gives us a stereo-olfaction smell of the (near) environment.

Touch is a mechanical, thermal or electrical stimulus of piezo-proteins close to the skin, that transform it into an electrical pulse which is sent up the neurological infrastructure.

3.1.3 The nervous system and brain

Information from sensors and instructions to muscles are transmitted via the nervous system – an extensive network of "electrical wire" ('axons'). The core of the neuron (made of electrically conducting carbonaceous material) acts as a waveguide for the electro–magnetic waves that run along the outside, in turn coated by myelin to insulate them from the surroundings. Where the nerve cell-end ('synapse') meets a sensor or muscle, chemicals called 'neuro-transmitters' are used to pass on information. A difference in concentration (c) of electro-lytes (ionic solutions) on either side results in – or from – an electric potential difference (ΔV; Nernst's equation):

[1] This is not too dissimilar from the reception of electro–magnetic waves by a radio telescope, in which case the signal is converted to a different frequency range using a 'local oscillator', which is subsequently analysed with a 'correlator'.

$$\frac{c_1}{c_2} \;=\; \exp\left(-\,\frac{e\Delta V}{kT}\right), \tag{3.8}$$

where e is the elementary unit of charge (1.6×10^{-19} C) and k is the constant of Boltzmann. Neurons can "fire" at 0.001–0.01 s intervals, but the potential remains for 0.01–0.1 s.

In more advanced organisms, the nervous system has a central hub called the 'brain', which is where information is processed, stored, and instructions are being sent out from. We shall delve into a little more detail in Chapter 4.

Not surprisingly perhaps, organisms we think of as more intelligent generally have larger brains. However, brain mass also increases with body mass, though not as much as in direct proportion, e.g., among mammals: $M_{\rm brain} \propto M_{\rm body}^{0.7}$. But while the blue whale is a rather intelligent creature one would probably rate its intelligence inferior to our own. In fact, humans stand out from the above average relationship in that their 'Encephalization Quotient', $EQ \equiv M_{\rm brain}/M_{\rm body} = 0.020$ is the highest among all – if even at a meagre 2%!

3.1.4 Motoric attributes

The movement of animals or certain mirco-organisms is subject to the laws of mechanics. Movement requires force (or torque, or pressure), and equilibrium means the net force (torque, pressure) is zero. Movement is performed of – and with respect to – the centre-of-mass, particularly tricky for bipeds. The force is provided either by rigidity or by muscle contraction upon electrical impulses and chemical stimulants. While it is well known that calcium is essential for healthy bones, calcium ions also play an important rôle in triggering muscle contraction. Muscles relax rather slowly so it is beneficial to have opposing pairs of muscles so one can pull back the other.

Limbs have been perfectionized for their purpose. For instance, animals generally stride in an elegant manner, to maximise agility, efficiency and effectiveness. Friction is crucial to gain the desired "reaction" from an opposed "action" in the opposite direction of (intended) movement. Optimal use is made of the ability of the body to absorb and react, with the spring force being proportional to the displacement: $F_{\rm spring} \propto -x$ (Hooke's Law). Jumping insects such as fleas have had to develop mechanical linkage – including gear trains – to synchronize their legs in order to guarantee balanced launches.

Animals moving through water use flippers where the increased surface area provides more pressure to push against (Newton's 3$^{\rm rd}$ Law of "action = −reaction"), as $F = PA$. The viscosity of the fluid matters, too. The same principle is used in the wings of flying animals such as birds. The latter also make use of Bernoulli's Law to generate lift, by creating an under-pressure above the wings as air there flows faster as it has to avoid the curved barrier at the front of the wing.

Drag on the body when moving through a fluid with a certain viscosity is either minimised or maximised, by skin-type and shape and following:

$$F_{\rm air} \;\propto Av^2 \text{ (macroscopic)}, \tag{3.9}$$

or:

$$F_{\rm liquid} \;\propto A^{1/2}v \text{ (microscopic)}. \tag{3.10}$$

Note that the work done, $W = Fx$ and the energy in motion, $E = \frac{1}{2}mv^2$ so the energy expended in movement depends on mass, speed but also the way the body parts move (Fx really is $\vec{F} \cdot \vec{x}$).

Body parts have a certain rigidity and flexibility and can only withstand a limited force until they fail (e.g., tear, break). This is true for limbs (bones, tendons, muscles, flesh, skin) but also organs, eyeballs, eardrums... even cells. Young's modulus (cf. the bulk modulus used in Equation 3.7) quantifies how stiff the body part is – it is the ratio of stress (applied pressure) over strain (relative deformation). The tensile strength quantifies how strong it is – it is the stress at which it fails. The shear modulus is similar to Young's modulus when forces act along the surface, in opposite directions at opposite sides of the object.

For example, human bone has a Young's modulus of $\approx 10^{10}$ Pa and a tensile strength of $\approx 10^8$ Pa. When diving, the pressure increases (from the weight of the water column above you) by one atmosphere (10^5 Pa) every ten metres. At the bottom of the Mariana Trench, therefore, at a depth of more than ten kilometres, human bones will be deformed by $\approx 1\%$ and break. In reality, sustained high pressure (or shear force) will lead to failure already below these threshold values if given enough time – this is called 'creep'. Likewise, repeated exposure to high pressure (or shear force) eventually can lead to failure – this is called 'fatigue'.

3.2 Growth and decline

Multi-cellular organisms start (almost) from scratch, often from a single cell (animals). Cell division and specialisation work from a "blueprint" encoded in the DNA to replicate the product of billions of years of evolution in a matter of days, weeks or months. Initial, embryonic development often takes place within a protected environment with guaranteed nutrition, such as the womb, an egg or a seed or bean. Once more autonomous after birth, hatching or sprouting, further development takes place until maturity. In some cases this requires a switch of functioning and thus a fundamental change in physiology ('metamorphosis') which requires activation of the growth of hitherto dormant facilities (e.g., wings, proboscis and proper eyes of a butterfly) whilst now-redundant organs die off (e.g., crawling feet and light sensors in a larva).

Once reached maturity usually a process of "aging" takes over, leading to a gradual deterioration of our physiology and sometimes also our mental capacity. If we can identify and understand the processes that limit life span then we might find a way to postpone – or even prevent – death.

The longest living animals known on Earth are sea dwellers – Greenland Shark, Bowhead Whale, sea urchins... They have a lifespan of a few centuries. The longest living plant known is a kind of pine, of 5,000 years old. Colonies of connected plants can live much longer, up to 80,000 years for a kind of Aspen. Very slowly metabolising microbes have been found in 100,000-year old sediment underneath the ocean floor, but it is unclear in what state exactly they are. On the other extreme end, some insects don't even live another day. Humans, therefore, are relatively long-living creatures. But our finite lifespan may constrain the ultimate reach of our intellectual capacity and our ability to

forge further technological revolutions. On the other hand, evolution through natural selection requires limited life spans, otherwise the population growth would eventually become unsustainable.

While we all take death for granted, should we? If we could maintain our body and prevent damage, perhaps replace parts if and when necessary, should we not be able to continue living as long as we wish? One could argue that one cannot keep repairing and replacing parts of the human body, to the point at which we would not have a single original cell left in our body. But this is exactly what is happening already: almost all atoms in our body are replaced within a decade, some (e.g., in the gut) within less than a day! So the human body is not a one-off creation, but a transient phenomenon.

So why, then, does life end? Some argue that the causes of death at old age are no different from those that could affect us at a younger age: illness (infection) or damage. In humans, the production of growth hormones declines with age and the body will take longer to regenerate, to the point at which it can no longer keep up with accumulating cell damage. Free radicals, causing unwanted chemical and physical reactions inside cells, are being blamed for some of that damage, though the efficacy of anti-oxidants to slow down aging is disputed. However, some organisms do not age, in the sense that their mortality rate does not increase with time – they are biologically immortal. An example is the small Hydra animal, which can regenerate fully indefinitely. They would still fall prey to predators, just like we do: while we have mastered our macroscopic fiends, we are still fighting a battle with our many microscopic enemies.

With complexity comes the need to manage. Apparently this requires a state of suspended activity ('stasis'), in animals referred to as 'sleep'. We will return to this critical necessity within the context of the human brain, but for now it suffices to note that all animals die of sleep deprivation before they would die of starvation, and that even very simple organisms take rest. The daily pattern of life is regulated by the brain and through secretion of hormones, in a 'circadian rhythm'.

3.3 Extremophiles

Some organisms thrive under conditions which are too extreme for most other organisms. The most common types of extremophiles, all simple organisms, are:

- ⋆ 'Thermophiles': temperature loving (be it hot or cold).
- ⋆ Acidic loving.
- ⋆ 'Halophiles': salt (alkali) loving.
- ⋆ Radiation loving.

The latter, radiation-loving extremophiles are considered as a potential solution to the waste disposal problem associated with nuclear power plants.

Most proteins change structure or are destroyed above 50 °C. Organisms adapted to such temperatures rely on strong, saturated fatty acids, while hyper-thermophiles do not rely on

fatty acids at all. On the other hand, cold-loving organisms rely on flexible, unsaturated fatty acids. Freezing is dangerous to most cells, as it causes expansion of the plasma; freezing also presses minerals out of solution and thus increases the salinity – which could cause sharp crystals to form which may damage tissue.

In the context of finding life elsewhere in the Universe, the 'bacillus infernus' reminds us that even if the most obvious places to consider are found to be hostile, life might exist in perhaps more exotic places. This bacterium is found to thrive at depths of several kilometres below the Earth's surface, a sweltering place devoid of sunlight. For all it cares, Earth's surface and atmosphere might be most inhospitable.

Note the important difference between thriving, meaning being able to grow and reproduce, and merely tolerating. For instance, the ultimate survivor, tardigrades (more affectionately called 'water bears') are considerably more complex organisms a fraction of a millimetre tall. They have been known to withstand temperatures as low as 1 K and as high as 151 °C, pressures as high as 6000 atm., immersion in liquid alcohol and dehydration for 120 yr. They survive by assuming a state of suspended animation: they switch off whilst being able to switch back on when conditions become more pleasant again. Tardigrades are therefore not extremophiles.

Chapter 4

Mentality

4.1 The brain

While the workings of the brain are probably based on the same electro–chemical processes as elsewhere in the nervous system, it is not yet understood *how* information is processed, stored and recovered, nor is the concept of 'consciousness'. An in-depth treatment of the workings of the brain goes well beyond the scope of this course, but we shall here consider a few aspects related to mental states (the 'mind') and artificial intelligence.

There are a couple of non-invasive techniques available that can be used to examine the functioning of the brain. An electro–encephalogram (EEG) can detect and locate brain activity, for instance in the form of electrical patterns. It essentially treats your head as a battery, and by sticking two (or more) electrodes on your skull it registers electrical currents. But unlike with DNA we have yet to "decode" the signals; and like with DNA, we do not yet know *how* those particular signals cause a specific effect (we have identified the genes that determine pigmentation but do not know how that particular sequence of nuclear base pairs results in differences in the production of pigment). A computed tomography (CT) scan reconstructs a 3D image of bones and bodily fluids on the basis of X-ray scans, while a nuclear magnetic resonance (NMR) scan does this by using a magnetic field and radio pulses to map atomic nuclei with a certain magnetic moment. Within the context of understanding the brain, such scans mainly help identify which areas are responsible for functions that are adversely affected by trauma (e.g., a tumor or hæmorrhage) – though there are many reports that some of these functions can, in some cases, be taken up in another part of the brain.

Chemicals can alter the chemical or electrical functioning of the nervous system including the brain. The rôles of agents such as dopamine and adrenaline in the mood and alertness states are well documented and chemical treatment of depression can be successful (as is performance enhancement). Drugs can also cause hallucinations, which might tell us something about how our perception and imagination works, but their effect can be lethal. Surprisingly, still very little is known about the possible risks to our brain from external electrical and/or magnetic fields, such as for instance frequent and prolonged mobile phone usage.

4.2 Mental states

The functionalist view of mental states is that they are causal, abstract functions that link sensory input and behavioural output or other mental states. This makes sense from an evolutionary point of view: they have a clear rôle to play, and benefit the overall functioning of the organism. A simple example is when tissue is being damaged. This would send electro–chemical signals through the nerves into a region of the brain where 'pain' is experienced. As a reaction, electro–chemical signals are sent back through nerves to the relevant muscles where they induce body motion aimed to remove the cause of the tissue damaging. Ancillary reactions such as yelling may also serve a purpose, to bring relief or alert others of possible danger. While this all makes perfect sense, the question arises as to how the decision is made for a certain "output" given an "input", and why one must be *aware* of it. It would be wonderful if the input–output chain of events could work in a purely mechanistic way, as in a switchboard, without feeling pain; but people who do not feel pain lack a vital link and as a result have no impulsive response to harm.

In an attempt to conceptualize this, the idealised 'Turing machine' realizes input–output relationships by means of internal states and causal links. Hilary Putnam, the father of functionalism, recognised that those internal states themselves can be realized in different ways. For instance, the above example of pain mediating a reaction to harm would be triggered and executed in a different biological and physiological way in a human and in an octopus. But the internal (mental) state is the same, namely 'pain'.

But what *is* pain? Is it a physical object? An electrical charge on a capacitor, a concentration of chemicals, a wave of some sort... If it is, then it ought to be accessible to an outsider, such as a neuro-scientist. But so far it has eluded detection. This suggests that such experiences are 'epi-phenomenal', i.e. not physical. While the cause of the pain and the signals rushing through the nervous systems are very real, the experience of pain is virtual. Likewise, seeing that a tomato is red has two aspects: (1) wavelength-dependent brightness recorded by the retina and coded as a neurological signal is registered in the brain as a property of the tomato that is like the property of other objects the brain would recall as being "red"; (2) you experience a 'sensation' associated with the colour "red". It is impossible to describe this sensation, and it may well differ from how someone else experiences seeing something red. It is the sensation, called 'quale' (plural 'qualia'), that is the epi-phenomenal state, whilst the recorded information is treated in an underlying physical state. This suggests there are different levels, or kinds, of mental states.

One might think that sensory mental states related to perception are the more ready to understand, as it seems to be a mere matter of detection coupled with some simple information processing resulting in appropriate motoric response. Yet it is associated with mental states that are subjective: only accessible to the subject. Then what about propositional, or intentional, mental states such as 'belief' and 'desire'? Surely those must be even more subjective and less grounded in the physical? Perhaps not. A belief is a concept, *data*, not a sensation. It can be described. It might therefore be data that can be accessed from the outside, by someone else. While this may sound unsettling, the location of these kinds of mental states (in humans) is firmly established in the prefrontal cortex just under the forehead. And if that is not disconcerting enough, if these are – or supervene upon – physical states, abiding by the laws of Physics, then the way these

data interact with other mental states and, more importantly, mediates behaviour may be governed by the sequences of information entering through our sensory organs and out of our control.

4.2.1 Consciousness

Consciousness is an ill-defined term, because it comes in different guises, some of which we would perhaps not call a state of "mental awareness". For instance, one can be fully awake, even alert, yet perform complicated tasks without realising it, or remembering it. For instance, driving a car. This "zombie" state can hardly be called "conscious", yet the body was fully aware of what was going on around it, and fully prepared to respond. So this was not an "un-conscious" state either. We call this type of consciousness 'access consciousness', as it is directly linked to the Turing machine processing of input–output, in opposition to 'phenomenal consciousness' which is linked to the kind of awareness we "notice" (while we may have been aware of the red car in front of us, and knew it was red, the quale "red" was not sensed – i.e. its mental state was not instantiated).

Bernard Baars described consciousness as a "theatre of the mind", an arena of thoughts and sensations in which phenomenal awareness (perception) "speaks" to functional (cognitive) awareness. It is not evident, though, that qualia affect cognition or effect mental states that mediate action. The sense of a "nice" smell (an epi-phenomenal mental state) may *seem* to instantiate the belief (a propositional mental state) that something is "delicious", which in turn instantiates the desire (an intentional mental state) to eat it, which in turn causes the right neurological stimuli of the relevant body parts in order to engage in the act of eating. But how can a non-physicality (an epi-phenomenal state) interact with a physical system (all the rest)? It might well be that the quale is a mere by-product, and the real interaction takes place directly, between the neurological translation of the composition of the air as sensed by the nose on the one hand, and the information-processing cognitive states (belief, desire and the command to action) on the other.

Indeed, it appears that only qualia require phenomenal consciousness, whilst all other mental facilities only require access consciousness. But what about the 'sub-conscious', meaning without direct access by cognitive mental states? The rôle of the sub-conscious is not clear. It could range from a mere passive repository of superfluous data to an active, powerful processor where our personality and geniality are formed and shaped. Its existence seems in little doubt, as evidenced by the restoration of lost memory and beliefs and actions that seem unexplained on the basis of conscious mental states. So information *is* exchanged, somehow, between the conscious and the sub-conscious, both ways. But how, and when? Is the sub-conscious active while we are awake, or exclusive to our dormant states? Are we ever fully 'un-conscious', unless brain-dead?

One requirement for consciousness at a level of self-awareness may be the facility to conceptualize. It may seem that qualia would be eminently suited to this task, being abstract but recognisable experiences. But qualia cannot be described or quantified. Linguistic abilities would naturally offer a terrific benefit to describing, memorizing and communicating conscious engagements. When we think, we often "talk" to ourselves – it is hard to convince ourselves we are thinking without such "internal conversation". Speech may not be necessary; visual language might do – in that case we would perhaps

be "writing and reading" in our heads. It may be no coincidence that linguistic and intellectual development closely correlate, both throughout the late evolution of hominids and through childhood. Human infants tend not to hold conscious memories until they are a couple of years old, when they start mastering language in earnest. The use of language among (other) animals may therefore offer us an intriguing "look" into their minds. Despite their lack of conceptual awareness, infants do learn prodigiously, but mainly by copying and trial-and-error – like other animals do. Do infants experience qualia – does one need conceptual awareness to know one is experiencing a quale? To develop reasoning, the capacity to deduce, imagine and generate original thought, self-awareness may not just be a great aid – it may be essential. We cannot, therefore, brush it aside as a mere curiosity as we tried to do with qualia. It must have a physical connection, and therefore be physical. This gives us hope of the possibility to apply the scientific method to study and eventually understand it.

4.2.2 Sleep

'Sleep' is a fundamental necessity for all mental life. That seems odd if sleep were a mere period of inactivity, of absolute rest. If nothing happens during sleep, then how can it be that important? Surely, something *must* be happening while we are asleep, and it must be vital.

While our body performs its many critical and optional functions, chemicals are taken up, transported, modified and stored or expelled. It is not surprising, therefore, that maintaining a stable environment within a body is crucial but presents a significant challenge. Just as garbage is collected and streets are sweeped after a busy time in town, in addition to the continuous processing of sewage, the body needs to perform clean-up operations. Whether it is lactates after intense and prolonged muscle activity, or rubbish clogging up our brain matter, disposing of it takes some time and would benefit from the temporary shutdown of non-critical systems. Hence sleep is accompanied by partial paralysis.

Likewise, the electro–magnetic signals that have been communicating throughout the nervous system and been used to store information into memory will no doubt have saturated some capacitance leading to hysteresis or buffer overflow, and have upset some chemical balance through electronic separation. To reset and restore the system it would benefit from drastically reducing the rate of sensory inputs. In particular visual data are substantial, hence why closing the eyes even for a brief moment can be very relaxing.

We know that sleep prioritises information processing, as this is what happens first. Clearing up memory, removing superfluous data (perhaps moving some of it to sub-conscious compartments) to make space for the storage of new data, apparently is super critical, even in non-human animals. We all know that one can only spend so many hours revising for an exam before saturation kicks in, gazing at painting after painting of your favourite style before becoming numb to the experience, or playing a computer game through level upon higher level. But it is sobering to realise that the body cannot cope with information on the go, and that information overload and processing strain can kill. When you feel your head hurts when concentrating or thinking hard, it means we are in a mental state that alerts us to possible harm. Reducing and reorganising data in memory during sleep causes us to 'dream', a form of hallucination.

In the second phase of sleep, physiological recuperation takes place including regeneration under the influence of growth hormones, and chemical cleansing. One particular peptide of amino acids, β-amyloid is critical in growing and repairing neural connections but it can become corrupted and then inhibit internal communication. This is linked to Alzheimer's, a debilitating form of dementia. Selectively removing the corrupted version but not the intact one might cure the disease.

While asleep, one is "unawake". But not (fully) "unaware". An external trigger can cause one to wake up. And dreams are experienced and sometimes remembered. Yet it is a small miracle that one wakes up from this comatose state, day after night.

4.2.3 The rôle of neurons in memory and consciousness

Plants don't have a mental life, alright? The presence of neurons seems a requirement for behaviour that we associate with some level of consciousness, and with memory. But one neuron does not contain information for very long. Chemical interactions lead to potential differences that "fire" the neuron. The neurological *network* might contain information, but this, too, is highly dynamic. The brain's inherent plasticity sees neural connections changing all the time. This is beneficial because it facilitates learning, moving data from short-term to long-term memory, and resilience against damage. One could also see a neuron, or neuron state, as an image pixel from the instantaneous "snap-shot" of a movie. But how then can memory be kept? There must be a more static configuration, too. Memory is lost when neural connections weaken, as a result of degradation or repurposing; these connections are strengthened (again) when learning (remembering).

Memory also is a condensed, incomplete account of events, often making use of mnemonics. The eye receives too much information to process or store, too, and a selection is made. This is done unconsciously – sleep-walking or 'blindsight' people use their eyes without "seeing". Socrates said it: "He who sees with the eyes is blind." (Meaning it is the mind that interpretes the world.) Reconstructing the vivid counterpart of a memory or data coming from the retina is akin to constructing an image from radio-interferometric data that represent a Fourier transform of the sky but one that is incomplete in its phase coverage. In fact, 'phase' is associated with relations, and if these pertain to neurological "traffic" then it may be the phase differences between electrical signals – rather than the electrical charges or currents themselves – that obtain emergent phenomena such as awareness and qualia. Phase is only defined dynamically, and this might explain why we sense the passing of time if this is what awareness is based upon.

4.3 Automata

Could consciousness emerge as a consequence of a system of causal links? Does it require a biological basis – and even if it does: could the physical mechanisms be emulated in a non-biological context? If this fails even if only for some aspects of consciousness, then that would imply that there are mental states that do not supervene on the physical. This would be difficult to reconcile with mental states requiring a physical body, and re-kindle debate on Descartes' mind–body substance dualism. There is thus a strong philosophical and meta-physical incentive to attempt to develop "true" artificial intelligence.

Computing machines, whether using gears, semi-conductors, capacitors or quantum states, process information. This can be in analogue form, but typically it is coded in discrete ('digital') format as it allows for sharper definition and a margin for error. If some of the mind's workings rely on the more subtle interplay between analogue signals then a digital computer approach may fail to replicate it. Computers can store data, retrieve it on demand, quantify, recognise and correlate properties, apply conditional processing and perform pattern recognition and formulaic calculations. These are all algorithmic, working from a script. Making computers autonomous is a challenge. Memory, neural networks (be it wired in hardware or programmed in software) and input–output causal relations are part and parcel of how computers operate. If propositional mental states can be reduced to the conditioning of causal relations as a result of previous triggers of mental states in combination with memory, then implementing this seems within the realms of possibility. But if you ask a computer whether it is self-aware, and it responds positively, do you believe it? The computer will have registered the question and analysed its meaning; conditional relations may then have triggered some diagnostic checks of memory and currently running processes, which would have returned a satisfactory level of activity; this produced the output. Did the computer "feel" it was awake, or just worked it out clinically? Did it "know" it existed, or just existed?

'Automata' (singular 'automaton') are abstract systems (machines or schemes) that have a degree of autonomy based on formal language. It is characterised by internal states, with transitions of those states occuring on the basis of input and current states following transition functions. The concept of an automaton is a mathematical object and forms a basis for computational theory, but automata are also implemented in practice. Board games are not unlike automata and the reason some people like board games is probably because it helps them hone their own logical skills, which can be beneficial to one's place in society and thereby one's wealth and well-being, and ultimately in an evolutionary sense. While automata usually take in one input at a time it is possible to construct a composite automaton which can deal with multiple, simultaneous input streams as we experience in real life. The transition functions can be deterministic (only one possible outcome) or non-deterministic (any one of multiple outcomes), and likewise input may be accepted strictly or with a degree of probability. A non-deterministic, probabilistic automaton could possibly allow for such human behaviour and conditions as mistakes, anxiety or imagination. The transition functions are often described in a language based on context-free grammar, introduced by Noam Chomsky to formally describe how sentences are built recursively from smaller blocks. This allows for simple logic, insensitive to input or memory, but we do not know whether this is true also for the mind.

Automata can also be used to represent populations and ecological systems, not just the workings of an individual organism or component. An example, the 'cellular automaton' describes the growth, competition and evolution of groups of "cells", which could also represent organisms within a society. John Conway's 1970 'Game of Life', based on John von Neumann's concept of a self-replicating automaton, consists of an initial configuration of cells on a grid, after which simple relational rules determine whether a cell persists, perishes or creates an adjoining cell. The rules can be changed to deal with additional dimensions or to modify the sensitivities to population stress, starvation and reproduction. These purely mathematical automata have resulted in surprising examples of self-organisation and emergent behaviour.

Chapter 5

The emergence of life on Earth

5.1 How did it start?

5.1.1 Tree of Life

Great progress has been made over the past half a century in deciphering the DNA genetic code. Without knowing what all the different bits ('genes') of the DNA mean, they can be extracted ("mapped") and simply compared between organisms. In this way it has been possible to establish likely evolutionary lineages: the 'Tree of Life'.

The picture that has emerged is of three main groups of lifeforms:

- ⋆ Bacteria.
- ⋆ Archæa, most of which are extremophiles.
- ⋆ Eukarya, to which humans belong.

Bacteria and Archæa are able to exchange genomes so they cannot really be considered completely separate branches of the tree. They are both prokaryots, which excrete enzyms and ingest the products, rather than the internal metabolism made possible by the organelles (and nucleus) within eukaryots.

The last common ancestor of animals existed ≈ 0.8 Gyr ago. Curiously, the oldest animal fossils date back to ≈ 0.6 Gyr ago. Perhaps the ancestors of modern animals did not have body parts that would leave fossils. Animals form just a tip of the Eukarya branch, and the common ancestor of all Eukarya existed much earlier in the history of Earth. Eukaryots may have found their origin in prokaryots wrapping themselves around other prokaryots – this predatory process is called phagocytosis.

Intriguingly, the genetic lineages place the Archæa in general, and the hyperthermophiles (high-temperature loving) organisms in particular, nearest to a possible common ancestor. It is thus quite possible that the first lifeforms to emerge on Earth were hyperthermophiles living near hydrothermal vents on the bottom of the ancient Earth's ocean floor. These were probably not the famous 'black smokers' but rather the cooler, alkaline vents that are found further away from tectonic ridges, where the right pH gradients would have provided the right conditions for chemi-osmosis before cells had formed.

5.1.2 The Miller–Urey experiment

In 1953 (a productive year for life sciences!), graduate student Stanley Miller and his thesis director Harold Urey performed a seminal experiment. They attempted to recreate conditions as they might have prevailed on the early Earth: an atmosphere of methane (CH_4), ammonia (NH_3), molecular hydrogen (H_2) and water vapour (H_2O), with electric discharges mimicking lightning (lightning often occurs in volcanic eruptions). To their amazement, various organic molecules were created in the process, quasi-spontaneously. These included amino acids such as glycine (the simplest one).

Now, the Miller–Urey experiment has lost some of its "spark", in that most scientists accept that simple organic molecules are the natural consequence of simple processes taking place in common mixtures under ubiquitous conditions, but that it is a long, non-trivial way from these simple molecules to the organized structures that become alive.

5.1.3 One of life's mysteries: chirality

One of life's mysteries is the fact that it has a "sense": in a multi-atomic molecule the atoms can often be arranged in different ways, resulting in different versions ('isomers') of the same molecule. If a mirror-copy of the molecule can be turned such that it looks identical to the original version then the molecule is said to be 'achiral'. If the mirror-copy can never be turned such that all atoms overlap with those in the original molecule then the molecule is said to be 'chiral', with the isomers called 'enantiomers'.

As it turns out, 19 out of the 20 amino acids used by Earthly lifeforms are chiral (only glycine isn't). They have a sense: when placed on the palm of the hand with the carboxyl group resting on the fingers and the hydrogen atom on the thumb, if this is the left hand then the isomer is 'left-handed', and *vice versa*. Earth life appears to only use left-handed (L) amino acids; and only right-handed (D) sugars. It may be responsible for the helical structure of DNA, giving it stability, and making it easier to copy. Circularly polarized light might have led to preferential use of one over the other for reasons of energetics.

5.2 When and where did it start?

5.2.1 The changing conditions on Earth

Earth has gone through dramatic changes since it formed:

- ⋆ The Sun formed 4.55 Gyr ago. We know this from the ages of meteorites, which are left-overs from the formation process of the Sun. These ages were determined (in 1953) from isotopic abundance ratios where at least one of the isotopes is radio-active and therefore decays, for instance the uranium–thorium–lead (U–Th–Pb) method.
- ⋆ The oldest rocks on Earth are 3.8 Gyr old. But zircon minerals ($ZrSiO_4$) are found as old as 4.4 Gyr, and their oxygen isotopes indicate that water was already present.
- ⋆ Oceans and landmasses covered Earth by ~ 0.8 Gyr, ending the Hadean æon. The final 0.2 Gyr of this period were characterised by the 'Late Heavy Bombardment' of the Earth and Moon by debris, evaporating any reservoirs of surface water.

- ⋆ Earth's primordial atmosphere – formed by outgassing of volatiles from the interior – must have been composed mostly of CO_2, H_2O, and N_2. Oxygen may have been present in trace amounts, from photo-dissociation of water. Comets impacting upon Earth might have supplied water – but surface water was also lost by the impacts and Earth still contains seven times as much water in its interior as on its surface.

- ⋆ Any O_2 would be removed from the atmosphere, through the oxidization of rocks.

- ⋆ CO_2 would also be removed from the atmosphere, dissolved into oceans and deposited onto the ocean floor as carbonates, becoming part of Earth's crust.

- ⋆ But atmospheric CO_2 has been continuously replenished by volcanoes. Plates of Earth's crust float on top of molten, humid rock (magma), and shove underneath each other in subduction zones. This tectonic activity feeds volcanoes – blisters forming at the rim of these zones – with fresh supplies of carbonate-rich magma. Over time, the fraction of Earth covered by dry land has slowly increased.

- ⋆ The Archæan æon soon saw the emergence of life: stromatolites (bacteria colonies) have been found in rocks which radio-active dating tells us are 3.4 Gyr old.

- ⋆ The high concentration of CO_2 in the atmosphere would have led to a 'Greenhouse effect', warming Earth more than nowadays. This may have been exacerbated by the presence of methane (CH_4) from anærobic metabolism in microbes. This could have prevented Earth from freezing over, at a time when the Sun was fainter than it is now (this observation is called the 'faint young Sun paradox'). A darker surface (without ice or vegetation) may also have kept Earth's temperature mild.

- ⋆ Photo-synthesis took place by 2 Gyr (in cyanobacteria), supplying the atmosphere with oxygen, and resulting ozone. This 'Great Oxidization Event' (oxygen levels only reached 1%, though, compared to 21% today) started the Proterozoic æon. The availability of oxygen may have been critical for the appearance of eukarya, which have much greater energy needs than prokaryots do.

- ⋆ Fossil records show a dramatic change, with complex life emerging quite suddenly after ~ 4 Gyr in the 'Cambrian Explosion', starting the Phanerozoic æon. Perhaps extreme climatic change prompted this evolution, ending periods of global glaciation. Curiously, the Cambrian Explosion was preceded by the Ediacaran era of organisms that seem to bear no semblance to anything that has lived since – perhaps a "failed experiment" of Darwinian evolution? It could indicate that different outcomes might be possible on other worlds.

The early Earth was hotter than it is now:

- ⋆ Heat was released through shocks and friction, in the assembly of the planet by the agglomeration of smaller bodies (called 'planetesimals'), and by heavy bombardment with debris from the Sun's formation. The Earth was molten throughout, and covered by a magma ocean as late as 600 Myr after the formation.

- ⋆ The early Earth continued to contract. Differences in buoyancy between lighter and denser material led to internal differentiation ('gravitational settling'), leading to an iron core and a mantle rich in silicates. Gravitational potential energy was released in this process and, through increased pressure and friction, transformed into heat.

- ⋆ Radio-active decay continues to the present day, keeping the iron core liquid. The photons or particles emitted in the decay process are absorbed, heating the material. Initially, ^{26}Al dominated, brought by the winds of massive stars and their explosions.

- ⋆ Tides as a result of the gravitational pull by the Moon and the rotation of the Earth, through friction, also contribute heat but this is not important in Earth's case (though the Moon was closer to Earth in the past and currently recedes at 4 cm/year – the Earth's day was also much shorter).

5.2.2 The significance of the other planetary bodies

Two other Solar System objects (besides the Sun) may have played a critical rôle in the formation and evolution of life on Earth: the Moon and Jupiter.

Jupiter is the most massive planet in the Solar System and also closer to the Sun and Earth than any of the other massive planets in the Solar System. Not only does it affect the orientation of the Earth's rotation axis and orbit – and can thus be blamed (in part) for the coming and going of ice ages – but it also helped shape the early history of the Solar System. Its gravity caused and/or affected the migration of proto-planets and may have stabilised Earth's orbit to be where it is now, and made it roughly circular. Its gravity also may have diverted many rocks and planetesimals, avoiding their impact upon Earth, sparing it the worst of the bombardment that typified the young Solar System.

On the other hand, one major impact within the first 0.1 Gyr, of a Mars-sized planetesimal (called Theia) with the proto-Earth caused the removal of a fair fraction of the Earth's mantle, which condensed to become our Moon. Hence Earth is accompanied by a Moon that is relatively large (and massive) compared to Earth, making it the only comparable planet–moon system in the Solar System (Pluto, with large moon Charon, is not a planet).

The Moon may be responsible for at least three important effects on Earth:

- ⋆ The Moon stabilises the tilt of the Earth's rotation axis (and in the process it has synchronised its own spin and orbit so it always faces the same side to Earth). This has ensured modest variations between the seasons and a relatively stable climate.

- ⋆ The tidal force the Moon exerts on Earth causes regular rises and falls of water levels. Tidal pools may have provided critical habitat for early complex life. Tides also stimulate tectonics. The Sun also exerts tides on Earth, but they are weaker.

- ⋆ The impact that created the Moon may have caused the release of large amounts of water from inside proto-Earth. The oceans may be courtesy of this "disaster". In fact, it may be a rare occurrence to find *both* oceans *and* landmasses on a planet.

Furthermore, Theia may have left its core inside that of Earth. This could have made an important difference to the ability of Earth's core to generate a magnetic field.

5.3 The origin of life on Earth in a nutshell?

Looking at the evidence for the evolution of life from simple building blocks to highly sophisticated systems, there seems to be a fairly monotonic picture emerging, passing through several key stages of development (Fig. 5.1). These transitions may be linked to global changes in the character of the Earth and its atmosphere – be it the cause of those changes (e.g., the Great Oxidization Event) or the result (e.g., global glaciation or the end thereof). One may wonder how Universal this sequence of events is, how inevitable, and whether this is where evolution ends, or how it will proceed from here.

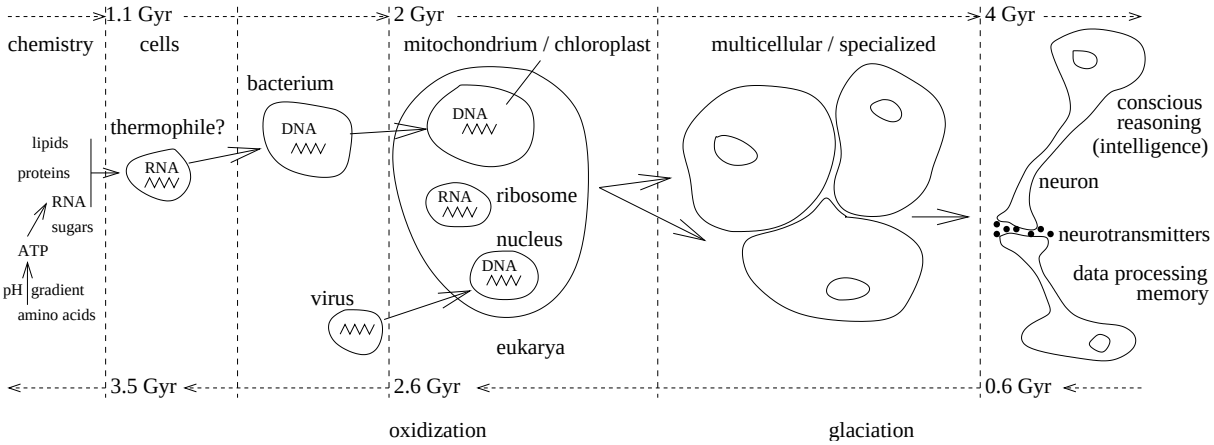

Figure 5.1: The emergence and evolution of life on Earth – a blueprint for other worlds?

Not surprisingly, the earliest stages of the emergence of life on Earth are still very sketchy. It is generally believed that RNA formed first. RNA is more stable in a cold environment such as may have existed between the cooling of the molten Earth and the reinstatement of a Greenhouse effect. DNA is more stable in general but it is broken down by ultraviolet light which was only blocked by atmospheric ozone since the Great Oxidization Event.

An early period of glaciation may also have helped concentrate chemicals and thus sped up the synthesis of complex organic molecules (as is the case in ice in space) and the formation of structures which were able to produce copies. Likewise, the global glaciation periods that preceded the Cambrian Explosion could have brought single-cellular organisms into prolonged contact, thus stimulating their bonding and eventual cooperation within multi-cellular organisms which eventually may have triggered the end of global glaciation.

Chapter 6

Habitable Zones and Earth's atmosphere

6.1 Conditions for a planet to be habitable

The Earth and Moon occupy essentially the same place in space, yet the Earth is teeming with life while the Moon is a barren, inhospitable piece of rock. Whence the difference?

We had realised that the presence of liquid water is an essential prerequisite for a planetary body to be able to support life (as we know it). The thermo-dynamical state of water is determined by the two environmental parameters temperature and pressure; this is mapped in the phase diagram which is presented schematically in Figure 6.1.

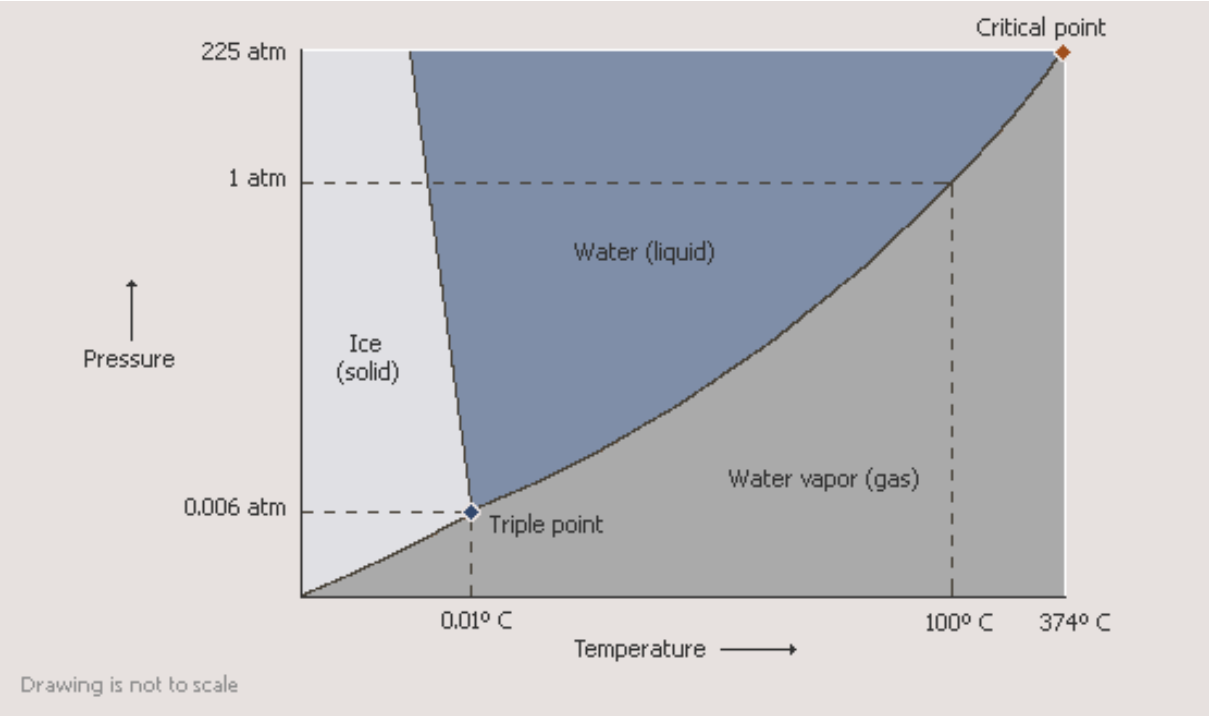

Figure 6.1: Phase diagram of water (H_2O).

In space, the pressure is much below the triple point, and hence no matter the temperature, water will never be found in liquid form. Sufficient pressure is therefore paramount. An atmosphere or surface crust functions as a "lid", preventing liquid water being in direct contact with the near-vacuum of space.

6.2 The temperature of a planet

6.2.1 Radiative equilibrium

A lone planet in space would be a cold place, but a planet orbiting a star is illuminated. The energy carried by the photons is absorbed by the planet, heating it. If that was all that happened, the planet would get increasingly hot until it would melt and evaporate. But a warm body radiates, carrying away energy. The rate at which a body emits increases with increasing temperature. If the temperature is sufficient for the planet to radiate away energy at the same rate as it absorbs starlight, an equilibrium is established.

Picture a star with a luminosity (energy emitted per unit time) $L_\star$ and a planet orbiting it at a distance d. The flux (energy flowing through a unit area per unit time) of starlight at the position of the planet is determined purely by geometry of 3-dimensional space as:

$$F_{\star,\circ} \;=\; \frac{L_\star}{4\pi d^2}. \tag{6.1}$$

A spherical planet, with a radius $R_\circ$, intercepts starlight flowing through a cross-sectional area $\pi R_\circ^2$ at a rate:

$$L_{\star,\circ} \;=\; F_{\star,\circ}\, \pi R_\circ^2. \tag{6.2}$$

To be in radiative equilibrium, the planet must emit energy at the same rate:

$$L_\circ \;\equiv\; L_{\star,\circ}. \tag{6.3}$$

If the surface, of total area $4\pi R_\circ^2$, emits at a rate $F_\circ$ per unit area (on average), then the luminosity of the planet is:

$$L_\circ \;=\; F_\circ\, 4\pi R_\circ^2, \tag{6.4}$$

where the flux is equivalent to that of an isothermal surface of temperature $T_\circ$:

$$F_\circ \;=\; \sigma T_\circ^4, \tag{6.5}$$

where $\sigma = 5.67 \times 10^{-8}$ J s^{-1} m^{-2} K^{-4} is the Stefan–Boltzmann constant. Hence we obtain a relationship between the planet's temperature and the star's luminosity:

$$T_\circ \;=\; \left(\frac{L_\star}{16\pi\sigma\; d^2}\right)^{\frac{1}{4}}. \tag{6.6}$$

Note that this depends on the distance from the star, but not (directly) on the star's temperature: a cooler, but bigger star can be as luminous as a smaller, but hotter star – in either case the planet's temperature would be the same.

6.2.2 Albedo

Equation 6.6 yields a temperature for Earth of $T_\oplus = 278$ K, or +5 °C, commensurate with the abundance of liquid water found on our planet. However, this assumes that all incident Sunlight is absorbed, which clearly it is not: clouds in the atmosphere, the polar ice caps, even foliage and sand reflect some of it.

The fraction of incident light that is reflected (in any direction) is called the 'Bond albedo', A. The true amount of absorbed light is then a fraction $(1 - A)$ of the incident light. Equation 6.2 now becomes:

$$L_{\star,\circ} \;=\; (1 - A)\; F_{\star,\circ}\; \pi R_\circ^2. \tag{6.7}$$

The efficiency with which material reflects often depends on the angle of incidence of the incoming light. For instance, a water surface is very poor at reflecting light falling on it under a right angle, but it reflects all light grazing it under a small angle.

The ice and snow dominating Earth's landscape at the poles have a high "natural" albedo. Light there always falls under a small angle, further increasing the "effective" albedo. The oceans have a very low albedo near the equator, where the Sun is generally high in the sky. The mean Bond albedo of Earth facing the Sun is $A_\oplus = 0.3$. The calculation for the Earth's temperature now yields $T_\oplus = 255$ K, or −18 °C. So the prospects to find liquid water on Earth actually look rather bleak.

6.2.3 Temperature variations

The temperature is not the same everywhere on Earth, for various reasons:

- ⋆ The latitude-dependency of the albedo results in the poles being colder than the equator. This is exacerbated by the fact that light incident under a small angle is spread over a larger area, thus diminishing the flux (remember: the flux is the rate at which energy flows *per unit area*). The lower incident flux necessitates a lower local flux to radiate the energy away, and hence a lower temperature to do so.
- ⋆ Earth's rotation causes a regular alternation of illumination by, and shielding from Sunlight. The side of Earth facing the Sun warms up, while the side facing away cools down. To maintain an average, effective temperature of 255 K, the day-side will be warmer than 255 K and the night-side will be cooler than 255 K.
- ⋆ Earth's rotation axis is tilted with respect to the axis of its orbit around the Sun. This causes the Northern hemisphere to be facing the Sun more directly during one half of the orbit (Northern Summer), whilst the opposite is true for the Southern hemisphere. These seasons lead to annual variations in local temperature, especially near the poles which are immersed in darkness for part of the year.
- ⋆ Earth's orbit is not perfectly circular, but slightly eccentric. This leads to annual variations in the incident Sunlight.
- ⋆ Earth's atmosphere traps warmth, thus heating up the Earth's surface (see below). However, at high elevations there is less atmosphere above the surface to do so, and hence the temperature drops with increasing elevation.

The oceans and atmosphere on Earth maintain a strong circulation pattern, transporting warmer water and air to colder places thus moderating the effects. But sometimes the opposite happens: the record low and high temperatures recorded on Earth are −89.2 °C (Antarctica) and +70.7 °C (Iran, in 2005)!

6.2.4 The Greenhouse effect

Earth's atmosphere is composed mainly of molecular nitrogen (N_2), oxygen (O_2), carbon–dioxide (CO_2), and water vapour, with traces of other molecules including ozone (O_3). CO_2 and H_2O have a similar molecular shape, like a wedge. These molecules can vibrate and rotate, and these extra degrees of freedom (in addition to translation of the molecule – a.k.a. thermal motion) are associated with the storage and release of internal energy, typically corresponding to electro–magnetic waves (or photons) with a wavelength in the infrared regime. Incident light is therefore efficiently absorbed by these molecules in specific infrared bands, rendering the atmosphere opaque at those wavelengths.

Di-atomic molecules such as nitrogen and oxygen are symmetric, and lack the internal modes of CO_2 and H_2O. They can still absorb light through electronic transitions as in single atoms. This happens at very specific wavelengths giving rise to narrow spectral lines. For di-atomic molecules these lines are broadened somewhat, because the electronic energy levels in one atom are broadened by the potential of the other atom. Electronic transitions and their resulting spectral lines are associated with higher energies and correspondingly shorter wavelengths (higher frequencies), generally in the optical regime.

The tri-atomic ozone is an extremely efficient absorber of still higher energy photons, rendering the atmosphere opaque to most ultraviolet light.

The large concentration of the relatively inert and transparent nitrogen has a stabilising effect on Earth's atmosphere.

The result is that the transparency of the atmosphere depends on the wavelength of the incident photon (Figure 6.2), with the following important consequences:

- ⋆ The Sun's photon energy distribution is quite well approximated with a Planck curve ('Black Body') of an effective temperature $T_\odot \sim 5700$ K, peaking in the optical. This is where Earth's atmosphere is relatively transparent, and hence much of the Sun's light reaches Earth's surface. Both factors, the Sun's temperature and the Earth's transparency, are key to many animals on Earth having developed eyes. Were the Sun to be much cooler, most of its radiation would be in the form of infrared light; were it to be much hotter, most of it would be ultraviolet light. In either case, less radiation would reach the surface and it would be colder.

- ⋆ Using Wien's law for the peak of the spectral energy distribution of a Black Body:

$$\lambda_{\rm peak}\, T \;=\; \text{constant}, \tag{6.8}$$

  one readily estimates that the Earth radiates mainly at infrared wavelengths, $\sim$ 5–20 μm. Earth's atmosphere blocks much of this light, only leaving a meagre few spectral "windows" of transparency. The result is that the radiation Earth emits is absorbed by the atmosphere, which then heats up until it is hot enough to radiate this energy

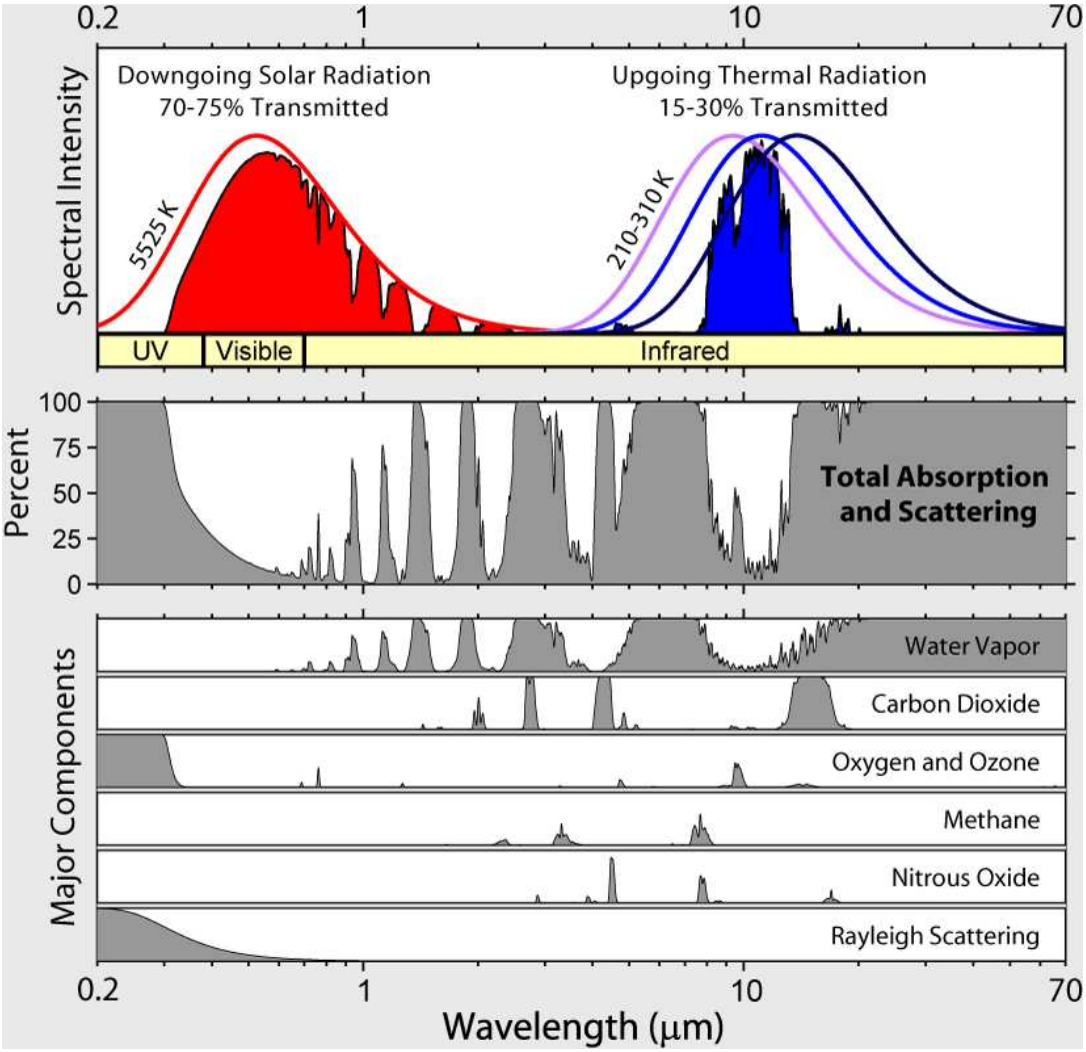

Figure 6.2: The opaqueness of Earth's atmosphere, with individual contributions.

away at the same rate it absorbs it. The radiation emitted by the atmosphere partly escapes into space, but partly illuminates the Earth. Thus, besides direct Sunlight, the atmospheric (infrared) radiation provides an additional source of warming the Earth. Hence, the average temperature on Earth's *surface* is higher than 255 K.

The warming effect the atmosphere has on Earth is called the 'Greenhouse effect'.

6.3 The Habitable Zone

Together, the starlight and the effect of the atmosphere determine the temperature of the planet's surface and hence the possibility for liquid water. This depends on the distance to the star, depending also on the star's luminosity. The range in distances from the star, where liquid water may be found on a planet's surface, is called the 'Habitable Zone'. For the Solar System, it ranges from just inside Earth's orbit to about the orbit of Mars. Only Earth is situated comfortably within it. It is important to realise that the exact boundaries of the Habitable Zone, and whether a planetary body within it is in fact habitable, depend very much on the properties and evolution of the planetary body itself.

6.4 The planet's atmosphere

We have seen that the presence of an atmosphere is a great asset for the prospects of life:

- ⋆ The atmospheric pressure can keep water liquid, provided the temperature is right.
- ⋆ The atmosphere, through the Greenhouse effect, provides an additional source of heat. This can make a difference in whether water can be liquid, making life bearable on an otherwise cold planet.
- ⋆ If ozone is present in the atmosphere, then harmful ultraviolet radiation is filtered out, protecting life against damage of its genetic material and hence malfunction.

6.4.1 Hydrostatic equilibrium

The pressure in an atmosphere depends on the weight of the atmosphere, i.e. its mass and the gravity of the planet. It depends also on the temperature of the atmosphere, as atoms in a hotter gas carry more momentum and thus transfer more impulse on whatever they encounter on their path. The vertical profile of the atmospheric pressure can be calculated fairly simply, if we assume that the atmosphere is static, that is, it does not change as a whole (for instance, evaporate into space). This is called hydrostatic equilibrium.

Consider a cubic parcel of gas in the atmosphere. For it to be static, the net force acting upon it must vanish:

$$\sum \vec{F} \;=\; 0, \tag{6.9}$$

for each of the three components separately. Choosing the altitude z to increase upwards, the downward force acting upon the bottom of the cube, with surface area A, is:

$$F_{\rm down} \;=\; -Mg - P_{\rm top}\; A, \tag{6.10}$$

where M is the mass of the gas inside the cube, g is the planet's gravitational acceleration (at that altitude), and $P_{\rm top}$ is the pressure exerted by the atmosphere on top of the cube. The upwards force acting upon the bottom of the cube is similarly given by the pressure exerted by the atmosphere below it:

$$F_{\rm up} \;=\; P_{\rm bottom}\; A. \tag{6.11}$$

Rearranging gives for the pressure difference between the top and the bottom of the cube:

$$\Delta P \;\equiv\; P_{\rm top} - P_{\rm bottom} \;=\; -\,\frac{Mg}{A}. \tag{6.12}$$

What we want is to get all dependencies on P and z explicit in the equation. The mass in the cube is given by the product of gas density ρ and volume $A \times \Delta z$. Replacing the macroscopic differences (Δ) by infinitesimally small ones (d), we thus obtain:

$$\mathrm{d}P \;=\; -\rho g\; \mathrm{d}z. \tag{6.13}$$

This is a differential equation, the solution of which is an expression for the pressure as a function of the altitude z. But we are not there yet: the density may still depend on the pressure, or on z (or both). The equation of state of a gas tells us the relationship between the density, pressure and temperature T. For an ideal gas, which is a good approximation for most planetary atmospheres, the equation of state is:

$$P \;=\; \frac{\rho}{\mu}\, kT, \tag{6.14}$$

where $k = 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant, and μ is the mean mass of the molecules that make up the gas. Using this equation of state, we obtain:

$$\mathrm{d}P \;=\; -\,\frac{\mu g}{kT}\, P\, \mathrm{d}z, \tag{6.15}$$

or if you prefer it ordered this way:

$$\frac{\mathrm{d}P}{\mathrm{d}z} \;=\; -\,\frac{\mu g}{kT}\, P. \tag{6.16}$$

This is a well-known type of differential equation, so you may remember the standard solution. Otherwise, simply integrate $\mathrm{d}P/P$, which returns a natural logarithm of P which can then be inverted to yield P. However, this is only this straightforward if we can assume that none of the quantities μ, g, and T depend on z:

- μ: This means almost with certainty that the gas be uniform in its composition. This is not strictly true for the Earth's atmosphere – for instance, ozone is found mainly close to the surface due to smog (where it acts as a toxic to our health) and high up in the stratosphere (where it protects us from ultraviolet radiation).

- g: Above the surface of a planet, the gravitational acceleration declines with distance r to the centre of the planet as $g \propto r^{-2}$, so it definitely changes with altitude z. But the bulk of the atmosphere is in a very thin layer close to the surface, compared to the planet's radius, and the approximation $g \neq g(z)$ is generally a good one.

- T: This means that the atmosphere be isothermal. This is generally not true. In the Earth's atmosphere the temperature declines with altitude in the troposphere, up to $z \sim 10$ km, because it is heated from the bottom. Then, it increases in the stratosphere up to $z \sim 50$ km, because there the ultraviolet radiation from the Sun helps convert oxygen into ozone, which in turn absorbs more ultraviolet photons: this heats the gas.

If we do make these assumptions, then:

$$P \;=\; P(0)\; \exp\left(-\,\frac{\mu g}{kT}\, z\right) \;\equiv\; P(0)\; \exp\left(-\,\frac{z}{H}\right), \tag{6.17}$$

where the scaleheight H is defined as:

$$H \;\equiv\; \frac{kT}{\mu g}. \tag{6.18}$$

The scaleheight is a measure for how quickly the pressure drops with altitude. In the lower Earth's atmosphere, $H \sim 8$ km. Note that rearranging Equation 6.18 yields $\mu g H = kT$: the work done against gravity to ascend by a scaleheight equals the thermal (kinetic) energy. This energy balance arises from the conservation of energy in a closed system.

6.4.2 Evaporation

In reality, the atmosphere of a planet is not static. It continuously evaporates into space and, if not replenished at the same rate, it will eventually disappear. The reason for this is that in a gas, there is always a chance for an atom to gain a velocity that exceeds the velocity necessary to escape from the planet's gravitational potential – the escape speed, which is obtained from equating the kinetic energy with the potential energy gained by travelling from a distance r from the centre of the planet to infinity:

$$v_{\rm esc} \;=\; \left(2\frac{GM_\circ}{r}\right)^{\frac{1}{2}}, \tag{6.19}$$

where $G = 6.67 \times 10^{-11}$ J m kg^{-2} is Newton's gravitational constant, and $M_\circ$ is the mass of the planet. If the atom travels in a direction away from the planet at such high speed then it will indeed be lost.

The most likely velocity of an atom in a gas is determined by equipartition of energy in the frequent collisions of the atoms. This is characterised by the temperature, setting the thermal energy equal to the mean kinetic energy:

$$kT \;=\; \frac{1}{2}mv^2_{\rm likely}, \tag{6.20}$$

where m is the mass of the atom. Note that in a mixture of different atoms and molecules, m is not simply replaced by μ as the equipartition of energy would result in different velocities for the lighter and heavier species.

Not all atoms will have this speed. Rather, there is a distribution of velocities, $N(v)$, which must reach $N(0) = 0$ and $N(v_\infty) = 0$ and peak at $v = v_{\rm likely}$. In thermo-dynamic equilibrium, when collisions are frequent enough, and in the non-relativistic limit (both which are valid for the atmospheres of planets that are even remotely habitable), this is the Maxwell–Boltzmann distribution:

$$N(v)\,{\rm d}v \;=\; 4\pi N\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}}\; v^2\; \exp\left(-\;\frac{m}{2kT}\;v^2\right)\;{\rm d}v, \tag{6.21}$$

where N is the number of atoms. Note that there is always a tail with $v > v_{\rm esc}$ no matter how high $v_{\rm esc}$ or low $v_{\rm likely}$. Atoms in that tail could leave the atmosphere (before they collide with another atom on their path) and thus deplete the Maxwell–Boltzmann distribution at those velocities. Meanwhile, collisions between the remaining atoms will lead to re-establishing the Maxwell–Boltzmann distribution, populating again the tail with $v > v_{\rm esc}$. So the evaporation of a planet's atmosphere is an ever-continuing process, and after some time the lightest constituents may have disappeared from it altogether.

Chapter 7

Life elsewhere in the Solar System

7.1 Where it's definitely not

7.1.1 Mercury

The innermost planet of the Solar System, Mercury is inhospitable:

- ⋆ It is close to the Sun (~ 0.4 au, where the astronomical unit is the average distance between the Earth and the Sun, 150 Mkm), and hence it is hot. This is exacerbated by its low albedo.
- ⋆ Tidal interaction with the Sun has led to near-synchronity between its spin (taking 59 Earth days) and its orbit around the Sun (taking 88 Earth days). As a result, one side of Mercury faces the Sun for an extended period of time and becomes very hot (≈ 420 °C), whilst the side facing away becomes extremely cold (≈ -220 °C).
- ⋆ Mercury's orbit is very eccentric, and its distance to the Sun varies correspondingly, between 0.30–0.47 au. This leads to significant temperature variations.
- ⋆ Mercury is small and, despite its relatively high density (66% is iron!), its gravity is only 40% that of Earth. This, combined with the high temperature and thus rapid evaporation, is the reason for the almost complete lack of an atmosphere.
- ⋆ Liquid water can therefore not exist on Mercury. (But water *ice* might exist in the eternal shade on the bottom of craters at the poles.)

7.1.2 The Moon

Earth's natural satellite, the Moon is also inhospitable. The single most important reason being that its gravity is only a sixth that of Earth. Hence it cannot retain an atmosphere.

7.2 Venus – our close twin

Venus might be considered inhospitable as well:

- ⋆ At 0.7 au, it is closer to the Sun than Earth, so it is definitely hotter than on Earth.

- ⋆ It is of similar size and density as Earth, and therefore the gravity is sufficient to retain an atmosphere. In fact, the atmosphere is extremely dense, enveloping the entire planet in a permanent thick cloud layer. Composed mainly of CO_2 (97%), and N_2 (3%), the Greenhouse effect is dramatic. The temperature at its surface is a scorching 467 °C, uniformly due to the strong atmospheric circulation patterns. Despite the enormous pressure of 92 bar[1], liquid water is not present at such high temperatures (see Figure 6.1). But lead would melt, and CO_2 is super-critical!
- ⋆ Venus's magnetic field is rather weak, and therefore provides little protection against the wind of fast particles emanating from the Sun, and against even more energetic particles that enter the Solar System (called cosmic rays).

However, in the upper troposphere of Venus's atmosphere, at an altitude of ∼50–65 km, the temperature and pressure have dropped into a range $T = 0$–100 °C around $P \sim 0.2$–1 bar, quite pleasant for life! This is roughly where the top of the cloud layer is situated. Organisms there must like sulphuric acid, UV light, and constant atmospheric motions.

Venus might have been habitable until ∼ 700 Myr ago, before it had such dense atmosphere. In fact, Earth has enough carbonates and nitrogen within its minerals for an atmosphere exactly like Venus's. There is no trace of ancient life or liquid water, because Venus is resurfaced regularly by global lava eruptions as heat built up underneath for lack of tectonics (lubricated by water) – that would otherwise have released heat more gradually.

Venus must have had some turbulent youth, as it spins opposite to its orbital motion – the only planet in the Solar system to do so! While the spin of the planet's body is very slow (243 Earth days) its atmosphere spins in just 4 Earth days! This is unexplained.

7.3 Mars – our little brother

7.3.1 Canals?

People have imagined civilizations living on the Sun (the dark Sunspots taken for windows in an otherwise luminous cover), and on a hypothetical planet Vulcan close to the Sun. But the most enduring have been suggestions of intelligent life on Mars. This received great impetus due to unfortunate and woefully optimistic interpretations of dubious data.

In 1888 the Italian astronomer Giovanni Schiaparelli sketched maps of Mars, as seen by eye (photographic plates were too insensitive). He noticed dark streaks running across the surface, which he referred to by the Italian word *canali*. The wealthy American entrepreneur and amateur astronomer, Percival Lowell took this as meaning "canals" – as in man-made water-filled trenches. His drawings of Mars exaggerated these features, and he added small dark patches which he interpreted as oases in an otherwise dry world. Eugenios Antoniadi first believed to be seeing canals too, but when he got access to the then most powerful telescope he realised that what had seemed like coherent features were in actual fact mere chance coincidences of mottled and ill-defined surface patterns[2].

[1] 1 bar = 100 kPa = 0.987 atm, with "atm" the standard atmospheric pressure at Earth's surface.

[2] Despite Lowell's persistent perpetuation of a myth, he left a huge positive legacy in building some of the most powerful telescopes in his day.

7.3.2 Water?

That said, the conditions on Mars do not preclude the presence of liquid water:

- ⋆ At 1.5 au from the Sun, it is cold but sometimes, somewhere temperatures reach $T \sim 0$–10 °C. Note also that salty water can remain liquid well below 0 °C.
- ⋆ The gravity is less than that on Earth, but a colder atmosphere evaporates less quickly. Indeed, Mars maintains a thin atmosphere, with 95% CO_2 and a few % N_2 very similar in composition to that of Venus. The pressure near the surface is ~ 6 mbar *on average*. This is very near the triple point of water! (Fig. 6.1)

There do in fact exist various pieces of indirect evidence for liquid water to be present:

- ⋆ Spaceprobes that have been sent to orbit or land on Mars have taken pictures of sand (a product of erosion by water), fluvial erosion patterns along the slopes of hills or the ridges of plateaus, flat basins which suggest alluvial sediment deposits, and meandering narrow valleys and canyons reminiscent of dry river beds.
- ⋆ Mars's rotation axis is inclined by about as much as that of Earth, causing seasonal variations in temperature but allowing permanent ice caps at the poles. This is mainly CO_2: it is common in the atmosphere and at the poles freezes out of the atmosphere. But the reflective properties of certain patches of ice suggest some of it to be water. The seasonal variations drive atmospheric meridional circulation by which ice that sublimates at the pole experiencing Summer is transported in vapour form to, and condenses as ice at, the pole experiencing Winter. The rotation axis of Mars wobbles on timescales of Myrs, with the inclination angle varying by some $\pm 10°$ (it has no large Moon to stabilise this). Hence the positions of the poles have varied in latitude, and water ice could have been present nearer the current equator.
- ⋆ Simple calculations suggest the conditions are favourable for sub-surface liquid water to exist at some depths and latitudes on Mars. These places can be located in a pressure–temperature diagram by comparison with the phase diagram of water. The sub-surface temperature and pressure depend on the depth z as:

$$T(z) = T_{\text{surface}} + \frac{q}{k}\, z, \tag{7.1}$$

$$P(z) = P_{\text{surface}} + \rho g\, z, \tag{7.2}$$

  where q is the thermal heat flow ($q_{\text{Mars}} \approx 0.03$ W m^{-2}), k is the thermal conductivity of the material ('regolith', with $k \sim 0.02$–0.3 W K^{-1} m^{-1}), ρ is the density of the regolith ($\rho \approx 2$ g cm^{-3}) and g is the surface gravity ($g_{\text{Mars}} = 0.38 g_{\text{Earth}}$).
- ⋆ Ground-penetrating radar, and γ-ray and neutron spectroscopic mapping have indeed sensed the presence of sub-surface water not just near the poles, but even near the equator, equivalent to a water column of ≈ 14 cm in the top 1 m.
- ⋆ The sub-surface water could be a permafrost. One of the robotic landers has scraped the top-soil layer, exposing what seemed like frost (which on subsequent Martian days sublimated as it was exposed to the low pressure and Sunlight).

- ⋆ Near the equator, fields have been discovered that look like broken ice floes, covered by a layer of dust. The "geo"logical age for these fields is only a couple of Myr.
- ⋆ Infrared mapping by Mars orbiters has revealed clay layers and large fields rich in carbonates. These are believed to be due to deposits from liquid water into sediments. The Curiosity rover found hydrated soil, corroborating this scenario.
- ⋆ Gullies have been found emanating from near the top-ridge of ravines. Elsewhere, swaths of different soil along steep slopes resemble landslides. In February 2008, a landslide was caught on camera as it happened. These observations have now been interpreted as evidence for the presence of sub-surface liquid water. Increased pressure, for instance by seasonal heating of the surface by Sunlight, could have molten sub-surface water ice and make it burst through the walls of ravines, releasing steam and dust and turning some of the soil into mud which slides down the slope.

7.3.3 Maintaining the CO_2 atmosphere

A problem with Mars's CO_2-rich atmosphere is that it would have disappeared through dissociation and subsequent escape, as well as freeze-out, unless there is a source which replenishes the atmosphere with CO_2. On Earth, this is tectonically-powered volcanism. Although huge shield volcanoes are known near Mars's equator, the highest of which, Olympus Mons towers 27 km, Mars is not tectonically active and may never have been.

One scenario to form CO_2 is based on the absence of ozone in the Martian atmosphere. Hence, all ultraviolet Sunlight penetrates all the way to the surface. This leads to the photo-dissociation of CO_2 (and water) *ice*. The oxygen which is produced in this process oxidizes CO vapour, to form CO_2. Some of the oxygen reacts with rocks, to produce the rusty red colour (iron oxides) for which Mars is known. The ultraviolet light also generates hydrogen peroxide (H_2O_2), a bleaching agent which is harmful to organic matter.

7.3.4 Evidence for life

There have been three lines of evidence that have supported the presence of life on Mars. However, all of these can also be explained in non-biological ways.

- ⋆ The Viking landers in 1975 performed experiments aimed at detecting metabolic activity. The outcomes of some of the experiments were negative, some inconclusive (positive results in sterile comparison samples), and some positive. But the positive results could also be explained by simple a-biotic photo-chemical reactions.
- ⋆ Methane gas (CH_4) has been detected in the equatorial regions of Mars. Methane is a common waste product of metabolism (e.g., in cows). Yet it could also be explained by "geo"-chemical processes, with methane released when sub-surface clathrate hydrates fail.
- ⋆ There was a near-certain false alarm when the discovery was claimed of fossilised organisms in a meteorite coming from Mars (see Section 7.5.2).

Still, all hope is not lost, yet:

- ⋆ Simple lifeforms might exist in sub-surface water reservoirs, protected from harmful ultraviolet light.
- ⋆ Life may have existed in the past, when the atmosphere was denser and permanent reservoirs of surface liquid water may have existed. Indeed, high concentrations of manganese suggest an erstwhile oxidizing environment; SO_2 from early volcanism may have contributed to a Greenhouse effect.
- ⋆ Interestingly, boron and molybdenum–oxide are more prevalent on Mars than they were on early-Earth. Both have been linked to the synthesis of DNA.

7.4 The Giants: Jupiter and Saturn

The largest planets of the Solar System, Jupiter and Saturn are inhospitable:

- ⋆ They are gaseous spheres without solid surface.
- ⋆ Like the Sun, their atmospheres are composed mainly of hydrogen ($> 90\%$ H_2) and helium (He), elements which have evaporated from the atmospheres of the small rocky planets we have discussed so far.
- ⋆ The clouds that are visible from far are located at a pressure around one bar. But because Jupiter and Saturn are so much further from the Sun, at 5.2 and 10 au, respectively, the temperatures are very low, $T < -100$ °C.
- ⋆ There are traces of methane, ammonia and water, but deep circulation throughout the atmosphere destroys any complex molecules that might have formed.

7.4.1 Europa

One might think Jupiter's moon Europa must be inhospitable:

- ⋆ Being at 5.2 au from the Sun, it is cold: $T \sim 80$–140 K.
- ⋆ Being small, the gravity is very low.
- ⋆ Hence no atmosphere is present: it would either have evaporated or frozen out.

But there is evidence for a sub-surface ocean of liquid water to exist:

- ⋆ The morphology of a large impact basin discovered on its ice-covered surface suggests that the impact cracked a crust lying on top of some more fluid material.
- ⋆ The morphology of certain terrain (called 'chaos terrain') closely matches that of ice floes which have been shattered and upheaved. This suggests an underlying ocean.
- ⋆ Titan has a magnetic field, which implies electrical currents. Induced by Jupiter's strong magnetic field, these electrical currents suggest a sub-surface liquid zone.
- ⋆ In 2016, the Hubble Space Telescope imaged Europa against the backdrop of Jupiter, detecting plumes in silhouette. These are suspected to be water geysers.

Sub-surface water would need a source of heating, for which two mechanisms are proposed:

- ⋆ Tidal heating: Jupiter exerts strong tidal forces on Europa, and its rotation is not entirely synchronous with its orbit, making the tidal waves move through the moon's body. Internal friction then releases heat.
- ⋆ Radio-activity: heat could be produced by the capture of photons and particles emitted when unstable isotopes decay.

It has been speculated that life might be able to survive in a Europan sub-surface ocean.

7.4.2 Ganymede & Callisto

The largest moon in the Solar system, Ganymede, and Callisto – both moons of Jupiter – also have sub-surface oceans. However, these are deeper and colder, and lain on top of dense ice. The lack of a rock–liquid interface and associated chemical reactions may render the conditions for life less favourable.

7.4.3 Titan – the moon that ought to be a planet

The next-largest moon of the Solar System, Titan[3] orbits Saturn. There is no chance of liquid water, as it is much too cold. But replacing water, methane here performs many of the functions water provides in more temperate places:

- ⋆ Being 10 au from the Sun, the temperature at the surface is $T = 94$ K.
- ⋆ Being a fairly large "planetary" body, the gravity of Titan is strong enough to maintain a uniquely (for a moon) dense atmosphere: $P = 1.5$ bar at the surface.
- ⋆ The atmosphere is composed mainly of N_2 (95%) and CH_4 (methane, 5%) instead of CO_2 (which is frozen out – see the phase diagram in Figure 7.1).
- ⋆ The temperature and pressure are very near that of the triple point of methane: $(T = 90.7 \text{ K},\ P = 1.6 \text{ bar})_{\text{triple}}$.

Nitrogen would not be expected to have survived in the atmosphere unless it were somehow replenished. Two scenarios have been suggested:

- ⋆ A primordial origin, being trapped in ice cages ('clathrate') in the formation of Titan, subsequently sublimating into the atmosphere.
- ⋆ Through photo-dissociation of ammonia (NH_3): this yields N_2 and H_2 where the latter escapes from the atmosphere through evaporation.

The concentration of nitrogen relative to that of the inert gas argon could decide between these two scenarios:

[3]Titan has a mean radius of 2576 km, cf. 2634 km for Ganymede. Titan appears larger, though, because of its dense atmosphere which prevents us from seeing all the way to the surface.

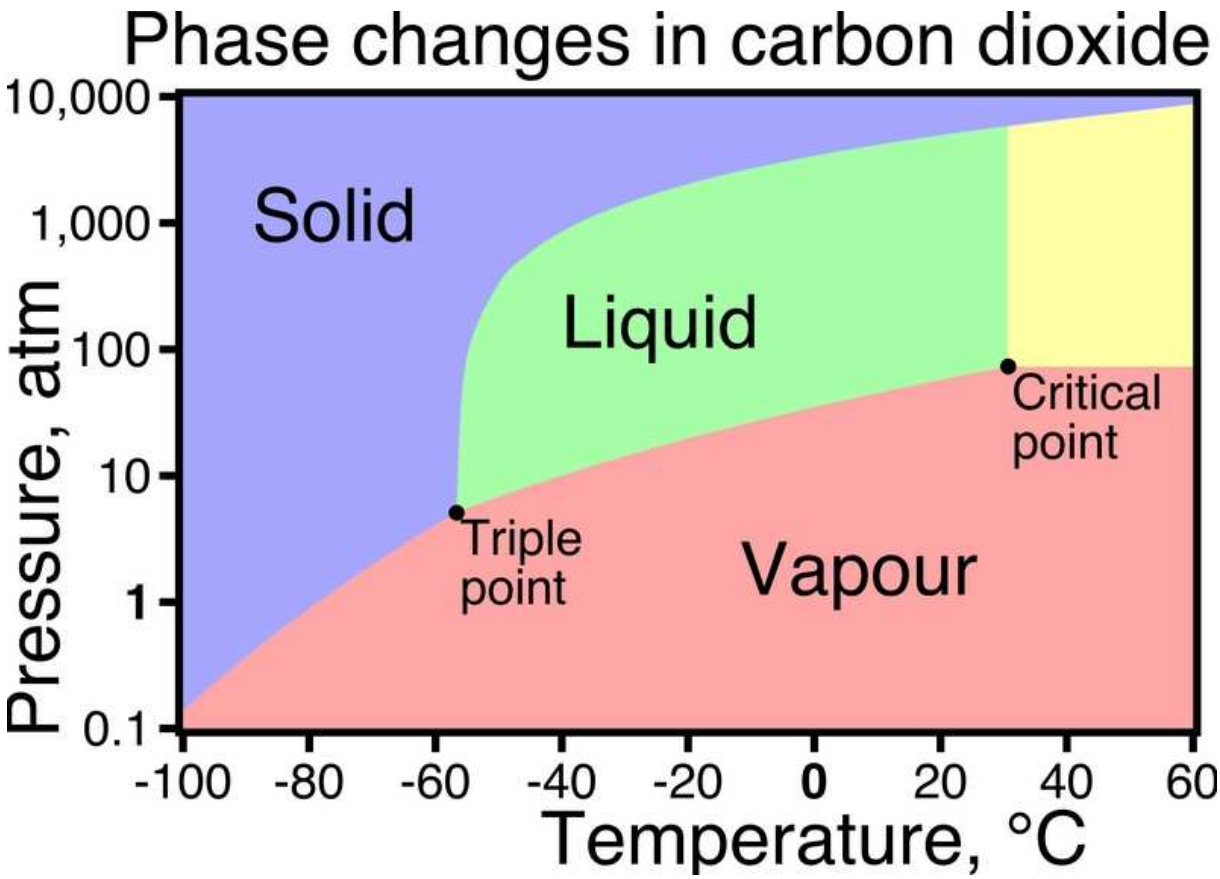

Figure 7.1: Phase diagram of carbon–dioxide (CO_2).

- ⋆ If primordial, the N:Ar ratio is set by that of the ambient interstellar medium from which stars and planets form.
- ⋆ If nitrogen were produced from ammonia, the nitrogen and argon would have led independent lives throughout the formation and evolution of Titan. This would have resulted in a very different N:Ar ratio.

The Huygens lander carried by the Cassini spacecraft measured the N:Ar ratio, which confirmed that ammonia is the primary source of Titan's atmospheric nitrogen.

Likewise, the methane in Titan's atmosphere must be continuously replenished. The source for this is believed to be a sub-surface methane–water ocean (on top of dense ice); flexure of Titan's surface occassionally leads to the release of methane into the atmosphere.

Until the arrival of Cassini/Huygens, Titan's thick cloud cover prevented a clear view of its surface. The Cassini spacecraft imaged infrared light, which in the absence of water and carbon–dioxide can penetrate quite well through the nitrogen–methane atmosphere. Its radar sensed, through differences in reflective properties, the presence of "pools", or "lakes", probably of methane or ethane, although it could not be said whether this was liquid or solid. The best such pictures show "rivers" leading to large basins.

Titan's atmosphere has a lower troposphere throughout which the temperature diminishes with growing altitude. Clouds of methane and ethane form, and it is quite possible that

it occassionally "rains", with the methane/ethane precipitating onto the surface, being collected and flowing through temporal rivers ("wadis") into seasonal lakes.

The Huygens lander sent pictures taken during its descent, showing dramatic views of hilly terrains of ice (water ice at those low temperatures behaves as rock) and "sediment" probably of the methane/ethane kind. It survived the landing, and sent us a postcard of the Titan landscape: pebbles of ice (likely water and hydrocarbons[4]), scattered on a soft surface – a methane "beach".

There was no sign of life, but microbial life cannot be excluded. However, hydrocarbons are poor solvents of large molecules – already not helped by the low temperature. And the lack of oxygen forces Titanic life to do without sugars, amino acids and nucleic acids.

7.4.4 Enceladus

Another of Saturn's moons, Enceladus is a mere 500 km in diameter but amazingly it too sports a body of sub-surface liquid water. The stresses caused by the varying gravitational pull in its elliptical orbit – kept that way by the moon Dione – heat the upper layers of the icy interior and power geysers that spew a spray of ice particles, methane, ammonia, organic matter and salt. Crucially, the ocean lies on a bed of rock, providing a surface for interesting physical and chemical reactions that could have spawn life. Would this world be the lively one, besides Earth?

7.5 Minor bodies

Even the free-floating sub-planetary bodies in our Solar System may be relevant to life.

7.5.1 Dwarf planets, asteroids and meteorites

Asteroids are rocks generally found between the orbits of Mars and Jupiter. The largest, Ceres measures 950 km in diameter and is in fact classified as a 'dwarf planet'. Evidence of clay, salt and water ice on its surface suggests there may have been liquid water. Ceres is believed to harbour a sub-surface ocean sandwiched in between a rocky core and a thick ice shell with a brittle crust. It is composed of 40% water (Earth has 0.01%!). The Dawn spacecraft, having visited the asteroid Vesta in 2011/12, arrived at Ceres in 2015. It observed haze above ice in the Occator crater, and signs of cryo-volcanism.

The smallest pieces are prone to depart from the main asteroid belt. If the Earth crosses their path they enter the atmosphere, are heated and some of the trapped gases are ionised – lighting up as a meteor. Any surviving bits landing on Earth are called meteorites.

One famous example is the Murchison meteorite:

- ⋆ Tiny, μm-sized particles of graphite and silicon–carbide (SiC) were discovered inside this meteorite. These grains date from before the formation of the Sun; they were formed in the ejecta from certain types of cool giant stars or exploded massive stars.

[4]Hydrocarbons only contain carbon and hydrogen, whereas carbo-hydrates also contain oxygen.

- ⋆ The meteorite also contains water, which must have been ice when still in space.
- ⋆ Intriguingly, many types of amino acids were also found inside. Their concentrations are very similar to those produced in the Miller–Urey experiment (Section 5.1.2), which just goes to show that these simple organic molecules are easy to make and commonly found even where conditions may not be favourable to support life.

7.5.2 Meteorites from Mars

Impacts on Mars have ejected rocks from its surface with such high speeds that they were flung out into interplanetary space. By chance the Earth has intercepted some of these rocks, some of which have been found. Inclusions of gases in these meteorites have been analysed and their relative concentrations provide an exact match to those in the Martian atmosphere – and very different from Earth's.

One such Martian meteorite, ALH 84001 (found in 1984) caused quite a stir. In 1996, it was announced that inside, electron-microscopic images revealed the fossilised remains of worm-like microbes that must have lived on Mars. Although it has been impossible to refute entirely, it has also been impossible to confirm that the object really once was a living creature. Meanwhile, it has been demonstrated that very similar structures are formed, under certain conditions, in the crystallization process of minerals.

Interestingly – while less likely – terrestrial meteorites may reach Mars, or even Europa! The transfer of rock bearing biotic material from one planetary body to another could possibly fertilize or "contaminate" the latter – this mechanism is called 'lithopanspermia'.

7.5.3 Asteroids from Outer Space

In 2017, an asteroid was – for the first time – confirmed to have entered the Solar System from interstellar space. The object, 'Oumuamua, or less prosaïcally 1I/2017 U1, is highly elongated, a few hundred yards long, and tumbling. At any moment, there must be many smaller interstellar asteroids crossing our path. In 2019, the first interstellar comet, 2I/Borisov was discovered, measuring ~ 2 km across. Could these objects carry life?

7.5.4 Comets

Comets are icy, generally km-size bodies that originate from beyond the orbit of Neptune. When they come close to the Sun some of the volatile material sublimates, leaving a gas tail pointing away from the Sun (because of the radiation pressure by the photons upon the atoms) and a dust tail trailing the comet in its path (dust grains are much heavier than atoms and the acceleration due to radiation pressure is therefore less).

- ⋆ Comets contain a lot of water ice, and comets crash-landing on Earth – an especially frequent spectacle in the first half Gyr since the formation of the Solar System – may have supplied large amounts of water. The recent landing of the probe Philæ (carried by the Rosetta spacecraft) on comet Churyumov–Gerasimenko in November 2014 has cast doubt on this scenario: it found an isotopic composition of water (deuterium) on the comet higher than that in Earth's oceans – though this can be due to selective sublimation. Comet impacts may explain noble gas isotopic ratios.

- ⋆ The impact of comets upon Earth has been suggested to facilitate the formation of amino acids out of methane, via the formation first of HCN and C_2H_2 (toxic molecules, very commonly found in space) and then ammonia. However, it has since been argued that this would not work in the presence of CO_2 (other chemical reactions would take precedence), as in the early Earth's atmosphere.

- ⋆ The Stardust spacecraft took samples from comet Wild 2, and returned these to Earth where they were scrutinised in laboratories with sophisticated instruments. By heating the material and timing the release of volatiles, one can deduce the composition. Some amino acids were found. But, as said before, amino acids appear to be so common and easy to make that their presence is no longer seen as strong evidence for conditions suitable to life.

7.5.5 Outer solar system dwarf planets

Beyond Neptune's orbit starts the realm of the Kuiper Belt and scattered-disc objects – a large population of icy left-overs from the formation of the Solar System (most comets originate a thousand times farther afield, in the putative Oort cloud). Some of these are large enough to be considered dwarf planets. Among the most famous examples are:

- ⋆ Pluto. First considered a planet when it was discovered in 1930, it "became smaller" when it was found to have a large moon, Charon. It is now classed as a dwarf planet mainly on the basis of its composition and orbit. The strong gravitational tides within the compact Pluto–Charon system have resulted in orbital circularization and spin–orbit synchronization ("tidal locking") and hence these tides no longer cause internal heating as there is no more friction. However, radiogenic heating may keep a sub-surface water layer liquid beneath Pluto's nitrogen-ice surface.

  Pluto was visited by the New Horizons spacecraft in 2015. It measured its diameter to be 2375 km, and revealed a blue haze due to ærosols scattering Sunlight, young (relatively uncratered) terrain formed by the flow and convection ("bubbling up from the interior") of nitrogen "lava", and possible (dormant) volcanoes. Large areas on Charon (diameter 1211 km) are covered with ammonia.

- ⋆ Triton is similar to (but larger than) Pluto, but it was once captured by Neptune which it now orbits as one of its moons. Geysers were observed in action when the Voyager 2 spacecraft flew by, leaving organic material on the surface. Radiogenic heating is the likely power of these geysers and a sub-surface methane–water ocean.

- ⋆ The scattered-disc object Eris was only discovered in 2005, but it is as large as Pluto. Its discovery prompted the International Astronomical Union to revisit the definition of a planet, and in 2006 to demote Pluto (and Eris, and to promote Ceres) to the status of dwarf planet. Eris may well also have a sub-surface water ocean.

There are expected to be many more objects like the above, many of which are much more distant most of the time, and which may have sub-surface water oceans resulting from radiogenic heating (which does not require the presence of a companion or starlight). If these oceans provide habitats for life then the Universe might be teeming with it!

Chapter 8

Searching for exoplanets

Exoplanets may be detected via a variety of methods, for example:

- ⋆ Direct detection of the planet, separated from the star, via:
  - Starlight reflected by the planet.
  - Thermal emission of the planet itself.
- ⋆ Detection through occultation, either:
  - The planet transiting in front of the stellar disc.
  - The planet itself being occulted by the star.
- ⋆ Through variations in the timings of the transits of other planets.
- ⋆ Detection through the star's reflex motion, via:
  - Astrometric motion of the stellar position.
  - Doppler motion of the spectral lines in the spectrum of the star.
- ⋆ As a gravitational lens magnifying a background star (or its lensed image).
- ⋆ Through pulsar timing (if in orbit around the pulsar).

8.1 Direct detection

8.1.1 Reflected starlight

Stars are generally bright at optical wavelengths, and planets have generally a significant optical albedo. One may thus attempt to detect starlight reflected off a planet's surface. Like the phases of the Moon, the size of the illuminated surface of the planet varies:

- ⋆ The inclination angle of the planet's orbit, i is defined as the angle between the planet's orbital axis (which itself is perpendicular to the planet's orbital plane) and the line-of-sight (which is the line connecting the star and us). For the case of $i = 0°$, the planet would appear to orbit the star in the plane of the sky. No variations would be expected in the illuminated surface as seen from Earth.

- ⋆ For the case of $i = 90°$, the full disc of the planet would be seen illuminated if the planet is at the opposite side of the star (except when it is exactly behind the star in which case it would be occulted by the star). This alignment is called 'superior conjunction'. In the opposite case, when the planet is in front of the star no reflected light would be seen from Earth. This alignment is called 'inferior conjunction'. When the planet is seen at either side next to the star, half of the disc is seen to be illuminated – this alignment is called 'quadrature'.

It is thus obvious that the planet would be brightest if it is in superior conjunction. With the usual nomenclature, one can easily derive:

$$\frac{L_{\circ,\ \text{reflected}}}{L_\star} \;=\; \frac{A}{4}\,\left(\frac{R_\circ}{d}\right)^2, \tag{8.1}$$

where A is the *geometric* albedo, which differs from the Bond albedo as it depends on the angles of incidence and reflection (we are only interested in light reflected in our direction).

On the other hand, the planet's widest angular separation from the star occurs in quadrature, when one might attempt to take an image showing star and planet side-by-side. The brightness at quadrature is expected naïvely to be only half that at superior conjunction, but in reality the difference is larger because of the angular dependence of reflection.

Rather than resolving the star–planet system, the variation in phase of the illuminated part of the planet as seen from Earth might be large enough to be detected as a tiny modulation of the brightness of the star+planet system.

8.1.2 Thermal emission

The planet is heated by the starlight it intercepts but does not reflect, and the resulting planet's temperature is lower than that of the star. Although the thermal emission from the planet is very weak compared to the luminosity of the star, it is emitted at longer wavelengths than where the starlight peaks. Thus, by observing at longer wavelengths, the contrast between the light from the planet and that from the star improves.

It can be easily argued that the luminosity contrast is:

$$\frac{L_{\circ,\ \text{thermal}}}{L_\star} \;=\; \left(\frac{R_\circ}{R_\star}\right)^2\,\frac{B'_\lambda(T_\circ)}{B_\lambda(T_\star)}, \tag{8.2}$$

where the spectral distribution of the emitted light is B_λ (not necessarily an identical function of temperature for the star and for the planet). At wavelengths significantly longer than where both spectral distributions peak, B can usually be well-approximated by the Rayleigh–Jeans tail of the Planck curve, and:

$$\frac{L_{\circ,\ \text{thermal}}}{L_\star} \;=\; \left(\frac{R_\circ}{R_\star}\right)^2\,\frac{T_\circ}{T_\star}, \tag{8.3}$$

Note that the planet will never become brighter than the star, unless the planet is much larger than the star – for instance if the "star" is in fact a stellar remnant such as a white dwarf (young white dwarves are still hot).

8.1.3 Angular resolution

The ability to see the planet separately from the star is called the angular resolution.

The angular resolution α of an optical system is determined by the widest separation between parts of its light-collecting (primary) optical element: the maximum 'baseline', D. In the case of a monolithic lens or mirror this is simply its diameter. In the case of interferometers it is the distance between the optical elements that are furthest apart – this is a common technique at radio wavelengths where the signals from different radio dishes are more easily combined coherently, but optical interferometers also exist.

The principles of interference of waves (e.g., light) dictate that the angular resolution depends also on the wavelength λ, like so[1]:

$$\alpha \;=\; 1.220\;\frac{\lambda}{D}. \tag{8.4}$$

8.2 Detection through occultation

If the inclination of the orbit is sufficiently close to 90°, the planet will alternately transit in front of the star and be eclipsed by the star when it passes behind the star.

8.2.1 Planetary transits

Especially at optical wavelengths the planet's unilluminated surface can be considered black compared to the surface brightness of the star. Hence, when the planet is fully in front of the star, the brightness as seen from Earth is diminished by a fraction:

$$f \;=\; \left(\frac{R_\circ}{R_\star}\right)^2. \tag{8.5}$$

(in the unusual case $R_\circ > R_\star$ then $f = 1$) Of course, if the transit is grazing, i.e. the planet only partially passes in front of the star, the fraction is smaller than this.

The duration of the transit depends in a simple geometric way on the orbital inclination, orbital speed (in principle that of both planet and star – the latter moves in the opposite direction), and the stellar radius. The duration of 'ingress' – i.e. from when the planet starts transiting until its disc is fully in front of the star – depends on the inclination, speed, and the planet's radius (and identically so at 'egress' at the end of the transit).

Because only few planets will be seen to transit, finding planets by their transits requires monitoring the brightness of many stars, frequently enough not to miss the transit event. Keele University has built its own transiting exoplanet observatory, called SuperWASP-South, located in South Africa (SuperWASP-North is on the Canarian island of La Palma). It led the world in detecting transiting exoplanets until the Kepler satellite increased the number of known transiting exoplanets by more than an order of magnitude.

[1]The factor 1.220 arises from the location of the first zero in the Bessel function of the first kind, of order one – which describes the diffraction pattern – divided by π.

8.2.2 Planetary eclipses

When the planet is eclipsed by the star, the combined observed brightness of the star+planet system is diminished (compared to that just before and after the eclipse) by a fraction:

$$f \;=\; \frac{L_\circ}{L_\star + L_\circ}, \tag{8.6}$$

which is generally well-approximated by:

$$f \;\simeq\; \frac{L_\circ}{L_\star}. \tag{8.7}$$

The durations of the eclipse, ingress and egress are identical to those of the transit. Again, we have assumed that $R_\circ < R_\star$ (in the opposite case, the fraction f is smaller).

8.2.3 Transit Timing Variations

Additional planets cause variations in the timing of the transit because their gravitational pull affects the transiting planet's motion. The effect is small but has been measured in some cases – and planet Neptune was discovered in this way in 1846 by Johann Galle following predictions by Alexis Bouvard and Urbain Le Verrier based on perturbations of the orbit of Uranus (which had been discovered by William Herschel in 1781, even though it is visible to the unaided eye). Transit timing variations not only indicate the presence of additional planets but also offer a way to obtain the planets' masses.

8.3 Detection through the star's reflex motion

The planet and star move around the common centre-of-mass. The presence of a planet can therefore be inferred from an observed motion of the star, either via astrometric (positional) or radial velocity measurements. In both cases the astrophysical problem is described by the prescription for the location of the centre-of-mass:

$$M_\circ \, a_\circ \;=\; M_\star \, a_\star, \tag{8.8}$$

where a is the distance between the object with mass M and the centre-of-mass. For the sake of mathematical simplicity we assume circular orbits – otherwise, the centre-of-mass is located in the common focus of the elliptical orbits.

To measure the planet mass we need to know the stellar mass, and both $a_\circ$ and $a_\star$. The stellar mass can often be fairly accurately estimated, most commonly from the following:

- ⋆ The luminosity of a star is related to its mass, because more massive stars have a higher core pressure and temperature and consequently a higher rate of nuclear fusion reactions going on. This relationship is most accurately known for hydrogen-burning stars, on the 'main sequence' where they spend most of their lives and where conditions for habitable planets are generally most favourable. To know the luminosity of a star one will have to know its distance.

- ⋆ The temperature of a star is also related to the mass of a main-sequence star, because a more massive star requires a hotter surface to be able to radiate away the energy at the same rate as it is produced in its core – if it doesn't, then the star will either expand or contract until such radiative equilibrium is reached. The temperature can be estimated from the colour of the starlight or (preferably) from the relative strengths of absorption lines in its spectrum.

The laws of gravity and motion prescribe how $a_\circ$ is related to the stellar mass (we assume for simplicity that $M_\circ \ll M_\star$), as the acceleration perpendicular to orbital motion is:

$$\frac{\mathrm{d}v_\perp}{\mathrm{d}t} \equiv -\frac{v^2}{a_\circ} = -\frac{GM_\star}{a_\circ^2}, \tag{8.9}$$

where $v_\circ \equiv v_\parallel$ is the orbital speed of the planet. Hence:

$$v_\circ = \left(\frac{GM_\star}{a_\circ}\right)^{\frac{1}{2}}. \tag{8.10}$$

The orbital speed can also be expressed in terms of the orbital period P (the "year"):

$$v_\circ = \frac{2\pi a_\circ}{P}. \tag{8.11}$$

In both methods (astrometric or radial velocity measurements) the period is obtained simply from the cyclic variation that is measured. Hence we obtain Kepler's third law:

$$a_\circ = \left(\frac{P}{2\pi}\right)^{\frac{2}{3}} (GM_\star)^{\frac{1}{3}}. \tag{8.12}$$

What remains to be obtained from the astrometric or radial velocity measurements is $a_\star$.

8.3.1 Astrometry

If repeated measurements of the position of a star show a cyclic variation different from that due to Earth's motion around the Sun (the parallax, a useful method for determining the distance to a star), then a companion may be suspected.

The advantage of this method is that a measurement of $a_\star$ is obtained which does not depend on the orbital inclination: for a different inclination the projection of the orbit on the sky is more circular or more linear, but the full amplitude of the wobble remains $2a_\star$. However, in case of an elliptical orbit it does depend on the orientation of the orbit with respect to the line-of-sight.

The drawback is that the wobble is tiny, and harder to measure the further the system.

8.3.2 Doppler shifts

The wavelength of waves emitted by a moving source is modified from that at rest. In the non-relativistic case, this Doppler shift is obtained from simple geometric considerations:

$$\frac{\lambda_{\rm observed} - \lambda_{\rm rest}}{\lambda_{\rm rest}} \;=\; \frac{v_{\rm rad}}{c}, \tag{8.13}$$

where $c = 3 \times 10^8$ m s^{-1} is the speed of light, and $v_{\rm rad}$ is the component of the motion away from the observer (radial velocity; $v_{\rm rad} < 0$ if the source is moving towards the observer).

This shift is measured from the positions of narrow absorption lines in the spectrum due to transitions between electronic levels in an atom. Orbital motion results in cyclic variation in the Doppler shift as the star moves towards and away from the observer depending on where it is in its orbit. To detect minute stellar wobbles, a relativistic version of Equation 8.13 needs to be used where $v_{\rm rad}$ includes the varying motion of the observer.

The value for $a_\star$ could, in principle, be determined from the star's orbital speed:

$$v_\star \;=\; \frac{2\pi a_\star}{P}. \tag{8.14}$$

In practice, the orbital inclination results in a smaller measured velocity amplitude. The Doppler technique yields a value for $v_\star \sin i$, and hence $a_\star \sin i$. As a consequence, only $M_\circ \sin i$ is obtained – which means that the true planet mass $M_\circ \geq M_\circ \sin i$.

Accuracies of 1 m s^{-1} are achieved, equivalent to walking speed! The first planet around a normal star was discovered in 1995 by Michel Mayor and Didier Queloz, by applying this technique to the star 51 Pegasi. The planet has recently been named Dimidium.

8.4 Gravitational microlensing

Gravity is the distortion in the space–time reality caused by gravitational mass (which, somehow, is identical to the inertial mass). The propagation of waves is affected by this. As a result, light rays are bent when they pass close to an object. When they arrive at Earth, they seem to be coming from slightly different directions than from the position of the source. This was also the method used by Arthur Eddington during the 1919 solar eclipse to confirm Albert Einstein's general theory of relativity – using the astrograph that was carried by Keele Observatory's 31-cm refractor when it was still in Oxford!

Depending on the alignment of source, object and us, distorted images of the light source are formed. These images are usually amplified in intensity as more light is concentrated than would otherwise have reached Earth. This effect is called 'gravitational lensing'. When the object is of stellar or lower mass we speak of 'microlensing'.

When a star passes in front of a background star, the background star may thus be brightened (it is not currently possible to resolve micro-lensed images) and subsequently dim to its original brightness as the lensing star moves out of the way (see Figure 8.1).

If a planet is present near the lensing star, it too can lens the background star. It can do this directly, if the planet comes close to the line-of-sight towards the background star. But it can do this also indirectly, when the planet comes close to the direction in which the lensed image is seen (see Figure 8.1). The latter is much more likely because the lensed image is spread out more and moves around with a reasonable likelihood of sweeping past

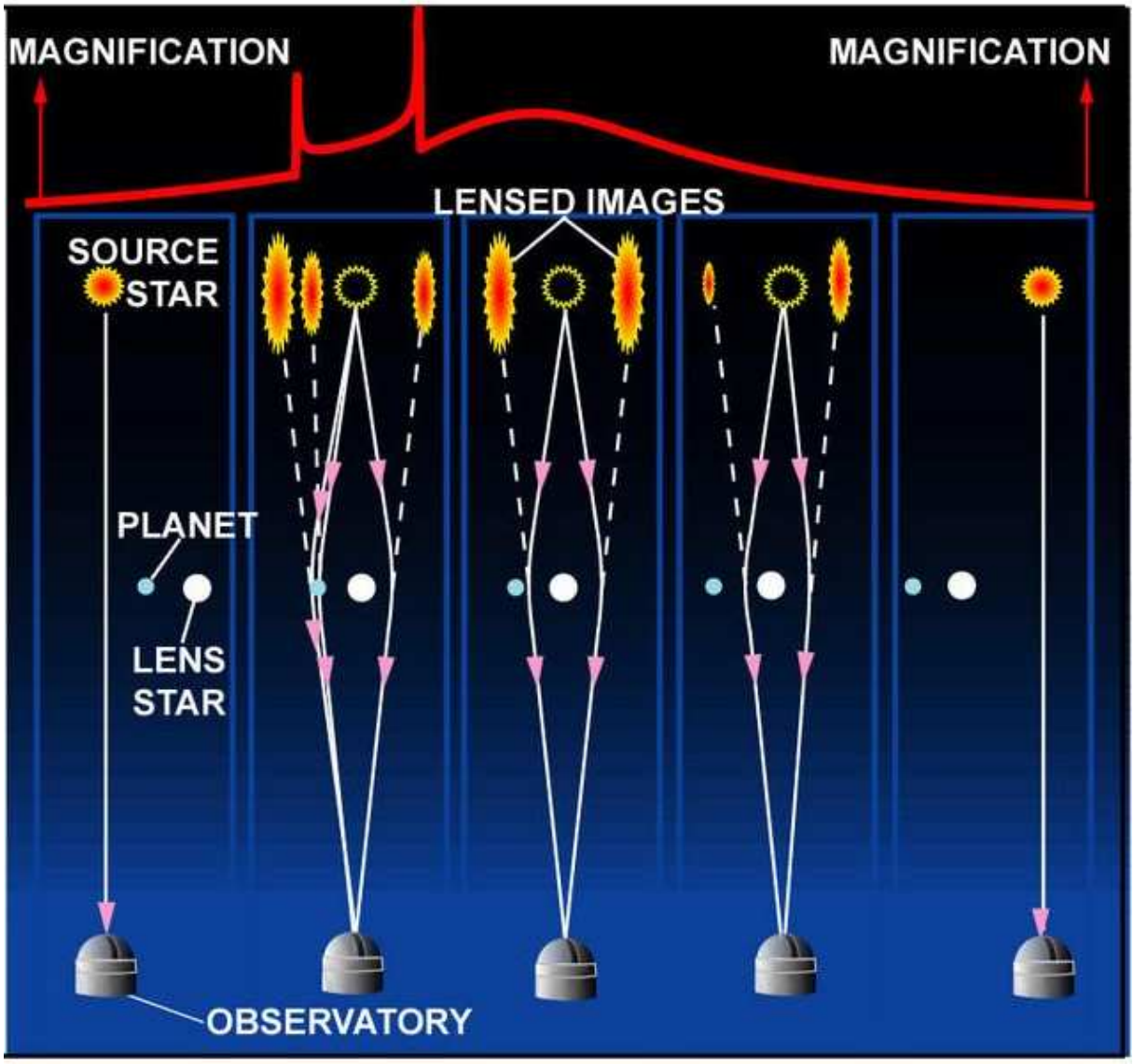

Figure 8.1: Schematic explanation of the effects during a microlensing event due to a star and due to a planet around that star. What is measured is the variation in time of the brightness of the background star, shown in red at the top of the diagram.

the planet at some point. The planetary lensing event can add a complicated signature to the stellar lensing event, although in the simple case it would be a sharp peak in brightness on top of the more gradual brightness variation due to the stellar lens.

Two things interest us here: the likelihood that lensing will happen, and the amplification of the background star's brightness. The likelihood of it happening depends on how close the lens and background source are seen in the sky, and on the mass of the lens.

8.4.1 Einstein radius

Significant lensing occurs when the angular separation between lens and background source comes within approximately the Einstein radius β (in radians):

$$\beta \;=\; \left(\frac{2R_{\rm s}}{d_{\rm lens}}\;\frac{(d_{\rm source}-d_{\rm lens})}{d_{\rm source}}\right)^{\frac{1}{2}}, \tag{8.15}$$

where $d_{\rm lens}$ and $d_{\rm source}$ are the distances from Earth to the lensing star and the lensed background star, respectively. The Schwarzschild radius $R_{\rm s}$ is defined as:

$$R_{\rm s} \;=\; \frac{2GM_{\rm lens}}{c^2}, \tag{8.16}$$

which is derived simply from equating the escape velocity to the speed of light. If the background source is much further away than the lensing star, then the expression for the Einstein radius simplifies to:

$$\beta \;\simeq\; \left(\frac{2R_{\rm s}}{d_{\rm lens}}\right)^{\frac{1}{2}} \;\propto\; (M_{\rm lens})^{\frac{1}{2}}\,. \tag{8.17}$$

8.4.2 Amplification

The amplification by the lens, μ, is not easy to derive. We simply list here the result:

$$\mu \;=\; \frac{\left(\frac{\alpha}{\beta}\right)^2+2}{\left(\frac{\alpha}{\beta}\right)\times\left[\left(\frac{\alpha}{\beta}\right)^2+4\right]^{\frac{1}{2}}}, \tag{8.18}$$

where α is the angular separation between lens and source. Significant amplification occurs for $\alpha \ll \beta$, in which case a simplified expression is obtained:

$$\mu \;\simeq\; \frac{\beta}{\alpha}. \tag{8.19}$$

8.5 Pulsar timing

Remarkably, the first exoplanets were discovered not around a normal star like the Sun, but near an exotic remnant of a massive star. In 1992, Alexander Wolszczan and Dale Frail announced the discovery with the world's largest radio telescope, in Arecibo (Puerto Rico), of two ~ 3 $M_\oplus$ planets orbiting the 'pulsar' PSR 1257+12.

Pulsars are rapidly rotating neutron stars, ~ 1.4 $M_\odot$ of neutrons packed together within a ball of ~ 10 km in diameter (an atomic nucleus on steroids!). When they are still young, neutron stars have strong magnetic fields, which accelerate particles. This "radiation" is highly beamed (emitted in a specific direction), and is detected only when one of the magnetic poles is facing us. The magnetic axis is generally inclined with respect to the rotation axis (as is Earth's), and the spinning of the neutron star thus causes the radiation to be pulsed. At a period of 6.2 milli-seconds, PSR 1257+12 is one of the fastest.

Like in the Doppler effect, the pulsation period varies as the neutron star moves towards and away from us due to the gravitational pull by the planet(s).

Chapter 9

Properties of discovered exoplanets

By early 2020, there were a total of > 4000 confirmed exoplanets, up from only 30 before the year 2000. It has thus become possible to take a statistically meaningful look at the properties of the discovered exoplanets[1]. The vast majority of exoplanets are found through the transit method (≈ 3000) or Doppler method (≈ 900); the Kepler satellite found ~ 2500 more exoplanet candidates but most of their host stars are too faint to confirm the planet masses with Doppler measurements.

9.1 Planet masses

Useful units to express exoplanet masses are the mass of Jupiter, the largest planet in our Solar System, and the mass of Earth: $M_{\rm Jupiter} = 0.001\ {\rm M}_\odot = 318\ {\rm M}_\oplus$.

Planets more massive than $\sim 13\ {\rm M}_{\rm Jupiter}$ are expected to develop such high pressures and temperatures in their cores that they sustain nuclear fusion of deuterium (hydrogen with an additional neutron) for at least some time since their formation. This would cause a significant increase in the energy output from the planet, and we therefore no longer call it a planet but a 'brown dwarf' star instead.

Most exoplanets that have been found have a mass similar to that of Jupiter, but an increasing number of planets with a mass around $10\ {\rm M}_\oplus$ are found; curiously, planets of about 40–50 $\rm M_\oplus$ (0.1–0.2 $\rm M_{Jupiter}$) are rare. The obvious observational reason for the predominance of Jupiter-mass planets is that massive planets are easier to detect:

- ⋆ More massive planets pull harder at the star, which therefore shows a larger radial velocity amplitude which is easier to detect in optical spectra of the star.
- ⋆ More massive planets are also generally larger, and therefore have the potential to occult a larger fraction of the stellar disc during a planetary transit, making it easier to detect the resulting dip in brightness of the star.
- ⋆ Larger planets are brighter; > 100 planets have been directly imaged using high-contrast cameras, or detected through reflected starlight in the lightcurves. While the largest planets are obviously the easiest to spot, it includes Kepler-70 b & c, two roughly Earth-sized planets in a very tight orbit around a hot, compact dwarf star.

[1] A good database is maintained at http://exoplanet.eu

9.2 Orbital radii

The orbital radii of the exoplanets discovered to date show a rather skewed distribution:

- ⋆ Many exoplanets are found at one or two au. This is not so surprising as we find planets in our Solar System at those locations.
- ⋆ The peak occurs for orbits much smaller than that of Mercury, the innermost planet in the Solar System. This was perhaps the biggest surprise emerging very soon after the first exoplanets were found: Not only is it quite common to find planets at < 0.1 au, but these are often massive planets as well. Remember: Jupiter is the massive planet in the Solar System closest to the Sun, at 5.2 au. The four planets in the Solar System closer to the Sun are all tiny compared to Jupiter.
- ⋆ There is a tail in the distribution out to very large distances from the star, much further than Neptune and Pluto are from the Sun.

The sensitivity of the observations affects the distribution over orbital radii:

- ⋆ Planets closer to a star are easier to detect because their pull on the star is stronger and they are more likely to transit (due to a larger margin in inclination as well as shorter orbital periods and hence more frequent transits). Hence nearly all super-Earths have been found within 1 au from the star, and the planets found around 1–5 au are often more massive than Jupiter.
- ⋆ Planets very far from a star may be detected via direct imaging – indeed this is how several of the planets at > 14 au were discovered. These are large (and thus massive), otherwise they would not have been bright enough to be detected.

The most pertinent robust results include:

- ⋆ Planets do exist very close to stars, and many of these planets are massive.
- ⋆ These 'Hot Jupiters' form a separate population from both the super-Earths and the outer gas giant planets.
- ⋆ Planets do exist with masses and orbital radii similar to planets in the Solar System.

9.3 Planetary systems

About a thousand of all known exoplanets are part of exoplanetary systems that have more than one planet. This allows us to look into the architecture of planetary systems.

With 8 planets, the Solar System remains the largest planetary system known. Besides the 8 planets, it also has at least 5 'dwarf planets' (e.g., Pluto). With masses < 0.01 $M_\oplus$, they are suspected to have formed in a slightly different way from the larger planets. Their detection is beyond the current capabilities of the optical radial-velocity and transit techniques. The richest exoplanetary system, Sun-like star Kepler-90 also has 8 planets; cool dwarf star TRAPPIST-1 has 7, as do naked-eye star HR 8832 and Sun-like star HD 10180. Systems with many more planets are probably unstable.

Four planets have been found to orbit HR 8799, through direct imaging. Not surprisingly these are all big, 7–10 $M_{Jupiter}$, and more distant from their host star than Saturn is from the Sun, 14–68 au. This planetary system is therefore rather unlike our Solar System.

The 5 planets in the 55 ρ^1 Cancri A system are also more massive than the planets in the Solar System at those same distances from the star – its lightest planet is 40× less massive than Jupiter but it orbits at just 0.015 au (in less than a day!). Like in the Solar System, the innermost planet is the least massive, and the planet outside 2 au is more massive than those within 2 au. 55 ρ^1 Cancri A also has a dwarf star companion, 55 ρ^1 Cancri B, at a distance of 1150 au (200× as far as the outermost planet of star A).

The masses of the planets around Kepler-90 have yet to be determined, while the planets around HD 10180 are not transiting. Both stars are slightly more massive than the Sun; the Kepler-90 planets all orbit within 1 au, but the HD 10180 system is more like ours – with an 0.2 $M_{Jupiter}$ gas giant planet orbiting at 3.4 au and smaller planets closer in.

Over the period 2015–2017, seven Earth-sized planets were found around the red dwarf star TRAPPIST-1 (just 40 light years away). Three of these may be in the habitable zone, though we do not know how hospitable the conditions are near a star of this kind.

9.4 Planet host stars

Planets are being discovered around stars which are selected *a priori*, as well as around stars which are observed as part of a "blind" survey. This induces both a spread in the properties of the planet host stars, such as the temperature of the star and its distance from Earth, as well as biases.

Some searches now target dwarf stars, which are lighter, smaller and dimmer than Sun-like stars, mainly because it can be easier to detect a planet around it than around a more massive, larger and brighter star (and the habitable zone is closer to a smaller star, too).

Most planet host stars are located at tens of parsec distance from the Sun, where a parsec (pc) is the distance at which 1 au subtends an angle of 1″ on the sky (1 pc = 3.26 light years = 206,265 au). The nearest stellar system, α Centauri and its companion Proxima Centauri are 1.3 pc away and each bear host to an Earth-mass planet; the next closest – and fastest moving star across the sky – Barnard's Star hosts a super-Earth. The centre of the Milky Way galaxy, on the other hand, is 8 kpc away. Obviously, nearer stars tend to be brighter and more amenable to accurate measurements.

There is a clear correlation observed between the frequency with which planets are found and the metal content of the planet host star: two in every three planet host stars have a metallicity above that of the Sun. This may not seem odd, but the vast majority of stars in the Milky Way galaxy contain *smaller* concentrations of metals than the Sun.

The metallicity is usually indicated by a logarithmic comparison with the Sun; often iron is taken as a proxy for all elements heavier than lithium:

$$[\mathrm{Fe/H}] \;=\; \log_{10}\left(\frac{N(\mathrm{Fe})/N(\mathrm{H})}{(N(\mathrm{Fe})/N(\mathrm{H}))_\odot}\right) \tag{9.1}$$

First believed by some to be a selection effect, metal-rich stars showing more spectral lines resulting in more accurate radial velocity measurements, it is now becoming widely accepted that metal-rich stars are indeed more frequently found to have planets. But the planets that *are* found around metal-poor and metal-rich stars do not appear to differ – except possibly for slightly smaller masses of the rocky planets at lower metallicity (but not of the gas giants, which are composed mostly of hydrogen and helium, not metals).

9.5 How complete is our picture?

The limitations of the current planet finding methods are generally well-understood. For instance, it is difficult to detect Earth-mass planets in Earth-like orbits around Sun-like stars. With better precision instruments, for instance from the more stable environment in space, this might eventually become possible. Planet-hosting stars continue to be monitored for radial velocity variations, which stand a good chance of revealing more planets around the same star and in particular planets in wider orbits.

About 10% of Sun-like stars that are being monitored for radial velocity variations are now known to have planets. The "ease" with which the Kepler satellite found candidate planets, suggests at least one in every three stars has planets. So the ultimate fraction of Sun-like stars with planets of any kind is possibly close to 100%.

On the other hand, blind surveys for microlensing events (which have detected nearly a hundred events so far) place a limit on the frequency of stars harbouring Jupiter-like planets, of $\lesssim 30\%$.

Likewise, while 'free-floating' planetary-mass objects (i.e. not orbiting a star) have been found in young star clusters (see also Section 13.2), microlensing surveys suggest these are scarce in the Milky Way as a whole.

It will be extremely hard to exclude a Mercury-sized planet to exist in a Jupiter-like orbit. Hence, in Drake's equation (1.1) the probability factor for Sun-like stars to host planets is close to unity – whether it is smaller for *Earth-like* planets has yet to be seen; some studies estimate one potentially habitable Earth-like planet for every three stars. In the next chapter we will look more carefully into empirical evidence of Earth-like planets in the Habitable Zone.

Chapter 10

Exoplanet atmospheres and exoplanets in Habitable Zones

What are the prospects for life on any of the exoplanets that have been found? To start addressing this question, we will look at what has been possible to find out about the atmospheres of exoplanets, whether some exoplanets are likely to be in the habitable zone, and how critical the dynamics in a planetary system are for the habitability of the planet.

10.1 Eroding atmospheres

10.1.1 Planet densities

For transiting planets not only the mass but also the radius can be measured. Together, these yield a value for the average density of the planet:

$$\langle \rho \rangle \;=\; \frac{M}{\frac{4}{3}\pi\ R^3}. \tag{10.1}$$

This has now been possible for about 800 exoplanets. We notice the following:

- ⋆ The distribution of transiting exoplanets over density appears log-normal, extending between ~ 0.1 and $\gtrsim 10\times$ that of water and peaking around the density of water.
- ⋆ Most transiting Jupiter-mass exoplanets are slightly larger than Jupiter; in fact, where the density of Jupiter is a little above that of water, the density of most of the transiting Jupiter-mass exoplanets is below that of water. Of all planets in the Solar System, only Saturn has such low density (≈ 0.7 that of water).
- ⋆ The super-Earths are generally denser than water (but some are not).
- ⋆ There exists a density–mass relationship, hinting at the processes of formation and evolution of planets. This correlation is positive for massive planets but negative for low-mass planets, with a minimum density for planets of ~ 0.4 $\mathrm{M_{Jupiter}}$.
- ⋆ The scatter is greater for low-mass planets; this may be due to measurement errors or real variations between these planets – e.g., the presence or lack of a deep ocean.

- ⋆ Likewise, a relation exists between surface gravity and mass; planets less massive than Saturn appear to have similar surface gravity to that of Earth ($\sim 10^3$ cm s^{-2}).
- ⋆ The radius of massive planets does not depend on their mass over two orders of magnitude; below 0.4 $M_{\rm Jupiter}$ a positive correlation exists between radius and mass.

Note: the density of water $\rho(H_2O) = 1$ g cm^{-3}; the density of rock ρ(rock) ~ 5 g cm^{-3}.

The higher density of the smallest planet may be understood in terms of a composition in which rocky material dominates over that of volatiles such as water, hydrogen, and helium. The higher density of the most massive planets may be understood in terms of their large masses compressing these objects so much that the densities become higher than usual. In fact, brown dwarves are even denser and smaller than the largest Jupiter-mass exoplanets. More massive objects, stars burning hydrogen in their cores become much larger and less dense again as the radiation pressure by the photons produced in the nuclear fusion counter-balances the gravity of the stellar mantle.

10.1.2 Hot Jupiters

The largest Jupiter-mass exoplanets are found close to their host star. These are the so-called 'Hot Jupiters'. Temperatures have been measured in excess of 1500 K – at which rocky material like silicates would sublimate. In some cases, brightness or even (using colours) temperature maps of the planets have been made, showing a hot spot at the dayside (the side of the planet facing the star). Their atmospheres are extended, hence the low overall density. Clearly, these planets are affected by the proximity of the star.

Could these Hot Jupiters be falling into the star? There is some evidence to support this scenario: the old star 16 Cyg A is rich in the lithium isotope ^{6}Li. This would normally be processed within the star, and disappear. Its companion star, 16 Cyg B is not lithium enriched. This suggests that 16 Cyg A has received a fresh supply of unprocessed material, for instance by a planet crashing into it. The difference in refractory elemental composition between the HD 240430 and HD 240429 pair may be explained in a similar way.

However, the frequency with which Hot Jupiters are found suggests that these are in more-or-less stable orbits, otherwise there would be far fewer so close to the star at any given time – though indeed the vast majority of planet hosts do *not* have a Hot Jupiter.

10.1.3 Radiative heating and evaporation

Considering these giant, gaseous planets, their atmospheres work in pretty much the same way as atmospheres around rocky planets like Earth, in the sense that stellar radiation will penetrate and heat the planet, but that the planet will have some difficulty radiating this energy away. This Greenhouse effect leads to storage of energy in the atmosphere and an increase in the planet's surface temperature. In the case of a gaseous giant planet, there is no hard surface, but there is an equivalent depth in the atmosphere that serves much like the surface in the case of a rocky planet: it is the depth down to which photons on average penetrate the atmosphere before they are absorbed or scattered by atoms.

Similar to the concept of the mean free path of an atom before it collides with another, optical depth τ is defined in terms of the fraction to which the intensity of light is reduced:

$$\mathrm{d}I \;=\; -I\,\kappa\,\mathrm{d}z \;\equiv\; -I\,\mathrm{d}\tau \;\Rightarrow\; I \;=\; I_0\,\exp(-\tau). \tag{10.2}$$

The absorbed radiation at the dayside of the planet may heat the atmosphere, increase its pressure, or cause the atmosphere to expand. If the atmosphere were not expanding, then it would heat up at a rate $\mathrm{d}Q/\mathrm{d}t = C_\mathrm{V}\,\mathrm{d}T/\mathrm{d}t$, where the heat capacity, $C_\mathrm{V} \simeq \frac{5}{2}Nk$ (diatomic gas). In reality, it will be a combination of these thermo-dynamical processes as the atmosphere is neither confined in volume, temperature or pressure.

The large temperature gradient between the day- and nightsides likely builds a large pressure gradient, which then drives bulk flow in the atmosphere (wind). This flow carries thermal energy with it, leading to some redistribution of heat over the planet's "surface".

The high temperature in the upper atmospheric regions leads to evaporation of volatile species. Such exosphere has indeed been observed in a few cases, when it passes in front of the star. The spectrum of the star then shows additional absorption lines by atoms in the planetary exosphere. In this way, the evaporation of the light elements hydrogen, oxygen and carbon has been observed in the exosphere of planet HD 209458 b, also named Osiris after the Egyptian god of the afterlife, whilst the heavier element silicon was not detected. Sodium was detected lower in another planet's atmosphere, HD 189733 b.

In both these planets' atmospheres, water was detected by comparison of the eclipse of the planet by the star observed at different wavelengths, with the spectrum of water. Obviously, this is super-hot water *vapour* (steam), and as such not of immediate interest for life. But it is reassuring that water does form and persist in exoplanetary systems.

It has become possible to directly record the infrared spectrum of exoplanets, including HR 8799 c – it is very noisy but appears to be shaped by the absorption by water vapour.

HD 149026 b is another fairly massive planet very near the star it orbits. But it is rather small, rendering it of surprisingly high density; it has been pondered that this is because it has lost much of its light-weight mantle leaving the denser core to contribute relatively more to the planet's mass. Possibly, with time, this planet will lose the remainder of its atmosphere, leaving exposed a large rocky core – scorchingly hot.

10.1.4 Radiation pressure

The exosphere is subject to radiation pressure exerted upon the atoms by the photons emitted by the star. From the usual meaning of pressure we obtain that the radiation pressure is given by the rate at which photon momentum flowing through an imaginary surface area in the exosphere is absorbed by, or scattered off, the atoms:

$$P_\mathrm{rad} \;=\; \frac{\kappa L_\star}{c}\,\frac{1}{4\pi d^2}, \tag{10.3}$$

where the momentum of a photon is given by its energy divided by the speed of light, and κ is the efficiency with which the atoms interact with the photon. Note that the 'rate' is obtained as the luminosity is the rate at which photons are emitted, and the 'flux' through the surface area is obtained by spreading it out over the surface of the sphere at distance d from the star.

The efficiency, $\kappa = 2$ for a perfect reflecting surface (after the interaction with the atom, the photon moves back in the opposite direction), but a tenuous gas behaves differently: $\kappa = \kappa(\lambda)$ where the dependence on the wavelength λ depends on the atom. Atoms generally only interact with photons in a couple of narrow energy bands. However, a mixture of different atoms and in different ionization states collectively interacts with photons over a larger portion of the energy spectrum. If the atoms collide frequently enough with each other they behave like a single fluid, and the radiation pressure will push the fluid as if it were one single entity. Thus, $P_{\rm rad} = \int P(\lambda)\ {\rm d}\lambda$.

10.1.5 Stellar wind pressure

The exosphere is also subject to a stream of particles emanating from the star. This stellar wind takes different forms in different evolutionary stages and on the main sequence is a clear function of the stellar mass. The Sun blows a wind of typically 400 km s^{-1} (compare this with the escape velocity of the Sun, with the orbital speed of Earth, and with the typical speed of an atom in Earth's atmosphere). In this way, the Sun loses mass at a rate ${\rm d}M/{\rm d}t \sim 10^{-14}$ M$_\odot$ yr^{-1}. Over its 10 Gyr lifetime this does not matter for the Sun. But it does matter for a tenuous gas in its path, as the wind will press upon it.

As before, this 'ram pressure' is obtained from the rate at which impinging momentum flows through an imaginary surface area placed in the exosphere:

$$P_{\rm ram} \ = \ \frac{{\rm d}M}{{\rm d}t} v_{\rm wind}\ \frac{1}{4\pi d^2}, \tag{10.4}$$

where the mass-loss rate is also related to the density in the wind, $\rho_{\rm wind}$, and the wind speed, $v_{\rm wind}$, through the continuity equation (a.k.a. "what goes in must come out"):

$$\frac{{\rm d}M}{{\rm d}t} \ = \ 4\pi d^2 \rho_{\rm wind}\ v_{\rm wind}. \tag{10.5}$$

Thus, the simple and familiar expression for ram pressure is obtained:

$$P_{\rm ram} \ = \ \rho_{\rm wind}\ v_{\rm wind}^2. \tag{10.6}$$

Note the strong dependence on the wind speed. This is also the reason why in the Earth's atmosphere, falling rain drops attain a constant speed as further acceleration would be counter-acted by the consequent increase in ram pressure. Note also that, although in Equation 10.6 there is no explicit dependence on the distance from the star, the density in the wind of course drops as the outstreaming wind dilutes. The wind also carries thermal gas pressure, nkT and potentially magnetic pressure, $B^2/2\mu_0$.

10.2 Have we found planets in their Habitable Zone?

Although super-Earths, of several M$_\oplus$, are being found, the Habitable Zone tends to be rather narrow and not very close to the star where it is easier to detect these small planets. For instance, the 7.5 M$_\oplus$ planet Gliese 876 d (also named Alpan) is only 1.8 times larger than the Earth in diameter, seems to have thick clouds and could sustain active volcanism.

It is, however, at a mere 0.02 au from its host star, and definitely inside of the Habitable Zone surrounding that star. But two more planets have been found orbiting the same star. One, Gliese 876 c probably spends some time within the Habitable Zone. However, the orbit of this planet is highly non-circular, with an eccentricity of $e = 0.26$ (see below). This takes it well outside of the Habitable Zone for some part of its year. The other, Gliese 876 b does have a fairly circular orbit, with $e = 0.034$, but it is near the outer edge of the Habitable Zone in a similar position where Mars is found in the Solar System.

Unfortunately, the Earth-mass planet found around α Centauri B orbits its solar-type host star at just 0.04 au. The Earth-mass planet around Proxima Centauri, on the other hand is with 0.05 au in the Habitable Zone of this cool dwarf. But erratic behaviour of the star in the form of energetic outbursts ('flares') and tidal locking may render it inhospitable. Photosynthesis is challenging: phytoplankton uses blue/green light, as it penetrates water, but cool dwarf stars are very red and very faint in the blue/green. Likewise, life on the three planets in the Habitable Zone around TRAPPIST-1 may face severe challenges.

To date, arguably the best candidate for habitability is Kepler-186f, which is both Earth-like and in a near-circular orbit within the Habitable Zone of a star half the Sun's size.

10.2.1 Moons around massive gas planets in the Habitable Zone

Jupiter-like planets have been found within the Habitable Zone of their host star. The most massive such planets might well have moons the size of Mars – or even Earth – and these could provide the platform for life to have developed and evolved (cf. Titan).

10.2.2 Planets around giant stars

When stars exhaust hydrogen in their core, they switch the mode in which nuclear fusion takes place in their interiors and consequently swell up to become cool giants. Planets, of a few $M_{Jupiter}$, have been found in orbits of ~ 1 au near stars that at some point will be nearly as big as – or bigger than – those orbits, a stark reminder of the fate of Earth when the Sun becomes such red giant star in a few billion years time. The red giant will shed its mantle, though, and leave its tiny, dense, and initially hot core: a 'white dwarf'. Some white dwarves have been "polluted" with refractory material thought to come from disintegrated erstwhile planets. On the other hand, some binary white dwarves are surrounded by a disc of material in which some believe new planets may be forming.

10.2.3 Planets around pulsars

The first convincing detections of exoplanets were made around pulsars. Life on those planets is highly unlikely. But, intriguingly, these planets include the smallest discovered so far, and in fact include examples of dwarf planets. These may have survived an event as violent as a supernova explosion of the massive star that created the pulsar.

One could perhaps imagine technologically advanced civilizations to establish an outpost on such planet, with a powerplant tapping into the mind-boggling energy resources locked in the magnetic field and rotation of the pulsar.

10.3 Orbits and gravitational perturbations

10.3.1 The importance of eccentricity

It is no good (from the point of view of aspiring lifeforms) for a planet to spend part of its orbit within the Habitable Zone, and part of it outside: elliptical orbits are problematic.

The degree of ellipticity of an orbit is measured by its eccentricity, defined as:

$$e \;=\; \frac{c}{a} \;=\; 1 \;-\; \frac{b}{a}, \tag{10.7}$$

where c is the distance between the centre of the ellipse and either of its two foci, and a and b are the semi-major and semi-minor axes (half the length of the widest and narrowest separation between two points on the ellipse). Thus, $0 < e < 1$ (zero for a perfect circle). If $M_\circ/M_\star \sim 0$ the closest approach to the star, 'peri-astron' (perihelion in our case) is $a(1-e)$, and the furthest, 'ap-astron' (aphelion) is $a(1+e)$.

Gravity balances orbital motion only at a specific separation, resulting in tides on either side of it. If the planet is spinning around its axis at a different rate than it travels around its orbit, the tides will move around the planet. As the tides deform the planet (according to Young's modulus), there will be a torque. Modulated deformation causes friction and thus energy dissipation. As a result, the spinning frequency and orbital period will change until they are equal: 'tidal locking'. Then, the planet will always face the star with the same side. The tidal dissipation is especially efficient if the orbit resonates with the Brunt–Väisälä buoyancy frequency ("bobbing") in the atmosphere of a gas giant. Around sufficiently cool dwarf stars, planets in the Habitable Zone will be tidally locked.

A planet in a non-circular orbit can never be fully tidally locked, as the orbital speed varies but the huge inertia of a planet prevents such quick adjustment of the spinning frequency. Hence, the tidal interaction will also force the orbit to become circular. The Moon is nearly there. Mercury has not quite reached this point – its day and year are in a 3:2 resonance (it revolves exactly three times over the course of two orbits).

Indeed, exoplanetary orbits close to the host stars tend to be circular, with highly eccentric orbits limited to distant planets. A remarkable observation, though, is that many planets *are* in fact on highly eccentric orbits, and that it is rather uncommon to find planets at an au or so from the star that have orbits as nearly circular as that of Earth (or an entire planetary system as little-eccentric as ours).

10.3.2 The importance of multiplicity

About every two in three Sun-like stars is part of a binary system, i.e. the star has a *stellar* companion. More massive stars are nearly all binaries (or higher multiples), but stars much smaller than the Sun are generally found to be single.

Stellar multiplicity will affect the orbits of any planets in the system, as well as the location (and shape) of the Habitable Zone. The most stable planetary orbits will be encountered either much closer to one of the stars than their separation, or keeping a wide berth around both stars.

Chapter 11

The formation of stars

11.1 The interstellar medium

Planets are believed to form during or immediately after the formation of stars. Very young stars are seen amidst dense clouds of gas and dust, whilst older stars do not have this association. This suggests that stars form from cold clouds. These clouds form from warmer gas, through thermal instability or cloud–cloud collisions, and will disperse on timescales of $\sim 10^7$ years to become part of warmer gas again. The interstellar space is a dynamic environment, in which phase transitions occur between different phases of the interstellar medium (ISM):

- ⋆ Cold neutral medium, typically 10–100 K and "dense" (more than $\sim$ 100 atoms cm^{-3}, sometimes by many orders of magnitude). Much of the gas is in molecular form, and dust grains play an important rôle in the behaviour of these clouds. Clouds in this phase are generally compact structures, that fill little space despite containing relatively significant amounts of mass. In the discs of galaxies like the Milky Way, an as yet unknown mechanism generates spiral arms, in which molecular gas accumulates. Molecular clouds contain a lot of substructure and internal motion.

- ⋆ Warm neutral medium and weakly ionized medium, 5000–8000 K and diffuse ($\sim$ 1 atom cm^{-3}). In the discs of galaxies like the Milky Way, most of interstellar space is filled with gas in this phase, and within which the cold clouds are immersed.

- ⋆ Hot, million K plasma, extremely rarefied at $\ll$ 1 atom cm^{-3}. This is found in localized pockets within galaxy discs, where there are local sources of energy input, and widespread in the exposed surroundings of galaxies (e.g., the Milky Way halo).

11.2 Gravitational instability

The dynamics of a cloud are described by the following set of four relations:

- ⋆ If no mass is created or annihilated, then if it moves out of (or into) a certain location within the cloud then it must have moved to (or from) somewhere else. This conservation of mass is more generally known as the 'continuity equation',

which in one spatial dimension (x) takes the form:

$$\frac{\partial \rho}{\partial t} \; + \; \frac{\partial}{\partial x}(\rho v) \; = \; 0. \tag{11.1}$$

⋆ Momentum obeys a relation similar to the continuity equation, the main difference being that there can be clear sources for momentum generation: forces. For now, we consider the force due to a pressure gradient, and the force due to gravity. The 'momentum equation', in one spatial dimension, takes the form:

$$\rho \frac{\partial v}{\partial t} \; + \; \rho v \frac{\partial v}{\partial x} \; = \; - \; \frac{\partial P}{\partial x} \; - \; \rho \frac{\partial \Phi}{\partial x}. \tag{11.2}$$

⋆ The relation between the thermodynamic quantities pressure, temperature, and density is called the 'equation of state':

$$P \; = \; \rho \; \frac{kT}{m}, \tag{11.3}$$

which we can also express in terms of the sound speed:

$$c_{\rm s} \; = \; \left(\frac{\gamma k T}{m} \right)^{\frac{1}{2}}, \tag{11.4}$$

where γ is the ratio of specific heats (~ 1.4), as follows:

$$P \; = \; \rho \; \frac{c_{\rm s}^2}{\gamma}. \tag{11.5}$$

Note the similarity with the expression for the ram pressure (Equation 10.6). This is not surprising, as both arise from the collision between matter.

⋆ The gravitational potential, Φ is determined by the distribution of matter. This is described by the 'Poisson equation', which in one spatial dimension takes the form:

$$\frac{\partial^2 \Phi}{\partial x^2} \; = \; 4 \pi G \; \rho. \tag{11.6}$$

Clouds create a gravitational potential which is minimal in its centre-of-mass. Hence, the cloud would contract until all matter is concentrated in that point... unless gravity is balanced by an outward force.

One such force is due to the pressure from the gas in the cloud. This thermal pressure appeared in the momentum equation and in the equation of state. Note that its units are the same as that of the energy density. Indeed, a cloud of mass M and radius R is held in balance if the thermal pressure equals the energy density in the gravitational field:

$$\rho \; \frac{GM}{R} \; \equiv \; \rho \; \frac{c_{\rm s}^2}{\gamma}. \tag{11.7}$$

To form stars, clouds must not be in balance: gravity must prevail over the thermal pressure. To see how, and on what scales, this would occur, we have to apply perturbation

theory to the above set of equations that govern the dynamics of the cloud. To that aim, we consider an initially uniform cloud in terms of its density, ρ_0, and hence gravitational potential, Φ_0, and for the sake of simplicity we assume it is at rest, $v_0 = 0$. We then consider a small perturbation, in the form of a plane wave traveling through the cloud, which can be described as:

$$\rho \;=\; \rho_0 \;+\; \delta\rho \; e^{i(kx-\omega t)}, \tag{11.8}$$

where k is the wavenumber and ω the angular frequency of the oscillation, and $\delta\rho$ is a small amplitude in density. Likewise, the gravitational potential then is:

$$\Phi \;=\; \Phi_0 \;+\; \delta\Phi \; e^{i(kx-\omega t)}, \tag{11.9}$$

and the velocity is:

$$v \;=\; \delta v \; e^{i(kx-\omega t)}. \tag{11.10}$$

We substitute this wave into Equations 11.1, 11.2, 11.5 and 11.6. We next realise that $\partial\rho_0/\partial t = 0$ (the cloud was initially static). We also realise that terms with products of two (or more) small amplitudes δ vanish compared to terms with just one such small amplitude – this is called linearisation of the equations. Some simple book keeping and crossing away of the exponentials then yields the following set of four equations:

$$-i\;\omega\;\delta\rho \;+ i\;k\;\rho_0\;\delta v \;=\; 0, \tag{11.11}$$

$$-i\;\omega\;\rho_0\;\delta v \;=\; -i\;k\;\delta P \;-\; i\;k\;\rho_0\;\delta\Phi. \tag{11.12}$$

$$\delta P \;=\; \delta\rho\;\frac{c_{\rm s}^2}{\gamma}, \tag{11.13}$$

$$-k^2\;\delta\Phi \;=\; 4\pi G\;\delta\rho, \tag{11.14}$$

In order to evaluate the stability of the cloud against the perturbing wave, we seek to derive the dispersion relation, which relates the angular frequency to the wavenumber. To this aim, we perform the following mathematical operations on the above equations:

- ⋆ Multiply Equation 11.11 by $-i\omega$.
- ⋆ Multiply Equation 11.12 by ik.
- ⋆ Subtract the latter from the former.
- ⋆ Substitute δP using Equation 11.13.
- ⋆ Substitute $\delta\Phi$ using Equation 11.14.

Hence the dispersion relation is obtained:

$$\omega^2 \;=\; \frac{c_{\rm s}^2}{\gamma}\,(k^2 - k_{\rm J}^2), \tag{11.15}$$

where the Jeans wavenumber $k_{\rm J}$ is introduced by definition:

$$k_{\rm J}^2\,\frac{c_{\rm s}^2}{\gamma} \;\equiv\; 4\pi G\rho_0. \tag{11.16}$$

The important result is that if $k < k_{\rm J}$ then ω is imaginary, and the wave will grow exponentially leading to the collapse of the cloud. This means that clouds are intrinsically unstable on scales larger than the Jeans wavelength, $\lambda_{\rm J} \equiv 2\pi/k_{\rm J}$:

$$\lambda_{\rm J} \;\equiv\; \left(\frac{\pi c_{\rm s}^2}{\gamma G\rho_0}\right)^{\frac{1}{2}}. \tag{11.17}$$

This corresponds to a Jeans mass threshold of approximately:

$$M_{\rm J} \;\sim\; \frac{4}{3}\pi\lambda_{\rm J}^3\,\rho_0 \;=\; \frac{4}{3}\pi\left(\frac{\pi c_{\rm s}^2}{\gamma G}\right)^{\frac{3}{2}}\;\rho_0^{-\frac{1}{2}}. \tag{11.18}$$

(conventions vary slightly between different authors) Clouds more massive than this are unstable and collapse. If they are significantly more massive than this, then they do not collapse monolithically but will fragment on the typical scale of the Jeans mass. This may be the reason why stars generally do not form in isolation but in clusters of hundreds or more, and it may explain the end product of star formation: the distribution function of stars over their masses – the 'Initial Mass Function' (IMF).

The IMF peaks near a solar mass – i.e. the Sun is a fairly common type of star. The IMF falls off steeply towards higher masses, luminous stars being rare. It also falls off towards low masses, brown dwarves not being as plentiful as once envisaged.

Besides gravity, other forces are at play in a collapsing cloud. These tend to counteract gravity, thus slowing, halting, or even reversing the collapse. In order to form a star, these forces must somehow be overcome. Eventually, it is the radiation pressure due to the nuclear fusion inside the young star which puts an end to the collapse. The efficiency with which a molecular cloud is turned into stars is usually just a few per cent.

11.3 Hydrostatic equilibrium and the need to cool

In the above analysis of cloud stability, the cloud was considered to remain isothermal, guaranteeing the invariance of the sound speed. In reality, the contracting cloud does not only become denser but also warmer as gravitational potential energy is converted into heat. The pressure thus increases, until it balances gravity and the collapse stalls.

A new, hydrostatic, equilibrium is established. The structure of such (spherical) cloud is imposed by the condition of force equilibrium:

$$\frac{\mathrm{d}P}{\mathrm{d}r} \;=\; -\rho\,\frac{GM(<r)}{r^2}, \tag{11.19}$$

where the mass distribution obeys:

$$\frac{\mathrm{d}M}{\mathrm{d}r} \;=\; 4\pi\,r^2\,\rho. \tag{11.20}$$

Solutions to these equations are named 'Bonnor–Ebert' spheres. One special case is that of the 'Singular Isothermal Sphere' (SIS):

$$\rho \;=\; \frac{c_\mathrm{s}^2}{2\pi G\,r^2}. \tag{11.21}$$

The "Singular" in SIS arises from the fact that the density in the centre ($r = 0$) tends to infinity. The "Isothermal" in SIS arises from the fact that in this case the sound speed is the same everywhere within the cloud.

The temperature structure of dense clouds can be probed in various ways, for instance:

- ⋆ Via the spectral energy distribution of thermal emission from dust within the cloud: at high densities, $\gtrsim 10^5$ cm^{-3}, dust grains and molecules collide frequently enough to lead to equipartition of energy – 'Local Thermodynamic Equilibrium' (LTE).
- ⋆ Via the thermal broadening (Doppler broadening due to thermal motions in the gas) of spectral lines arising from gas within the cloud. Internal kinematic structure or turbulence cause similar line broadening, so this "temperature" is an upper limit.
- ⋆ Via the intensity ratios of the spectral lines of atoms and molecules. As these lines correspond to different energy levels their ratios depend on the energy spectrum of the source of excitation. This is straightforward if collisions are responsible for the excitation, in which case the energy spectrum directly relates to the temperature of the gas. But it is not straightforward if the excitation is due to photons, as the energy spectrum of the radiation field may differ from that of the gas: $T_\mathrm{excitation} \neq T_\mathrm{kinetic}$.

The cloud may continue to collapse if it can cool efficiently:

- ⋆ Via radiation in atomic fine-structure lines, at infrared wavelengths where the cloud is relatively transparent so the energy can escape. Examples are lines of neutral oxygen at $\lambda = 63$ μm and of singly-ionized carbon (C^+) at $\lambda = 158$ μm.
- ⋆ Via radiation in rotational and vibrational transitions of abundant molecules, which also tend to have lines at infrared (and microwave) wavelengths. Examples are bands of water and lines of CO. The most abundant molecule, molecular hydrogen (H_2) is symmetric and thus lacks strong transitions.
- ⋆ Via radiation by dust grains, according to their temperature. Dust at temperatures $T \ll 1000$ K mainly shines at infrared and microwave wavelengths.

However, there are also mechanisms at play which lead to heating of the cloud:

- ⋆ Via the photo-electric effect. Here, an energetic photon hitting a dust grain releases an electron from the grain. The electron carries the excess energy in the form of kinetic energy, which it then shares with any atom it collides with.
- ⋆ Via ionization of atoms and molecules, and heating of dust grains, by 'cosmic rays'. These are fast particles which easily penetrate dense clouds. They arise from places in the Universe where very strong magnetic fields accelerate plasma to high energies.
- ⋆ Via radiation from stars, either outside the cloud or within.

11.3.1 Interstellar chemistry and the building blocks for life

It is obvious that the rate at which a cloud cools depends on the chemistry of the cloud:

- ⋆ Tiny dust grains (~ 0.1 μm in size) are formed in the outflows of cool evolved stars. Within dense molecular clouds, these grains can grow considerably by adsorption of molecules, coating the grains with ice mantles. The freeze-out from the gas phase leads to dramatic changes in the chemical make-up of the gas.
- ⋆ Grain coatings provide a surface on (or in) which chemical reactions can take place that in the gas phase are much slower. This is the dominant formation route for H_2.
- ⋆ Simple molecules such as CO, CO_2, water and ammonia are the building blocks for larger molecules, from formaldehyde (H_2CO) and methanol (CH_3OH) to complex organic molecules. Cold chemistry also leads to deuteration – an increased D/H. The molecular make-up and deuterium fraction of the gas act as a chemical clock.
- ⋆ Complex organic molecules include prebiotic molecules. Glycolonitrile ($HOCH_2CN$) has been found in protostellar clouds; it derives from formaldehyde and could react with water to form the amino acid glycine (NH_2CH_2COOH). The Z isomer (atomic arrangement within a molecule) of cyanomethanimine (HNCHCN), also found in these interstellar environments, is believed to be involved in forming the nucleic acid adenine ($C_5H_5N_5$). These molecules have a high energy threshold of formation, and the fact that they can be present already when planets form around young stars means life is given a – potentially critical – head start.

11.4 The next hurdle: magnetic fields

Because even the cold ISM is not completely neutral, moving charged particles create or maintain an interstellar magnetic field. A charged particle moving with respect to the magnetic field experiences a (Lorentz) force perpendicular to the magnetic field lines but also perpendicular to its original trajectory:

$$\vec{F}_{\rm B} \;=\; q\;\vec{v}\times\vec{B}, \tag{11.22}$$

where q is the particle charge and $B = |\vec{B}|$ is the magnetic field strength. Hence the particle is deflected from its path. As a result, gas will have difficulty crossing magnetic field lines. The magnetized gas can be said to exert a magnetic pressure:

$$P_{\rm B} \;=\; \frac{B^2}{2\mu_0}, \tag{11.23}$$

where $\mu_0 = 4\pi\ 10^{-7}$ N A^{-2} is the magnetic permeability of vacuum. Note that in an ordered magnetic field the magnetic pressure is not isotropic (the same in all directions), but directed perpendicular to the field lines. The momentum equation would need to include the magnetic pressure as well as tension, which requires a full 3D description.

The magnetic pressure might balance the ram pressure (Equation 10.6) resulting from infalling matter:

$$P_{\rm ram} \;=\; \frac{{\rm d}M}{{\rm d}t} v_{\rm infall}\ \frac{1}{4\pi r^2}, \tag{11.24}$$

where $\dot{M}_{\rm infall} \equiv {\rm d}M/{\rm d}t$ is the rate of infall. The magnetic flux,

$$\Phi_{\rm B} \;=\; \vec{B} \cdot \vec{S}, \tag{11.25}$$

where $\vec{S}$ is the surface[1]. Gauss's law insists that the total magnetic flux through a closed surface, $\Phi_{\rm B} = \oint \vec{B} \cdot {\rm d}\vec{S} = 0$. This implies that if the magnetic field is compressed, i.e. the surface area is reduced, the magnetic field must strengthen in proportion.

The magnetic field barrier is overcome by turbulence (as in "folding" when preparing batter) or by a process called 'ambipolar diffusion': the neutral particles within the cloud do not feel the effect of the magnetic field, and they can cross the magnetic field lines unhindered. They do collide with charged particles within the cloud, which cannot cross the magnetic field lines (easily). This exerts a drag, slowing the diffusion. Diffusion is already a slow process without this drag, compared to bulk flow as a result of external forces (e.g., gravity, pressure gradients), as it results from random-walk thermal motion of atoms and molecules. The neutral matter piles up, increasing the gravitational pull on the charged matter until it starts having some success in crossing the magnetic field lines.

Recent observations show the ubiquity of filamentary structure in molecular clouds, which formed as a result of the gravitational contraction along the magnetic field (which is perpendicular to the filament). In the densest, most neutral parts, gravity is strong enough to overcome the magnetic pressure, and further collapse proceeds *along* the filaments.

Magnetic fields can be measured, for instance:

- ⋆ Via Zeeman splitting of energy levels in atoms and molecules.
- ⋆ Via Faraday rotation of the polarization direction of electro–magnetic waves as they traverse the cloud.

Unfortunately, we're not done yet. There are further complications in forming a star. However, a by-product of this process is the formation of a disc of material from which planets (can) form.

[1]This may seem counter-intuïtive and one might have expected flux to divide a quantity by the area, just like radiation flux is the luminosity per unit area. But $\vec{B}$ is already a magnetic field *flux density*.

Chapter 12

Proto-planetary discs

12.1 The angular momentum catastrophy

Because of the dynamical nature of the ISM, clouds and portions of clouds will always have some degree of rotation; they are known to be turbulent. In the absence of external torques, the conservation of angular momentum (L, not to be confused with luminosity) dictates that the rotation rate ($\omega = v/R$) of a contracting cloud increases unless changes to the configuration of the system lead to sufficiently large increase in the moment of inertia (I):

$$L \;=\; I\,\omega. \tag{12.1}$$

Once the rotation velocity reaches the escape velocity the contraction stalls.

The moment of inertia can be computed:

$$I \;=\; \int_0^M r^2 \,\mathrm{d}m. \tag{12.2}$$

For instance, for a uniform sphere it is $I_{\mathrm{sphere}} = \frac{2}{5}MR^2$. For a uniform disc it is marginally larger: $I_{\mathrm{disc}} = \frac{1}{2}MR^2$. Thus a contracting cloud mitigates speeding up by assuming a flattened geometry, but only to a very small degree. The moment of inertia of a ring is $I_{\mathrm{ring}} = MR^2$, hence by moving some matter outward through the disc the angular momentum in the centre of the disc may be reduced further.

The moment of inertia of two clumps, each of mass $\frac{1}{2}M$ and each at a distance R from the centre-of-mass, is $I_{\mathrm{binary}} = MR^2$. Thus, fragmentation into clumps can also be a mechanism through which to redistribute the angular momentum so as to avoid very high rotation speeds. It is therefore not surprising that many stars are found to be part of binary or multiple-star systems.

12.2 Star–disc interaction

Proto-stars are fully convective balls of gas, in hydrostatic equilibrium. They contract by cooling, along what is named the 'Hayashi track' in the temperature–luminosity diagram (which is also named the 'Hertzsprung–Russell Diagram'). They become fainter.

Once the pressure is high enough to ignite nuclear fusion reactions, a pre-Main Sequence star is born. Its outer layers continue to contract and heat up until it settles on the Main Sequence. This happens on the so-called 'Kelvin–Helmholtz timescale' ($\sim 10^6$ yr), which is set by the ratio between the gravitational potential energy and the luminosity at which this is being radiated away:

$$t_{\rm KH} \;=\; \frac{GM^2}{RL}. \tag{12.3}$$

12.2.1 Accretion and feedback

Proto- and pre-Main Sequence stars continue to attract matter from the surrounding cloud or disc. In the outskirts this accretion happens on roughly the free-fall timescale:

$$t_{\rm ff} \;=\; \frac{R}{v_{\rm ff}}, \tag{12.4}$$

where the free-fall velocity is the same as the escape velocity, which in a hydrostatic equilibrium is very similar to the sound speed (see Equation 11.7):

$$v_{\rm ff} \;=\; \left(\frac{2GM}{R}\right)^{\frac{1}{2}} \;\sim\; c_{\rm s}. \tag{12.5}$$

The accretion rate can thus be estimated:

$$\frac{{\rm d}M}{{\rm d}t} \;\sim\; \frac{M}{t_{\rm ff}} \sim\; \frac{c_{\rm s}^3}{G}. \tag{12.6}$$

In analogy with Equation 10.5, the density can be derived from:

$$\frac{{\rm d}M}{{\rm d}t} \;=\; 4\pi r^2\;\rho\;v_{\rm ff}. \tag{12.7}$$

In the isothermal case ($c_{\rm s}$ constant) this leads to the following density profile of the cloud:

$$\rho \;\propto\; r^{-\,\frac{3}{2}}. \tag{12.8}$$

The gravitational potential energy of the accreting matter is radiated away at a rate:

$$L_{\rm acc} \;\sim\; \frac{GM}{R}\;\frac{{\rm d}M}{{\rm d}t}, \tag{12.9}$$

which emerges as radiation from an emitting cloud with an effective temperature $T_{\rm eff}$:

$$L_{\rm rad} \;=\; 4\pi R^2\;\sigma T_{\rm eff}^4. \tag{12.10}$$

Young stars generally have a strong magnetic field, and the accretion from the disc thus proceeds along the magnetic field lines, ending up on the magnetic poles – which are almost always closer to the poles of the star's spin than to its equator.

With this magnetically channeled accreting matter, angular momentum is also exchanged between disc and star. Young stars tend to spin faster than the disc, and the disc–magnetosphere locking thus leads to the transport of angular momentum away from the star into the disc. If the disc disperses quickly enough, then the star may spin up again as it further contracts towards the Main Sequence.

The infant star soon starts affecting the surroundings. This feedback takes various forms:

- ⋆ Radiation pressure.
- ⋆ Photo-evaporation.
- ⋆ Ram pressure due to a stellar wind, or due to diverted accreting matter in the form of collimated jets emanating from the magnetic poles. This is possibly accompanied by shocks, which dissipate energy through heat and radiation. These outflows stall if they sweep up a quantity of matter similar to that contained within the flow.

In addition to erosion from the inside, discs are also ablated from the outside; by radiation or winds from other stars in the young cluster or by collisions with other stars and discs.

The feedback halts the accretion and erodes the circumstellar disc, and on a grander scale it can halt the star formation process – but there are also mechanisms through which it can in actual fact *induce* the collapse of nearby clouds or cloud cores and henceforth give rise to further epochs of star formation.

12.3 Proto-planetary discs

Although not all circumstellar discs around young stars may go on to form planets, such discs are still considered the prime factory of planets. They are therefore also named ‘proto-planetary discs’.

The gravitational potential of a flattened mass distribution features non-radial gradients. Matter will thus not just flow towards the centre, but also towards the symmetry plane of the flattened geometry. Hence a flattened distribution will reinforce itself and become ever more concentrated towards its mid-plane.

The high density and settlement of dust in the mid-plane makes it a very cold part of the proto-planetary disc, as the dust shields the material from direct irradiation by the star. Beyond a certain distance from the star, the ‘snowline’, the dust grains will be coated with ice. This dust and ice has been detected spectroscopically at infrared wavelengths in a number of proto-planetary discs. It is in the mid-plane where planets are expected to form, and this explains why planetary orbits tend to be co-planar to a very high degree.

The skirts of proto-planetary discs are subjected to irradiation by the star, and they are thus relatively warm. One could think of these regions as the “atmosphere” of the disc, with the dusty mid-plane taking the rôle of the planetary surface. Indeed, the vertical structure of the disc envelope is well-described by an exponential profile with a scaleheight depending on the temperature of the gas and the local gravity. Discs are denser in their inner regions (closer to the star), and the resulting decreasing gravity in the outer parts of the disc leads to an increase in scaleheight of the disc envelope with distance to the star: the outer disc “puffs up” – this is referred to as ‘disc flaring’.

12.4 Observed disc properties

The proto-planetary discs are relatively easily detected at infrared wavelengths where the dust emission is strong. Some more detailed pictures have been obtained at shorter wavelengths, where starlight is seen scattered off the disc (the scattering efficiency of small grains generally increases towards blue wavelengths). It has even been possible to take a spectrum of the inner part of a disc, using spectral lines to probe the velocity field which appears consistent with a Keplerian disc much like the orbits in a planetary system.

The frequency with which discs are seen around stars declines with age (obtained for stars that form part of a cluster). Within a Myr of the formation of a proto-star, essentially all solar-mass stars have an accretion disc. Within 10 Myr, almost all have lost their disc. This sets very firm upper limits on the timescale over which planets can form.

Some – possibly many – stars appear to keep some of the left-over material from the planet formation process, in the form of a gas-free disc of rocks and grains (refractory material such as silicon, iron...). These 'debris discs' no longer provide favourable conditions for planet formation. In our Solar System, the asteroid belt may be thought of as a relic from a debris disc – comets (volatile material such as water, nitrogen...) originate from a more spherical distribution at larger distances, and may be more directly related to the original cloud from which the proto-Sun and its proto-planetary disc formed.

12.4.1 The showcase disc of β Pictoris

One particularly spectacular disc, which is scrutinized with every possible observational technique, is that surrounding the star β Pictoris. Its dust emission is bright at infrared and microwave wavelengths. There have been claims of the detection of H_2 gas, using spectral lines at infrared wavelengths, but this has been disputed on the basis of the absence of the corresponding spectral lines at ultra-violet wavelengths. It is therefore not clear whether this is a debris disc, or whether it is still a gaseous disc in which to witness the process of planet formation. The age of β Pictoris is not precisely known (it is not a member of a cluster), but it is believed to be in the range 8–20 Myr.

The detection of structure in the outer parts of the disc (which is viewed nearly edge-on) led to the interpretation in terms of gravitational resonances due to the presence of planets. The detection of a small secondary disc, inclined with respect to the main disc, also was interpreted as indirect evidence for a planetary object which would have caused the perturbance. In 2008, at last, the discovery was announced of a planet imaged in the disc of β Pictoris, with an estimated 9 ± 3 M_{Jup} (it is too faint for a spectroscopic determination) and in a Saturn-like orbit.

12.4.2 An atlas of proto-planetary discs

Powerful observing facilities have become able to detect and resolve a sizeable *sample* of nearby proto-planetary discs. This allows to identify patterns, dependencies on stellar host, and comparison with models for the structure of these discs as planets form in them.

The SPHERE instrument on the European Southern Observatory's Very Large Telescope in Chile can take very sharp optical images using adaptive optics, reaching the diffraction

limit (Equation 8.4). By blocking the star it can see faint structures surrounding it ('coronagraphy'). The optical light from proto-planetary discs is dominated by light from the star that is scattered off grains in the disc. The light then becomes polarized, and by using a polarizer the contrast can be enhanced with respect to the wings of the star's diffraction pattern. The structures that are seen with SPHERE include rings, disc flaring and central outflows.

The Atacama Large (sub-)Millimetre Array (ALMA) on the 5-km high altiplano of Chile has got the sensitivity and angular resolution to map the distribution of cold dust and gas in proto-planetary discs directly, and also measure the kinematics of the gas. Most proto-planetary discs imaged with ALMA show gaps, and sometimes local concentrations (see Figure 12.1). Both are believed to be directly linked to the formation of planets (see next Chapter). All rotate. Discs around binary stars tend to have spiral structure.

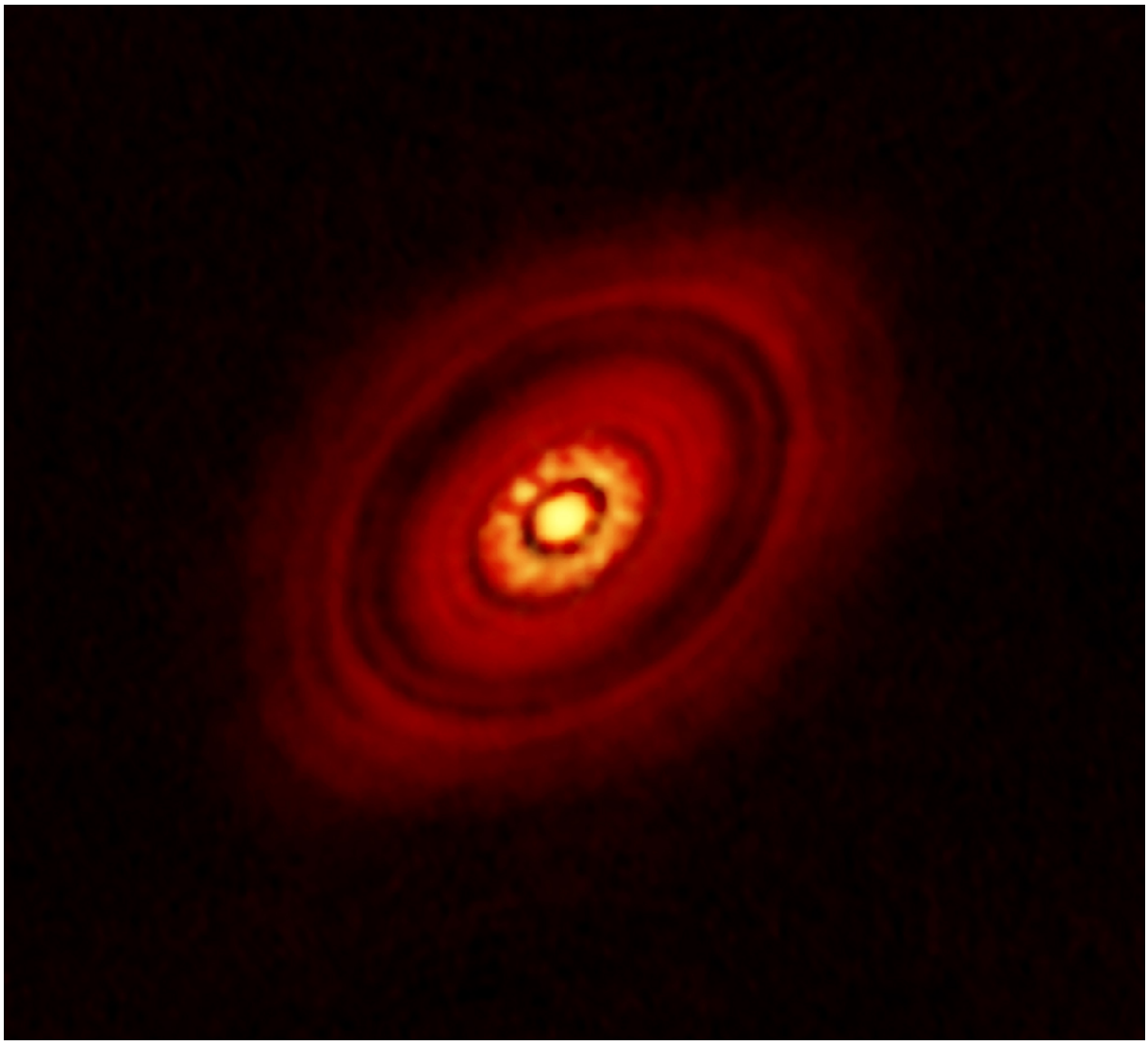

Figure 12.1: ALMA and Very Large Array (VLA) composite of the proto-planetary disc of HL Tauri. The clumpy structure in the bright central ring is only unveiled with the VLA because the dust is optically thick at the high frequencies at which ALMA operates.

Chapter 13

The formation of planetary systems

13.1 The formation of planets

13.1.1 Disc instability

Discs with a high surface density, Σ (in mass-per-area units), can become unstable due to gravity in a similar way as molecular clouds. This happens above a certain threshold – the Toomre criterion:

$$\Sigma_{\rm unstable} \; > \; \Sigma_{\rm critical} \; \equiv \; \frac{\kappa c_{\rm s}}{\pi G}, \tag{13.1}$$

where κ is the so-called 'epi-cycle frequency': orbits can be described with an additional circular motion (the epi-cycle) around a point moving on a circular orbit. For Keplerian (closed elliptical) orbits, simply $\kappa = \omega$.

The disc instability in combination with the disc rotation leads to shear of the instability and the generation of a spiral wave. As the instability intensifies, it fragments and the fragments contract further, until objects are formed that may be massive planets. Given the high disc masses where this mechanism could operate, the newly formed objects could be brown dwarves instead. Discs have also been detected around brown dwarves themselves, though, implying that they may also form in a similar manner as stars.

13.1.2 Planet formation "bottom up"

Observed discs generally are not massive enough for disc instability to occur, and the formation route for planets is believed to be through the co-agulation of small particulates, through pebbles and rocks to small planetary bodies, 'planetesimals'. This process is easier beyond the snowline, as icy grains stick better than bare grains. Electro-static and magnetic forces may also contribute. Boulders are much more difficult to assemble; vortices[1] are invoked to create 'rubble piles' that accumulate sufficient self-gravity to bind. Slow encounters between planetesimals then lead to the formation of molten Earth-size planets; a more violent impact created Earth's Moon.

[1] Perhaps arising from magneto–rotational instability in a differentially rotating magnetised disc.

Beyond the snowline, temperatures are low enough for the rocky proto-planets to acquire an extended gaseous atmosphere in hydrostatic equilibrium. The accretion of gas becomes a runaway process with the ever more massive proto-planet more capable of attracting surrounding gas. Thus the rocky proto-planet becomes the core for a gaseous giant planet. This runaway process explains the large difference in size and composition, observed in the Solar System, between the gaseous giant planets and the rocky planets.

One may wonder what the planet would have looked like, that could possibly have formed out of what is now the main asteroid belt – which includes dwarf planet Ceres – situated between the orbits of Mars and Jupiter. A super-Earth 'waterworld'?

The formation of planets at large distances from the star, in particular that of gaseous giant planets, is at odds with the frequent detection of Hot Jupiters: gaseous giant planets at very close proximity to their host star. How can this be reconciled?

13.1.3 Planet migration

The answer may be: migration. (Though disc instability in the denser central regions of the disc might possibly be invoked form them *in situ* – see Figure 12.1.)

The underlying process through which migration is believed to occur, is the transport of angular momentum through the disc. For this to happen, there needs to be some coupling between material on adjacent orbits, at slightly differing orbital speeds. This is called 'viscosity'. The faster material on an inner orbit will feel a drag from the slower moving material in the outer orbit. This will slow down the former and speed up the latter. This in turn will make the former move inwards and the latter move outwards, thus creating a gap. Because the outer material experiences the opposite effect with regard to material further away still, it does not actually move away but rather follows the inmoving matter. Thus the disc is accreting, by transporting angular momentum outwards. A planet within the disc would participate in this process and migrate inwards.

One distinguishes two flavours of this migration mechanism:

- ⋆ Type I migration: if the planet has a strong connection with the disc. This is the case if the planet is small and embedded within the disc, so that it is in strong, "direct" contact with the disc. As a result, this migration is relatively fast.
- ⋆ Type II migration: in the case of a massive planet, or a planet which has cleared a gap in the disc by the gravitational perturbation of its immediate surroundings (this is also more important for more massive planets), there remains only a weak connection between planet and disc – by gravity over distance and by confined accretion funnels through which the planet sucks away material from the disc. This migration is thus relatively slow.

The migration halts once the planet has become completely detached from the disc, and the planet will then have arrived in its final orbit, closer to the star than where it formed.

Mutual dynamical interaction (through gravity) between the newly formed planets in their initial orbits will also change their orbits until a pseudo-stable arrangement is established. This can even change the order of the planets (or eject planets altogether), as well as their

spin axis orientation (e.g., Uranus is lying "on its side") and tends to circularise the orbits. The outcome, a maximum-entropy state, may be an ordered, dynamically-quiet planetary system in which the orbital periods have simple ratios ('resonances') just as the 'Titius–Bode law' approximately describes the semi-major axes (a, in au) of the orbits in the Solar System: $a = 0.4 + 0.3 \times 2^p$ for $p = -\infty$ (Mercury), 0 (Venus), 1 (Earth), 2 (Mars), 3 (Ceres / asteroid belt), 4 (Jupiter), 5 (Saturn) and 6 (Uranus). While this law was "confirmed" with the discovery of Uranus, which then led to the prediction and discovery of Ceres, the discovery of Neptune led to the failure of the law.

13.2 Gravitational perturbations

Star formation is a dynamical process, generally leading to dense clusters of stars. A young star is thus subjected to irradiation by other young stars, and close encounters with another star can also lead to erosion of the disc and/or the ejection of the star from the cluster. Planets may be torn away from their host star, and become free-floating objects. There have been claims that such objects are indeed seen, but it has been difficult to establish accurately their distances, ages and hence their masses.

Within a planetary system, young or old, gravitational interactions between all members (planets, stars, moons...) continue to affect the orbits of the planets, the stability of planet–moon systems, and even the stability of the planets themselves.

A useful concept in this regard is that of equi-potential surfaces, or 'Lagrangian surfaces', which map the influence spheres of the planets, host star, any companion stars, *et cetera*.

13.2.1 The Hill radius

A disruptive effect by the star would be the detachment of a planet–moon system. The 'Hill radius' is defined as the distance from the planet within which a moon remains gravitationally bound to the planet (it can be modified to describe a star–planet system):

$$r_{\rm H} \;=\; d\,\left(\frac{M_\circ}{3M_\star}\right)^{\frac{1}{3}}, \tag{13.2}$$

where a circular orbit is assumed, and $M_\circ \ll M_\star$. For instance, for the planet Gliese 876 d the Hill radius is only $r_{\rm H} < 5$ R$_\circ$; it is highly unlikely that it has a moon that close.

13.2.2 The Roche limit

Another disruptive effect would be the tidal disruption of the planet by the star. The 'Roche limit' is defined as the distance from the star within which a planet would be torn apart by tidal forces (a similar version describes the disruption of a moon by a planet):

$$d_{\rm R} \;=\; \text{constant}\;\times R_\circ\,\left(\frac{M_\star}{M_\circ}\right)^{\frac{1}{3}}, \tag{13.3}$$

which can also be expressed in terms of the densities of the objects:

$$d_{\rm R} \;=\; \text{constant} \;\times R_\star \left(\frac{\rho_\star}{\rho_\circ}\right)^{\frac{1}{3}}. \tag{13.4}$$

The constant is of order unity, but depends on the kind of material the object is made of – the rigidity, described by the 'Love numbers' and related to Young's Modulus and tensile strength (see Section 3.1.4). For a solid object, the constant $= 2^{1/3} \sim 1.26$, but for a liquid the constant ~ 2.4. Hence, a solid planet could survive closer to a star than a gaseous planet.

Saturn's ring system offers a spectacular example of tidal disruption; with all four gas giant planets in the Solar System in possession of rings this is clearly a common process.

13.2.3 Lidov–Kozai cycles

Planets have been detected in stellar binaries. The planets are always seen to orbit the primary (more massive) star at distances (much) less than a $10^{\rm th}$ of the separation between the two stars that form the binary. Any current effect of the companion star is therefore relatively weak. But systems with Hot Jupiters are more prevalent among binaries.

The distribution of orbital eccentricities of planets in binary stars is indistinguishable from that of planets around single stars. Curiously, the relation between eccentricity and orbital radius extends to the binary star companions, suggesting that these may have been determined by similar processes, in the early phases of star and planet formation.

If the orbit of the companion star is inclined with respect to the orbit of the planet, then the modulation of the gravitational potential due to the companion star not only induces precession of the planetary orbit but also an oscillatory exchange between eccentricity and inclination. These modulations are called 'Lidov–Kozai cycles', and in extreme cases can flip the entire orbit.

13.2.4 The Rossiter–McLaughlin effect

The inclination between the orbit of a transiting planet and the rotation axis of the host star can be measured via the 'Rossiter–McLaughlin effect'. As the planet occults part of the star, that part of the star does not contribute to the formation of the rotationally-broadened spectral line (Figure 13.1). Hence, the centroid of the spectral line is no longer at the same velocity as when the fully illuminated disc is seen. As the planet transits, different parts of the star contributing to different parts of the line profile are occulted.

The resulting signature in the radial velocity curve of the star during the transit is different for different angles between the planetary orbit and the star's rotation. This has now been measured for > 150 systems. Many planet orbits are close to co-planar, i.e. their angular momentum is aligned with that of the star, but there are notable exceptions.

Although the star's rotation axis may not be perfectly aligned with that of the planetary orbits, it is more likely to be so: the angular momentum contained in the stars's rotation and that of the proto-planetary disc from which the planets formed both arose from the contracting natal cloud.

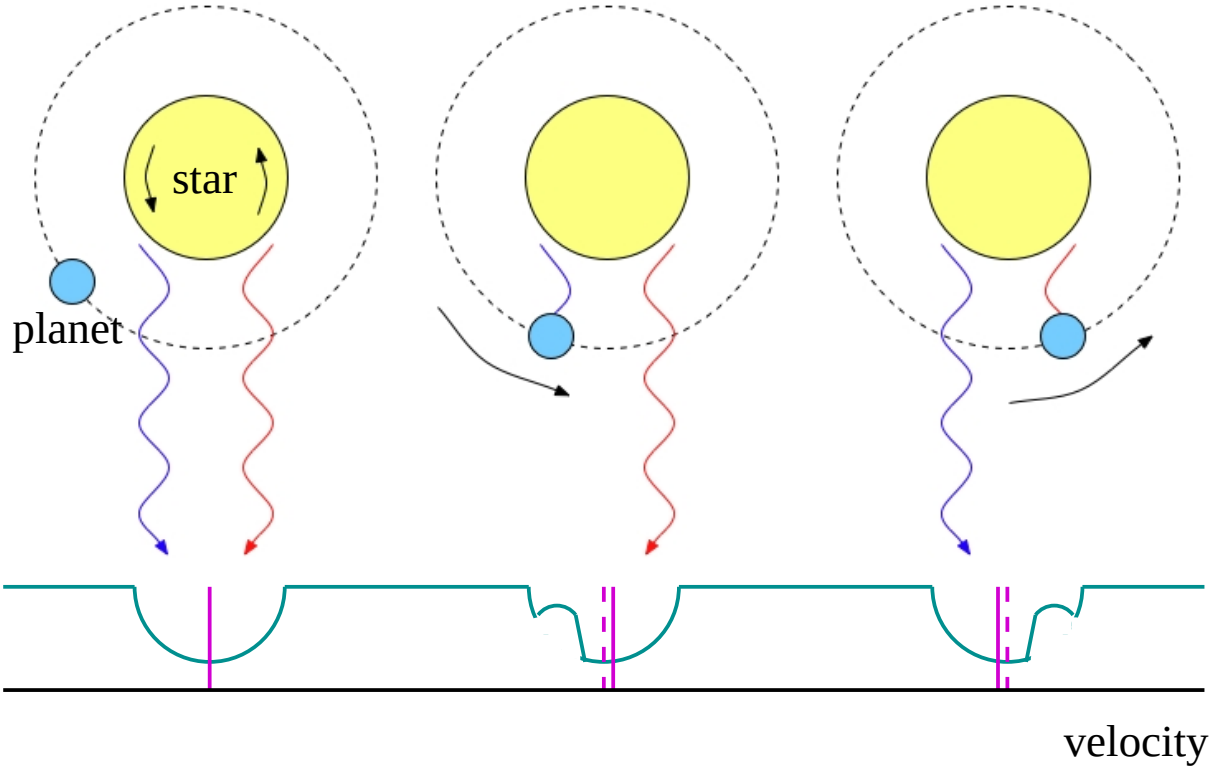

Figure 13.1: Schematic explanation of the Rossiter–McLaughlin effect: as a planet transits a star, it obscures part of the star moving with a specific velocity with respect to the observer due to the star's rotation. This is reflected in a "transit" of the planet through the rotationally-broadened spectral line in the star's spectrum. As a result, the velocity centroid of the spectral line shifts back and forth. The precise realization of this shift depends on the angle between the planet's orbit and the star's rotation.

If more than one planet transits the star, then the relative angle between the orbits of the two planets can be measured.

In the Solar System, the planetary orbits are co-planar to a high degree. The exception is Pluto, the furthest of the classical planets, which has both a relatively high inclination ($i = 17°$) and a high eccentricity ($e = 0.25$) – but this oddity has now been reclassified as a dwarf planet. Curiously, the most extreme "true" planet in the Solar System with the highest inclination and eccentricity is Mercury ($i = 7°$ and $e = 0.21$): this is both the smallest true planet and the one closest to the Sun. This is not understood.

In the Solar System, all planets have prograde orbits, i.e. the same sense as the Sun's spin. All planets spin in a prograde sense, as expected – except Venus, which is puzzling.

Chapter 14

Finding signatures of life

14.1 Planetary conditions for life

It is in many cases feasible to measure the following orbital and planetary properties of exoplanets, in particular if both radial velocity and transit photometry measurements are available or if the planet is imaged directly:

- ⋆ The orbital radius.
- ⋆ The orbital eccentricity.
- ⋆ The planet mass.
- ⋆ The planet radius.

From these, it is possible to determine the following factors that are relevant in the search for habitable planets:

- ⋆ The mean density of the planet, and thus whether it is a gaseous planet (difficult to sustain life) or a rocky planet with a solid surface (possibly covered with an ocean).
- ⋆ The gravity at the surface of the planet and the temperature of the planet, and thus whether the planet is able to sustain an atmosphere and the composition this atmosphere may have (in terms of evaporation rates).
- ⋆ The location and excursions of the planet with respect to the Habitable Zone around the host star, and thus the possibility for liquid water to exist on the planet, somewhere, sometime.

For instance, we have seen that the Earth is a rocky planet with a gravity strong enough to keep an atmosphere containing nitrogen, oxygen, ozone, water and carbon–dioxide. The temperature of Earth is neither too hot to evaporate the atmosphere, nor too cold to freeze out important constituents. Indeed, Earth is located near the inner edge of the Habitable Zone around the Sun, and liquid water covers most of its surface and is present throughout the atmosphere. The orbit of Earth is nearly circular, and the annual variations in conditions on the surface are relatively modest – there are nonetheless significant differences between arctic Winters and Summers.

Mars, on the other hand, is located near the outer edge of the Habitable Zone, and its eccentric orbit probably moves it in and out of the zone. The conditions at its surface underneath the thin atmosphere are near the triple point of water, and although liquid surface water is not common now, it may exist briefly sometimes, and large reservoirs may exist beneath Mars's surface. The (current) lack of tectonic activity and a magnetic field might be critically important for a surface biosphere, too.

Other planets in the Solar System are generally too hot, too cold, do not have enough gravity to sustain an atmosphere with sufficient pressure, or do not have a solid surface. Exceptions to these hostile environments may be found in 'micro-climates' such as the tropopause high up in Venus's atmosphere or a possible sub-surface ocean on Jupiter's moon Europa or Saturn's moon Enceladus.

Hence, depending on the exact conditions on a planet, and given limitations in our ability to model these conditions and the response of an atmosphere, the Habitable Zone in exoplanetary systems can be estimated to some extent but the location of a planet within or outside this formal Habitable Zone is not in itself a certain indicator of life nor the absence of life.

14.2 Atmospheric signatures of a biosphere

In order to sustain enough pressure to keep water liquid, the biosphere must either be sub-terranean or covered with an atmosphere. What does the structure and composition of an exoplanet atmosphere tell about the likelihood of a biosphere being present?

14.2.1 Detecting and probing an atmosphere

The presence of an atmosphere can be inferred in several ways:

- ⋆ Via absorption lines appearing in the host star's spectrum, arising in the planet atmosphere as it passes in front of the star during a transit.
- ⋆ Via the differences occurring in the spectrum of the star+planet system as the planet is occulted by the star during secondary eclipse (thus removing the contribution of the planet to the spectrum). Absorption due to for instance water vapour would show up as the planet will radiate less efficiently at those wavelengths.
- ⋆ Via a spectrum obtained directly from the light of an imaged planet – this would look similar to the spectrum obtained via the planet occultation method.

The optical depth (see Equation 10.2) determines how deep one looks into the atmosphere. If the optical depth is zero, then an unobstructed view of the planet surface is gained. If the optical depth is significant, then one sees the part of the atmosphere located at a height above the surface where optical depth unity is reached. The temperature at that location can be very different from that at the surface. If it is colder, like in the troposphere, then that part of the atmosphere will radiate less intensely, and hence it will appear dark against the emission from the planet surface or lower regions in the atmosphere. If it is warmer, like it can happen above the tropopause (location of a temperature inversion), then that part of the atmosphere will appear brighter.

Hence, spectral features may appear in absorption or in emission. They tend to appear in absorption because the density is higher close to the planet surface and the optical depth often only reaches unity quite close to the surface, well within the troposphere. But for instance, ozone is present very close to the surface (smog) and high above the tropopause (the product of photo-synthesis) but not throughout the troposphere. Depending on the temperature profile above the tropopause, and the optical depth in the spectral regions adjacent to the ozone band, bands of ozone may appear in emission or in absorption.

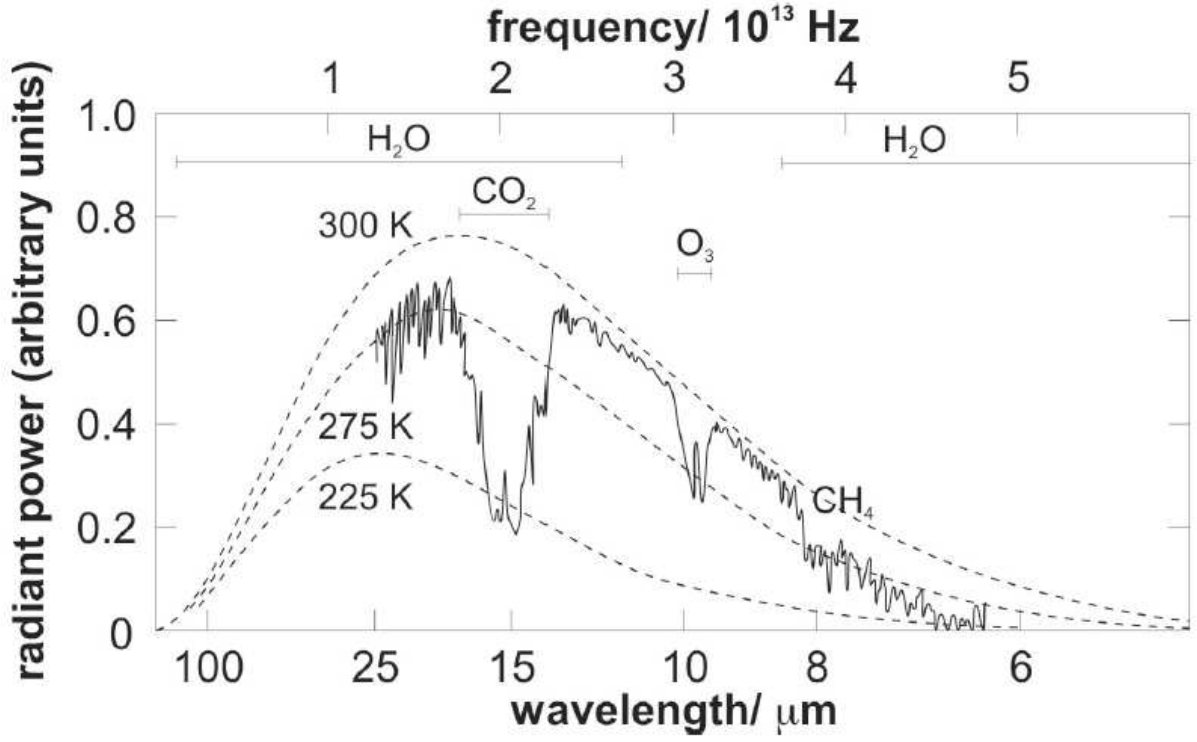

Figure 14.1: Earth's atmosphere seen from above. Credit: Barrie Jones, Open University.

The Earth's spectrum as viewed from above is displayed in Figure 14.1. It only shows the infrared portion of the spectrum, as this is where Earth shines most brightly as its temperature is $T \sim 288$ K. This is also where strong bands of water (depressing much of the spectrum), carbon–dioxide (around $\lambda \sim 15$ μm) and ozone (at $\lambda \sim 9.5$ μm) occur.

A few interesting observations can be made:

- ⋆ In between the absorption bands, Earth appears to have a temperature close to 300 K. This is higher than the 255 K estimated from the radiation balance with the solar irradiation of Earth, and a result of the Greenhouse effect (see Section 6.2.4).
- ⋆ The atmospheric layers seen in carbon–dioxide have a temperature of only ≈ 255 K. This is because carbon–dioxide is particularly good at absorbing radiation, and hence optical depth unity is reached well above the surface, but still within the troposphere where the temperature has dropped accordingly.
- ⋆ The ozone is seen in layers at a higher temperature, ≈ 275 K. This is because the ozone is also good at absorbing radiation, and optical depth unity is already

reached in the high-altitude layer, above the tropopause – where the temperature again increases with altitude – but still at a lower temperature than the Earth's surface. Note that the ozone band occurs in between water bands, which elsewhere in the spectrum leads us to observe the lower regions in the Earth's atmosphere rather than the Earth's surface.

14.2.2 Oxygen

Oxygen can have a biogenic origin, and indeed this is the main reason for it being a major constituent of Earth's atmosphere. Oxygen may be detected in several ways:

- ⋆ Oxygen molecules have absorption bands at optical wavelengths. However, these bands are narrow and the planet and its atmosphere are dim at optical wavelengths. To acquire a spectrum of an exoplanet with sufficient signal and spectral resolution to detect the oxygen bands is therefore extremely challenging.
- ⋆ Oxygen molecules in Earth's atmosphere are excited during day time. During night time, they relax to lower energy levels, radiating through many transitions at optical and ultraviolet wavelengths. This fluorescent light is detectable. However, it is dim compared to the reflected sunlight on the dayside of Earth, and it will therefore be extremely challenging to detect it on an exoplanet's nightside.
- ⋆ Ozone can be produced via photo-synthesis of oxygen. This makes ozone a proxy for the presence of oxygen. Ozone has a strong, fairly broad band at infrared wavelenghts. It would thus be relatively easy to detect in an exoplanet atmosphere.

14.2.3 Out-of-equilibrium chemistry as a biomarker

One argument for the presence of oxygen being a biomarker (a signature of a biosphere) is the fact that oxygen is efficiently removed from the atmosphere via oxidization of rocks (and via dissolution of carbonates, CO_3^{2-}, into the oceans). Hence the presence of significant amounts of oxygen in the atmosphere requires a mechanism through which to continuously resupply the atmosphere with fresh oxygen. This mechanism can be:

- ⋆ Biogenic, via photo-synthesis in green vegetation from carbon–dioxide and water.
- ⋆ Abiotic, via essentially the same process as in the biogenic case, but in the open.

How, then, can one distinguish between a biogenic and an abiotic origin of atmospheric oxygen? The key is again ozone. If the abiotic process operates in the atmosphere, then the subsequent photo-synthesis of ozone from oxygen occurs in a humid environment (as the production of oxygen requires water). But ozone is efficiently destroyed in the presence of water. Hence, if ozone is found, this would indicate the presence of atmospheric oxygen in a (relatively) dry environment and hence not produced in those atmospheric layers. The alternative explanation would be that (additional) oxygen was produced within organisms, so that enough oxygen would make it to the dry, high-altitude layers where ozone survives.

Hence in a biosphere, ozone and oxygen are found out of equilibrium with other chemical reactants. Out-of-equilibrium chemistry is generally a strong biomarker.

High concentrations of CO_2 produce very strong absorption bands across wide swaths of the infrared spectrum, obliterating any signature of ozone. Earth's atmosphere only contains about 0.03% CO_2 (much less than either Venus or Mars, perhaps because it is out of equilibrium?) and still its bands count among the most important contributors to the Greenhouse effect on Earth.

Carbon–monoxide (CO) is effectively removed from a biosphere, so detection of CO would indicate the *absence* of life.

14.2.4 Methane and ammonia

Methane also is a reactant to oxygen; they form a 'redox' (reduction–oxidization) pair, where the reducer accepts and the oxidizer donates an electron. Simultaneous detection of methane and oxygen (via its proxy, ozone) therefore strongly suggests a biosphere. Indeed, both are found in Earth's atmosphere (see Figure 14.1) – methane has a fairly strong band at $\lambda \sim 7.5$ μm.

In more exotic environments, the rôles usually performed by water may be performed by other agents. For instance, in cold environments this could be ammonia. Its bands at $\lambda \sim 9$ and 11 μm may be detectable in the spectra of exoplanets. It, too, reacts with other molecules (as on Titan, Section 7.4.3), and finding it out of chemical equilibrium would suggest that ammonia is generated biogenically.

14.3 Detecting vegetation

Green vegetation, through the aborbance and reflectance properties of chlorophyll, leaves a marked signature on the spectrum of a planet: the 'Vegetation Red Edge' (VRE). It can be considered a recognisable wavelength-dependent albedo. The VRE occurs around $\lambda \sim 700$ nm. Could this be detected from space?

The Moon reflects direct sunlight, but also sunlight scattered off Earth – 'Earthshine'. The brightness of the sunlit part of the Moon, with albedo $a_{\rm Moon}$, is:

$$I_{\rm Moonlight} \;=\; \psi \; I_{\odot} \;\times\; a_{\rm Moon}, \tag{14.1}$$

whilst the brightness of the Earthshine from the night side of the Moon is:

$$I_{\rm Earthshine} \;=\; \psi' \; I_{\odot} \;\times\; a_{\rm Moon} \; a_{\oplus}, \tag{14.2}$$

where ψ is a geometrical factor which depends on the fractions of illuminated Moon disc, and in the case of ψ' also on the distance between the Earth and the Moon. The spectra of Moonlight and Earthshine can be simply normalised to take away the wavelength-independent factor ψ'/ψ. Then, the Earth albedo is obtained from $I_{\rm Earthshine}/I_{\rm Moonlight}$.

The Earth's albedo of course contains signatures of the atmosphere as well as signatures of the surface material. Correcting for the atmospheric imprint, the VRE has indeed been detected in Earthshine measurements, but at a modest contrast level of only $\sim 7\%$.

Arnold, Bréon & Brewer (2009) have conducted computer simulations aided by calibration measurements of Earth from space, to investigate Earth's VRE signature today, during the Last Glacial Maximum (21 kyr ago), and during the warm inter-glacial Holocene Optimum (6 kyr ago). From this experiment one learns the following:

- ⋆ The VRE contrast has varied by a factor two between the Last Glacial Maximum and the Holocene Optimum. So, climate variations might be a factor to take into account when looking for the VRE signature on exoplanets, but it can hardly be a serious limit to our ability to detect it.
- ⋆ Large portions of Earth are not covered by green vegetation, but by oceans, ice caps, deserts, *et cetera*. This reduces the VRE contrast in a spectrum obtained of the entire Earth (from afar).
- ⋆ As a consequence, the viewing angle is important. If we were to view Earth from one of its poles, the VRE signature would be marginal at best.
- ⋆ Cloud cover masks the VRE signature.
- ⋆ Certain mineralogical materials also have albedo decrements around the position of the VRE, though generally not as sharp. An exoplanet spectrum should have sufficient resolution to distinguish between these.
- ⋆ Extra-terrestrial vegetation may have its VRE feature at a different wavelength.

An interesting possibility to confirm that a spectral shape that looks like the VRE indeed is the VRE, is through the expected circular polarization of the spectrum originating in biotic material: life on Earth has a strong sense of chirality (see Section 5.1.3), and light reflected off biotic molecules with a preferred isomeric orientation will thus become circular polarized.

14.4 Direct imaging of an exoplanet's surface

Direct imaging of exoplanets has become possible for gaseous giant planets in the outer regions of planetary systems. It is not unreasonable to expect that in future, this will also be accomplished for Earth-like planets at an au from a Solar-type star. If the planet can also be *resolved*, i.e. its disc spanning several angular resolution elements, then one may:

- ⋆ Detect the VRE signature as a function of position on the planet, maximising the sensitivity to regions where vegetation does flourish – and is not covered by a permanent cloud cover.
- ⋆ Map the scattering properties of the surface as a function of incident angle of starlight, to identify water surfaces.

This is possible, in principle, for an Earth at 3 pc distance, by means of an optical interferometre of ~ 200 mirrors of each 3m diameter and with a baseline (largest separation between optical elements) of ~ 200 km, situated in space.

Of course, any sub-terranean life would still go unnoticed until a manned or robotic spacecraft visited the planet and drilled through its surface.

Chapter 15

The Sun–Earth connection

The temperature at Earth is first and foremost determined by the luminosity of, and distance from the Sun. Besides radiation, the Sun also impacts upon Earth through a wind and through energetic particles as a result of magnetically driven activity.

15.1 The Earth's orbit and rotation

The temperature at the top of the Earth's atmosphere varies on timescales of millenia and longer due to several aspects of the Earth's orbit and rotation which vary as the other planets (primarily Jupiter) pull at it. These (multi-)periodic modulations are called 'Milanković cycles', after 20^{th}-century Serbian Milutin Milanković:

- ⋆ The eccentricity of Earth's orbit varies between $e \sim 0.005$ and $e \sim 0.05$ on timescales of 100 kyr. Although the annual-averaged distance to the Sun remains constant within $\sim 0.1\%$, the time spent closest to the Sun (around peri-helion) shortens and the time spent farthest from the Sun (around ap-helion) lengthens. The imprint of this on the Earth's temperature is apparent in climate records.

- ⋆ The Earth's rotation axis precesses as a result of torques exerted by the Sun and the Moon. This causes variation in the timing of seasons with respect to the position of Earth within its (slightly) eccentric orbit. This may be a factor in causing ice ages such as Earth experienced in the past tens of thousands years.

- ⋆ The Earth's orbit also precesses, adding to the variation in the timing of seasons.

- ⋆ The angle of Earth's rotation axis varies by $\sim \pm 1^\circ$ around a value $i \sim 23^\circ$. This change in obliquity has consequences for the seasons, especially near the poles where it affects the angle of incidence of the sunlight and the length of night and day.

Records of ice ages are based on isotopic ratios within polar water ice samples. The polar ice results from settling of water that originated at more tropical latitudes where it had evaporated. Lower temperature will disproportionally disfavour evaporation of the heavier isotopes, such as $H_2{}^{18}O$. The temperature record shows sharp rises to relatively brief interglacial periods.

15.2 The Sun's activity

Electrical currents within the hot plasma of the Sun's interior drive an essentially poloïdal, dipole magnetic field. However, the outer layers of the Sun rotate faster at the equator (once every 25 days) than near the poles (once every 35 days). This can be measured by following surface[1] features or by measuring the Doppler shift of spectral lines at different positions near the limb of the Sun's disc. Using helio-seismology techniques to probe the sound speed in these outer layers, we now even know that the equatorial surface layers rotate faster than equatorial layers deeper in the Sun. The differential rotation causes the magnetic field lines ('flux tubes') to wrap up into a toroïdal field where the field lines become increasingly parallel to the equator. The reason for this is that the magnetic field is "frozen" in the plasma due to the Lorentz force: plasma trying to move across field lines (as when they are aligned with the rotation axis) drags the field lines with it. The wrapping up of the flux tubes forces them to move towards the equator as well.

The magnetic field within the flux tubes adds magnetic pressure. To maintain pressure equilibrium with the surrounding plasma, the flux tube must expand to reduce the gas pressure within it. The resulting drop in density within the flux tube makes it buoyant and float towards the surface. When it reaches the surface, it can break through it. It thus forms a magnetic loop through the Sun's atmosphere, anchored to the surface in two spots with opposite magnetic polarity, namely where the magnetic field emerges and where it disappears into the Sun again. Because of the reduced gas pressure not only the density but also the temperature is reduced within the flux tube, to ~ 4000 K at the surface instead of ~ 5800 K elsewhere on the surface, causing Sunspots to look dark.

Sunspots were already seen in antiquity, for instance in the 4^{th} century by Theophrastus of Athens. The first known *record* of a Sunspot was not made until 1128, when John of Worcester made a rather cryptic drawing of two large, dark spots on the Sun. That these were indeed Sunspots has been confirmed by Korean records of the appearance five days later of a spectacular Aurora Borealis.

The process of migration of the magnetic field towards the equator, and its intensification as the magnetic field lines are increasingly concentrated, takes ~ 11 years on average (the Schwabe cycle). Once completed, a poloïdal field is re-established, but with opposite poles compared to the previous cycle. The cycle then repeats itself, but with the leading (in the sense of the Sun's rotation) Sunspots having opposite magnetic polarity compared to the previous cycle. Thus, a full cycle of solar magnetic activity really lasts ~ 22 years. The activity is measured by the number of Sunspots (which typically last a month), and this has been documented since the first use of the telescope in the early 17^{th} century. The location of the Sunspots on the Sun's disc is documented in the 'Butterfly diagram', so called because of the migration from intermediate latitudes towards the equator.

It has been possible in modern times to measure the total light output from the Sun accurately enough, to establish a very convincing direct correlation between magnetic activity and the Solar irradiance at Earth. The effect is less than 0.1%. Given also that it varies periodically on a timescale of a decade, any climatic impact of this is not

[1]The Sun is gaseous and as such does not have a surface. However, the optical depth generally increases very sharply from negligible to unity within a geometrically (or rather "heliometrically") thin layer, giving the appearance of a "physical" surface.

yet demonstrated. However, Sunspot numbers were very low during the entire second half of the 17$^{\rm th}$ century – the 'Maunder minimum' – a period which is also recorded in continental Europe as particularly cold. Indeed, the ultraviolet output of the Sun varies by 10%, which may affect the ozone and temperature inversion in the Earth's atmosphere.

15.3 Space weather

15.3.1 The solar wind

The Sun loses matter (at a slow rate of $\mathrm{d}M/\mathrm{d}t \sim 10^{-14}$ M$_\odot$ yr^{-1}) through a wind. Although the magnetic field lines at some distance from the Sun are essentially open, they are curved because of the coupling with the plasma in the wind and the conservation of angular momentum since the wind is launched at the rotating surface of the Sun. However, we will here approximate the outflow as purely radial and not affected by the magnetic field. The solar wind is driven by thermal pressure of the gas – the Parker model. It is described by a similar set of equations as used to describe molecular clouds (see Chapter 11), excluding for the time being the effects of electro–magnetic fields:

⋆ The continuity equation:

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}(r^2\rho v) \;=\; 0, \tag{15.1}$$

and hence:

$$r^2\rho v \;=\; \text{constant}, \tag{15.2}$$

where we realise that in spherical coordinates the radial component of the divergence $(\vec{\nabla}\cdot\vec{v})_r \rightarrow (1/r^2)\mathrm{d}(r^2v)/\mathrm{d}r$, and no mass is created or removed: $\partial\rho/\partial t = 0$.

⋆ The equation of state:

$$P \;=\; \frac{c_{\rm s}^2}{\gamma}\rho. \tag{15.3}$$

⋆ The momentum equation:

$$\rho v\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}(r^2v) \;=\; -\;\frac{\mathrm{d}P}{\mathrm{d}r} - \;\rho\frac{GM_\odot}{r^2}, \tag{15.4}$$

where we realise that in spherical coordinates the radial component of the gradient $(\vec{\nabla}v)_r \rightarrow (1/r^2)\mathrm{d}(r^2v)/\mathrm{d}r$, and we have assumed a steady flow: $\partial v/\partial t = 0$.

⋆ An "energy equation":

$$T \;=\; \text{constant}, \tag{15.5}$$

i.e. an isothermal wind, so the sound speed $c_{\rm s}$ is constant.

These equations have five different solutions, of which the Sun has "chosen" one in which the wind speed gradually increases from sub-sonic ($v < c_{\rm s}$) to super-sonic ($v > c_{\rm s}$) at a critical radial distance from the centre of the Sun:

$$r_{\rm c} \;=\; \gamma\frac{GM_\odot}{c_{\rm s}^2}. \tag{15.6}$$

The wind reaches a velocity of ~ 400 km s^{-1} at the distance of Earth.

15.3.2 Solar flares and Coronal Mass Ejections

The magnetic loops that stick out of the Sun are under considerable stress. When the opportunity arises, the magnetic field will reconfigure itself so as to relieve this stress. This occurs when magnetic field lines reconnect between themselves. Much of the magnetic energy that was stored in the initial configuration is thereby released, causing what is called a 'solar flare'. The free magnetic energy density is:

$$U = \frac{B^2}{2\mu_0}. \tag{15.7}$$

It is used to fling plasma that was initially confined to the flux tube, out to large distances and potentially into interplanetary space, by two means:

- ⋆ Radiation pressure. The energy flux of an electro–magnetic field is given by the Poynting flux:

  $$\vec{S} = \vec{E} \times \frac{\vec{B}}{\mu_0}, \tag{15.8}$$

  and the resulting radiation pressure is given by:

  $$P_{\rm rad} = \frac{\langle S \rangle}{c}. \tag{15.9}$$

  It is used to accelerate particles. These will then have a non-Maxwellian velocity distribution as it is no longer determined by equipartition of energy during collisions.

- ⋆ Ohmic dissipation. With the movement of charged particles a current is associated. When a current flows through a resistance, electric energy is converted into heat. This is usually not in itself capable of flinging plasma into interplanetary space, but it helps as the gas pressure in the heated gas will expand it and create buoyancy.

Ejections of matter from the hot extended atmosphere of the Sun, the corona, are called 'Coronal Mass Ejections' (CMEs). These are often triggered by very energetic solar flares, but sometimes they appear to be a purely coronal phenomenon without obvious causal link to activity on the Sun's surface.

15.3.3 Geo-magnetic phenomena

CMEs can hit the Earth. They are common and large, so this happens regularly. Luckily, the Earth's surface is protected from adverse effects by the Earth's own magnetic field. This field is probably generated in the liquid iron core, by a dynamo mechanism similar to that which operates in the Sun's interior. The field is reasonably approximated by a dipole, with an axis that is somewhat inclined (and offset) from the Earth's rotation axis. Charged particles from the Sun's wind and outbursts are trapped by the Earth's magnetic field in two bands – the 'Van Allen belts' – wrapped around the Earth's waist: an inner belt of protons at a few 10^3 km above the Earth, and an outer belt of electrons at a few 10^4 km (overlapping with the orbits of geo-stationary satellites).

When a CME hits the Earth's magnetosphere it can lead to the following phenomena:

- ⋆ An aurora. The charged particles of a CME are forced to follow magnetic field lines, which guides them towards Earth's magnetic poles. They impact upon high atmospheric layers before reaching the ground, creating a ring of fluorescent light. The characteristic red and green colours are due mainly to molecular oxygen (at 558 nm and ~ 630 nm); molecular nitrogen adds more red as well as violet.
- ⋆ A geo-magnetic storm. High atmospheric layers of charged particles, the ionosphere is used to bounce off electromagnetic waves so as to communicate across the globe. CMEs move faster than the solar wind. This results in a shock front preceding it. When the shock front hits the Earth's magnetosphere it can seriously disturb the ionosphere, upsetting radio communication. The 'Carrington event' in 1859 induced strong magnetic activity and auroræ visible from sub-tropical latitudes; more recently, the 'Halloween storms' caused wide-spread havoc in 2003.

The stand-off distance between the solar wind and Earth's magnetic field is determined by the balance between the ram pressure by the solar wind and Earth's magnetic pressure:

$$\frac{B_\oplus^2}{2\mu_0} \;=\; \rho_{\rm sw} v_{\rm sw}^2. \tag{15.10}$$

Earth's magnetic field can be approximately described as a dipole:

$$B_\oplus(r) \;=\; \frac{B_\oplus(\mathrm{R}_\oplus)}{r^3}. \tag{15.11}$$

Hence the stand-off distance is typically $r_{\rm so} \sim 10\ \mathrm{R}_\oplus$, following from:

$$r_{\rm so} \;=\; \left(\frac{B_\oplus^2(\mathrm{R}_\oplus)}{2\mu_0\rho v^2}\right)^{1/6}. \tag{15.12}$$

The Earth rotates as a solid body, so not the same cycles of activity occur as they do in the Sun. But the geomagnetic field is far from static. The magnetic pole wanders around by many degrees latitude on a timescale of centuries. Records of the magnetic field are preserved in magnetised rocks. They tell a story of frequent reversals of the magnetic field, not unlike what happens in the Sun. The magnetic field strength undergoes vast excursions, and it may have been absent during some reversals. That could have exposed Earth to the direct impact from CMEs. But the magnetic field dragged along by the solar wind might envelop the Earth and induce a protective magnetosphere around it instead.

15.3.4 Cosmic rays

The most energetic particles, cosmic rays can penetrate deep into Earth's atmosphere before colliding with molecules and thus giving rise to showers of secondary particles that do reach the ground. These secondary particles often still carry sufficient energy to cause harm to cells in organisms. Such harm may in fact be required for evolution of life as more successful mutants are selected to prevail preferentially over the original species.

Mars has such tenuous atmosphere that its surface is essentially sterilised by the primary cosmic rays. Venus lacks protection from a magnetosphere, but its atmosphere is so dense that not even secondary cosmic rays reach the ground.

There are two main sources of cosmic rays. In both cases, but in different ways, the rate at which they impact upon Earth is influenced by the Sun:

- ⋆ Cosmic rays accompany CMEs. Thus, sudden bursts of cosmic rays can hit the Earth when it is hit by a CME. This happens more often at the height of solar activity, but in between bursts the solar cosmic ray flux is rather modest.
- ⋆ Cosmic rays are also entering the Solar System from outside. There are probably multiple sources for these truly cosmic rays, amongst which have been identified supernova remnants and other places where magnetic fields are compressed making them effective particle accelerators. The solar wind makes it harder for these cosmic rays to reach the Earth. Hence, at solar maximum when the solar wind is also stronger, the cosmic ray hit rate on Earth is *lower* than at solar minimum.

The neutron flux arising in secondary cosmic ray showers leads to high levels of radioactive isotopes such as ^{14}C (half life of 5730 yr) and ^{10}Be (half life of 1.5 Myr). Indeed, the neutron flux measured in modern times correlates well with ^{14}C and ^{10}Be concentrations recorded in organic matter. The latter can thus be used as a proxy for solar activity, tracing it back to at least the Last Glacial Maximum. The post-glacial warming during the Holocene is mirrored in high levels of solar activity. Curiously, along with high levels of solar activity over the past century, today's geomagnetic field strength is rather weak.

15.4 The Sun's evolution affects the Habitable Zone

From the moment it started life as a hydrogen-burning star, the Sun has evolved as a consequence of changes in its fuel supply and internal structure. In the past, the Sun almost certainly rotated more rapidly than today, and its magnetic field was stronger. The more active Sun may have created too hostile an environment on Earth for life to develop in the first few hundred Myr, as Earth would have been exposed to higher levels of cosmic rays and ultraviolet radiation (there was no ozone layer then). Or perhaps it played a more positive rôle in creating precisely the right conditions for life to emerge. The Sun may also have been fainter than it is now, and Earth might have been glaciated all over some 2 Gyr ago (around the time of the Great Oxidization Event).

In another ~ 5 billion years, the Sun will become a red giant star, increasing in size and luminosity on a timescale of millions of years. It will become about as large as the orbit of Earth. Clearly, the Habitable Zone will have moved outwards. In the final phases, alternative mechanisms take effect which drive a vastly enhanced solar wind, making the Sun lose mass at a rate $\mathrm{d}M/\mathrm{d}t \sim 10^{-6}\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$. Apart from the devastating ram-pressure effects on the Earth's atmosphere, it also eventually halves the mass of the Sun. As a result of the reduced gravity, the Earth will move away from the Sun – but not far enough to catch up with the receding Habitable Zone.

The Sun is already increasing in luminosity, very slowly. With the Earth being located close to the inner edge of the Habitable Zone, it is quite possible that the Habitable Zone will have left the Earth behind well before the Sun becomes a red giant. Mars may then become a more attractive planet to settle on, especially if we identify a sustainable mechanism for increasing its atmospheric pressure and for creating an ozone layer.

Chapter 16

Natural and human hazards to life

Earth endured extended periods of severe glaciation 600–700 Myr ago, demoting the more recent ice ages as merely chilly moments. Immediately following the last global glaciation, multi-cellular life burst on the scene in an explosion of diversity[1]. Ever since, there have been frequent periods of mass extinction, wiping out entire families of species. In between, species have become extinct continuously. In fact, recent extinction rates have been at their lowest for over half a billion years – but humankind has started to reverse this trend. While this all sounds bad for the fate of individual species, regular disasters speed up evolution, and may be essential to avoid stagnation in the development of life.

Interestingly, a global glaciation period ~ 2.2 Gyr ago might have been triggered *by* the Great Oxidization Event. At a time when the main Greenhouse gas was methane, this would have been removed from the atmosphere by the oxygen that was being produced. Perhaps the more recent global glaciations were hickups triggered by biological evolution?

What causes mass extinctions?

- ⋆ Milanković cycles? Occasionally all orbital and rotational effects could "align", resulting in a greatly amplified impact on Earth's temperature, perhaps just enough to set in motion dramatic climatic changes.
- ⋆ Solar activity variations? It is not clear how much more inactive the Sun could possibly have been, compared to the Maunder minimum or the level that may have caused the Last Glacial Maximum, for instance. Whether *high* levels of solar activity could have led to climatic changes that induced mass extinctions is equally unclear.
- ⋆ Volcanism? The last major mass extinction, that wiped out the dinosaurs 65 Myr ago, was accompanied by a period of intense volcanic activity on the Indian subcontinent. Humankind may have narrowly survived the eruption of the Toba supervolcano on Sumatra 74,000 years ago, having passed through a genetic bottleneck.
- ⋆ Impacts?
- ⋆ Supernovæ?
- ⋆ Disease, competition, over-specialization? To wipe out large numbers of species they would have to have led to an avalanche of death amongst inter-dependent species.

[1] Our own exit from the last ice age allowed a sedentary lifestyle and the birth of complex civilization.

16.1 Are humans causing the next mass extinction?

Humans affect their environment in many ways, at the expense of others. If some species perish, this can have knock-on effects on other species and upset the entire ecological system. Pollution of rivers and oceans (microplastics!), deforestation, war, *et cetera* all impact on regional ecosystems. Perhaps the most global, dramatic and imminent danger from human activity is through reduced protection against ultraviolet radiation we receive from the high atmospheric ozone layer, and through changes to the Greenhouse effect which regulates the atmospheric temperature and resulting weather patterns.

16.1.1 The ozone hole

When a large hole was discovered in the high atmospheric ozone layer above Antarctica in 1985, it became clear that the protection against harmful ultraviolet radiation can not be taken for granted. When the hole was seen to increase in size from year to year, alarm bells started ringing. The main chemical processes which set the ozone concentrations in the cold, polar atmospheric layers were already known in the 1970s, as was the fact that certain chemical compounds used in industry and domestic appliances (such as refrigerators) end up in those layers where they would catalyse the destruction of ozone. The most infamous of these, chloro–fluoro–carbons (CFCs) have since been banned. Although the stratosphere takes many decades to recover, the ozone hole has in fact started to shrink.

The Antarctic ozone hole is a seasonal effect, appearing during September–November. The dramatic variations, apparent fragility, critical importance of ultraviolet protection, and demonstrable human impact have served as a wake-up call for humankind. But other processes, too, can alter the stratospheric ozone concentration. Volcanism, for instance, can cause a drop in stratospheric temperature which leads to a reduction in ozone.

16.1.2 Carbon–dioxide and the Greenhouse effect

One of the most important Greenhouse gases is carbon–dioxide, because it has such strong absorption bands in many parts of the infrared spectrum where the Earth shines most brightly. This is despite its low concentration of only $\sim 0.04\%$. There is a clear measured correlation between the atmospheric CO_2 concentration and the global average (annual) temperature, both in modern times as well as in ancient fossil records and ice samples.

There are several major sources of atmospheric CO_2:

- ⋆ Exhalation by organisms using respiration as their main source of energy intake. This is however a minor contributor.
- ⋆ Volcanism. Peaks in volcanic activity are known to have caused periods of greatly increased atmospheric CO_2 content.
- ⋆ Exhaust of CO_2 into the atmosphere by industry, transport and domestic activities using fossil carbon-based fuel has caused a 40% increase of the atmospheric CO_2 content since the start of the Industrial Revolution some two centuries ago.
- ⋆ Indirect human impact. Deforestation, draining of wetlands, and agriculture, for example, lead to the release of CO_2 that was stored in organic matter and soil.

But there are also major sinks (sequestration) of CO_2:

- ⋆ Photo-synthesizing organisms, such as plants and some types of bacteria (which populate oceanic surface layers), take in CO_2 to produce energy and grow. When these organisms die, carbon is stored in soil and sediment.
- ⋆ Oxidization of rocks binds atmospheric CO_2 into carbonates (compounds involving CO_3^{2-}). Erosion causes carbonates to end up in oceans where it settles in sediments.
- ⋆ CO_2 can dissolve directly in oceanic surface layers, where it enhances the levels of acidity by formation of carbonates and protons[2]. This causes the death of coral reefs by bleaching. However, warmer seawater *less* readily dissolves CO_2.

Other gases contribute to the Greenhouse effect, notably water vapour (its main agent) and methane (one of its most potent). A stronger Greenhouse effect causes warmer weather and thus increased evaporation of surface water into the atmosphere but also desertification. Hence it is non-trivial to predict the result: some consequences will re-enforce the change whereas others will undo it. This is called 'feedback'. There may be thresholds, which when passed give way to an unconstrained 'runaway' process.

16.2 Circulation patterns and weather

Ultimately, the atmosphere is a heat engine: the flow of warm air drives weather. If heat Q is added to the atmosphere at a temperature T_1, and given off at a lower temperature T_2, then the efficiency of the atmosphere in driving weather (performing work) is:

$$\eta \;=\; \frac{\text{work}}{Q} \;=\; \frac{T_1 - T_2}{T_1}. \tag{16.1}$$

Wind is driven by pressure gradients, often due to temperature differences – by directly moving air, or by suction into a low-pressure region resulting from rising warm air:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \;=\; -\,\frac{1}{\rho}\vec{\nabla}P. \tag{16.2}$$

In oceans, temperature differences drive diffusion of warm water towards colder areas. Above 4 °C, cold water is denser than warmer water, and thus sinks. Even colder water becomes less dense – this is a peculiar property of water, exemplified most dramatically in the solid form (ice) being lighter than the liquid form. But salinity also affects the density of water. Differences in salinity occur where fresh water reaches the sea, or where the deep ocean meets shallower, warm waters where evaporation increases the salinity. The temperature and salinity act together to drive 'thermo-haline' oceans currents.

In the interface between the water surface and the air, wind induces waves and currents.

The stratification of the oceans into layers can be beneficial to species as it creates stable, homogenous environments, however ecosystems also benefit from mixing these layers as

[2]Seawater is slightly basic, and the acidification makes it more neutral, not strictly speaking acidic.

it moves nutrients and oxygen to deprived areas. Circulation results from upwelling due to currents diverging following gradients, for instance in topography.

The rotation of the Earth causes airflows or water currents to deviate from their original path – as seen by an observer moving along with the Earth's rotation. This is a result of fictitious forces appearing when transforming between inertial and accelerating restframes. With respect to an observer in space, the Earth rotates with angular frequency Ω. In order to maintain its curved path circling the Earth's rotation axis, a fluid element (could be an air pocket) travelling with a speed v in the same direction undergoes an acceleration

$$a_{\text{inertial}} = \frac{(\Omega R + v)^2}{R}, \tag{16.3}$$

where R is the (shortest) distance to the Earth's rotation axis. For an observer on Earth, in the rotating restframe, the fluid element appears to have a velocity v only, and is thus in need of an acceleration

$$a_{\text{rotating}} = \frac{v^2}{R}. \tag{16.4}$$

Hence, fictitious accelerations have cropped up that must be added to make up the difference with respect to the inertial restframe acceleration:

$$a_{\text{fictitious}} = \Omega^2 R + 2\Omega v. \tag{16.5}$$

The first of these terms, $\Omega^2 R$ is the acceleration towards the Earth's rotation axis needed to follow Earth's rotation; in vector notation:

$$\vec{a}_{\text{centripetal}} = \left(\vec{\Omega} \times \vec{R}\right) \times \vec{\Omega}, \tag{16.6}$$

where $\vec{\Omega} \times \vec{R}$ is the linear velocity of the Earth's surface – we do not call it $\vec{v}$ here so as not to cause confusion with the velocity *relative* to Earth's surface. The second term, $2\Omega v$ (in the rotating restframe) is called the Coriolis acceleration; in vector notation:

$$\vec{a}_{\text{Coriolis}} = 2\vec{v} \times \vec{\Omega}. \tag{16.7}$$

Both the centripetal and Coriolis accelerations force the fluid element to deviate towards the equator. When flowing meridionally (along lines of constant geographic longitude), movement towards or away from the Earth's axis results in a change of angular velocity as the angular momentum must be conserved. The resulting deviation is in the same sense with respect to the original path as the centripetal and Coriolis accelerations: towards the right on the Northern hemisphere and towards the left on the Southern hemisphere. As a consequence, circular motions across the Earth's surface are made.

It then seems odd that cyclones – heavy tropical storms surrounding regions of extreme low pressure – rotate in the opposite sense of what is just described. The reason is that in cyclones, air streams towards the centre where the pressure is lowest. It deviates from a radial path by the centripetal and Coriolis accelerations and angular momentum

conservation, making it initially miss the centre of the cyclone and swirl around it in the opposite sense (before reaching it and being transported upwards into the atmosphere).

Jet streams at > 10 km altitude do not show the circular motion of cyclones, as in this case pressure gradients balance the Coriolis force sufficiently to withstand deflection. However, the decreasing temperature gradient between polar and equatorial regions weakens this constraint, allowing the jet stream to meander more violently (a form of a 'Rossby wave'), thus causing more extreme variations in weather – particularly around 30–60° latitude.

Warm moist air from around the equator rises and moves poleward before sinking again around 30° latitude (the 'Hadley cell'). Air then moving North will deviate in an Easterly direction, whilst air moving South will go Westwards. This explains the prevailing (but not exclusive) motion of wind and weather, with tropical storms (initially) moving Westwards but weather affecting the United Kingdom generally *coming* from the West.

In atmospheres of rapidly rotating planets, the angular momentum conservation prevents latitudinal transport. This may cause very cold polar regions and a hot equatorial zone.

A big uncertainty in weather predictions and climate models is the rôle and especially the formation of clouds. Cloud formation relies on the presence of nucleation seeds. These can be in the form of small dust particles or even large molecules – 'ærosols'. Super-saturated vapour then condenses onto these seeds to form ice crystals or water droplets depending on the temperature of the cloud. Because thermal motions disturb the condensation process, it is simpler to model the formation of high-altitude cirrus clouds than of rain clouds.

16.3 Ecosystems

Ecosystems are complex systems in which multiple components have causal relationships. These tend to maintain an equilibrium, with some stability against small perturbations. However, they can be brought out of equilibrium and it is important to understand how the ecosystem would respond.

16.3.1 Feedback

Many ecological processes, for instance those that regulate the atmospheric composition, work in both directions and depend upon each other. This creates a complex network of processes, where the rate and direction at which a given process operates depends upon the outcome of another process – or even upon itself.

Examples of simple mathematical descriptions of feedback include:

⋆ Exponential growth or decay:

$$\frac{\mathrm{d}N}{\mathrm{d}t} \;=\; \alpha N \;\rightarrow\; N(t) \;=\; N(0)\exp(\alpha t). \tag{16.8}$$

Depending on the sign of α, N either increases or decreases exponentially.

⋆ A harmonic oscillator:

$$\frac{\mathrm{d}^2N}{\mathrm{d}t^2} \;=\; -\alpha^2 N \;\rightarrow\; N(t) \;=\; N(0)\cos(\alpha t). \tag{16.9}$$

Note the slight difference in the differential equation with respect to the previous example, where $\mathrm{d}^2N/\mathrm{d}t^2 = +\alpha^2 N$. The solutions are very different, though.

An example of positive feedback is the melting of ice caps. As the ice melts, the albedo of the Earth's surface lowers. A darker Earth warms up, and thus radiates more energy into the atmosphere, which therefore warms up as well. As a consequence, the ice caps would melt even faster. Likewise, vast quantities of methane, locked up in 'clathrate' cages in permafrost and the ocean floor, may be released into the atmosphere due to Global Warming, thus intensifying it.

An example of negative feedback would be the formation of clouds. As the ocean warms, evaporation increases the water vapour content of the atmosphere. This, in principle, facilitates the formation of clouds. Clouds reflect sunlight before it reaches the Earth's surface, thus increasing its albedo. As a consequence, the Earth cools.

Feedback relations usually have a limited range of operation. If pushed (by a process external to those relations) beyond a critical threshold, some of the mechanisms may become ineffective or fail altogether, whilst other mechanisms might emerge. Any of these could provide positive or negative feedback, leading to a runaway situation or a new equilibrium. One example is the 'moist greenhouse threshold': once water becomes abundant in the *stratosphere*, it dissociates under the influence of ultraviolet light. The hydrogen escapes into space and thus becomes unavailable to reform water. As a result, water is lost. This may have happened to Venus.

16.3.2 Carrying capacity

If one species' population grows unexpectedly rapidly it will put pressure on the ecosystem to support it, until it reaches the ecosystem's 'carrying capacity' and the species finds itself in a pickle. This can happen when natural predators become ineffective (by elimination or defense) or when it exhausts finite resources (e.g., fossil fuels and rare Earth elements in our case).

As with other feedback processes, many of these dynamical relationships can be written down in the form of differential equations. For instance, the Lotka–Volterra equations describe a predator–prey coupled system model:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = +\alpha x - \beta xy \tag{16.10}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\gamma y + \delta xy \tag{16.11}$$

The population of prey, x increases by birth at a rate set by α; that of the predator, y decreases by starvation at a rate set by γ. When they meet (xy term), this leads to a certain rate of death among prey set by β, and a certain growth rate among predators set by δ. The solutions to these equations are either zero for both, or oscillatory with a 90° phase shift (can you guess in what sense?). Predator–prey co-evolution can accelerate evolution and make both more successful.

Chapter 17

Threats from outer space

Aside from alien attack, and the delicate Sun–Earth connection discussed previously, the Universe is not always kind to life. Most of the threats from outer space are rather haphazard, and perhaps we have been lucky... so far.

17.1 Impacts

In the first few hundred Myr, the Solar System was riddled with the debris left over from the planet formation process. Frequently, this debris would collide with the planets and moons. This bombardment has left its imprint in the form of craters on the faces of all planetary bodies that do not have (or had) atmospheres or tectonic activity which erode such features.

Debris is still around today, and comets enter the realm of the planets when the slightest disturbance deviates them from their normal residence at thousands of au from the Sun. Evidence for recent impacts include:

- ⋆ In 1994, comet Shoemaker–Levy 9 crashed into Jupiter. Being massive, Jupiter is a great attractor for minor bodies – one of the reasons for it to have so many moons orbiting it. It is also big and thus more likely to be hit. In fact, comet Shoemaker–Levy 9 had already passed very close to Jupiter on an earlier occasion. As a result of gravitational tides, it had broken up into more than twenty fragments. Like wagons on a train, each of these fragments one after the other crashed into the atmospheric "surface" layers of the rapidly rotating planet (Jupiter rotates in less than 10 hours), leaving a trail of crater-like patches. The strong atmospheric circulation washed out these features within months.

- ⋆ In 1908, Earth was hit by an object perhaps $\sim$ 100 m across. It happened in a remote region in Siberia (Tunguska), and no-one was hurt. Without very specific witness reports, it has been difficult to establish exactly what happened. Recent computer simulations suggest that a porous meteoroid or a fragment of a comet vaporised as a result of friction with the atmosphere, before actually reaching the ground. Hence there is no crater. The expanding super-heated air created a shock-wave which did impact upon Earth, causing wide-spread devastation in the heavily forested landscape.

A similar, smaller event happened in 2013 when a 10-m sized object detonated above the million-people city of Chelyabinsk, also in Siberia. The explosion had a strength of more than twenty times the atomic bomb that destroyed Hiroshima, but most of the energy did not make it to Earth's surface. The blast wave that did, though, caused significant damage and injury (but no fatalities).

⋆ The Barringer crater in Arizona is prototypical for the impact of a large meteoroid (~ 50 m diameter). Fragments of it have been recovered. The arid environment has well-preserved the crater for the 50,000 years since it was created.

⋆ The extinction of the dinosaurs 65 Myr ago has been speculated to have been caused by an impact orders of magnitude more energetic than that which happened in Tunguska or Arizona. One early hint at this scenario was derived from the presence of a thin layer of iridium covering much of the Earth, in layers that could be dated to around the same time, 65 Myr ago. Iridium is not uncommon in the Universe, but on Earth it has sunken towards the core. The surface layer of iridium was therefore interpreted as evidence for a large impact, of which the debris was scattered high into the atmosphere where it could be spread around the globe and slowly precipitate onto the ground. In the 1960s the "smoking gun" was discovered: the Chicxulub crater, 180 km in diameter, overlapping the coast of the Yucatán peninsula and the Gulf of México. It was discovered on very accurate maps of the Earth's gravity field, and is accompanied by quartz from the impact and soot from burnt vegetation. The object that hit Earth must have measured ~ 10 km in diameter.

⋆ A large impact crater in Western Australia has been found to be ~ 2.2 Gyr old. It is speculated that it might have ended global glaciation – could it also have initiated the Great Oxidization event? But see page 101; it is crucial that we improve the accuracy of the relative timings of these events.

So there are two immediate effects that should be considered when assessing the "impact" of a collision. Firstly, the frictional heating of the Earth's atmosphere, which is determined by the ram pressure (Equations 3.9 & 10.6) and depends sensitively on the (relative) velocity and the cross-sectional area of the object. Head-on collisions will be faster, but more compact objects will be smaller. Associated *internal* heating of decametre-sized objects causes rapid expansion of volatile material which can lead to detonation (larger objects take too long to heat all-through). Secondly, an actual impact on the surface, causing a crater, which is determined by the kinetic energy with the same dependence on relative velocity as the frictional heating and dependent on the overall mass of the object. It also depends on the angle of incidence, which – through the conservation of momentum and force balance – determines the response of the Earth's surface and sub-surface material to the impact.

Improved detection methods have resulted in a catalogue of asteroids that can come close to Earth, the 'Near Earth Objects' (NEOs). The second-largest of these is Eros, which was discovered in 1898. Its ellipsoïdal size is $\sim 34 \times 11 \times 11$ km^3. The vast majority of NEOs are smaller than a km across. Depending on where they land, these can wipe out cities and kill millions of people. By coincidence, on the day of the Chelyabinsk event all eyes were on Duende, a 30-m NEO approaching to within 4.3 $R_\oplus$ from Earth's surface.

Indeed, some asteroids become temporary moons of the Earth, 1% of which are estimated to crash onto Earth (at speeds below the escape velocity).

Predicting the locations of NEOs is hampered by the following:

- ⋆ There is no analytical solution for the trajectory of an object in the presence of more than one massive body. Already Jupiter exerts a noticeable gravitational pull.
- ⋆ On comets, localised escape of volatiles causes a rocket effect which pushes the comet in the opposite direction.
- ⋆ Another type of non-gravitational effect is the Yarkovsky effect. Minor bodies rotate. As the surface which had been warmed by sunlight moves into the shadow, it cools by radiating. This thermal emission carries momentum, so like the escape of gas this too creates propulsion in the opposite direction.

Likewise, shooting it to smithereens before collision may avoid a concentrated impact, but the fragments will have unpredictable trajectories and could potentially do more harm than the original threat. However, a small nudge is enough to change the course of an object enough to avoid collision, and such measured deflection technique is deemed to be more successful, predictable, as well as energetically less demanding.

With the discovery of a $\sim$ 100-m sized asteroid in 2017 and $\sim$ 2-km sized comet in 2019, both coming from *outside* the Solar System, the question is which ones to fear most. Recent, crude guesses suggest such nemeses might hit Earth once every 100 million years.

17.2 Supernovæ

Massive stars exhaust their fuel before they shed their mantles, and rather than simply leaving behind their bare cores (white dwarves), cooling and fainting into oblivion, they collapse under the weight of their mantles and depart in spectacular fashion as infalling material bounces off and blasts into space. Accompanied by the light of a billion Suns, 'Supernovæ' can be extremely dangerous if going off in your vicinity. Luckily, they happen only a few times per century in the entire Milky Way galaxy, which is vast (the Sun is 8 kpc away from its centre). The region affected by supernovæ is typically a hundred pc, much less if entrained within dense molecular clouds.

There exists a very different mechanism that can lead to a supernova, namely the thermonuclear detonation of a white dwarf, or the coalescence of two white dwarves. Because these so-called Type Ia events happen a long time after the death of a long-lived star, they are not limited to the immediate neighbourhoods of their birth sites.

We know of a handful examples of both types of supernovæ which have gone off within several kpc of the Sun in the last millenium.

Damage done by supernovæ may be realised in several guises:

- ⋆ The light, in particular the energetic portion of ultraviolet light, X-rays and γ-rays, can heat up and sterilise planets in its vicinity. The 'flash-ionization' of the atmosphere, on the one hand, causes the production of nitrogen–oxide (NO_2) which absorbs near-ultraviolet light (315–400 nm) but, on the other hand, depletes ozone.

- ⋆ The ejecta of supernovæ, initially traveling with $\sim 10,000$ km s^{-1}, can compress the Sun's influence sphere so as to expose the Earth directly to anything coming from interstellar space. It could even damage directly the Earth's atmosphere, by ram pressure. The ejecta slow down as they sweep up interstellar matter. They typically stall by the time they have swept up a similar mass in ISM as originally within the ejecta (of order 10 $M_\odot$) and/or dissipate their kinetic energy by cooling.
- ⋆ Before being engulfed by the ejecta, anything in their path is first hit by a shock, as a result of the super-sonic motion of the ejecta. This shock also compresses and thereby intensifies the interstellar magnetic field. It thus becomes a powerful accelerator of charged particles, and these become the cosmic rays that continuously hit the Earth's atmosphere. The cosmic rays from a nearby supernova can destroy DNA, ozone or lead to cloud formation resulting in global cooling.

The odds are, that we have escaped the flash-ionization effects of a supernova for billions of years. Among the most powerful supernovæ, though, are those that are non-isotropic. Instead, these Gamma-Ray Bursts (GRBs) are beamed, and while this makes it likely to miss us, if it does hit it will be much more intense. GRBs as far away from us as the other side of the Milky Way could do very significant harm – perhaps once every billion years.

Finally, the mild effects from supernovæ could be beneficial, causing genetic diversity.

17.2.1 The local interstellar environment

The solar wind offers protection against the hostility of interstellar space. Its influence sphere, the 'heliosphere' reaches out to ~ 100 au, beyond which the ISM increasingly manifests itself. The stand-off boundary, the 'heliopause', is accompanied by a mild bow-shock, as the Sun moves slightly super-sonically with respect to the local ISM.

Initially, the Sun moved in a nearly circular orbit around the centre of the Milky Way galaxy along with the molecular cloud from which it had formed. Over billions of years, though, even fairly remote encounters with other stars and molecular clouds incrementally deflected the Sun from its original path. This is called 'dynamical heating', and it causes the Sun to have a 'peculiar' motion with respect to the molecular clouds, both within the thin (~ 100 pc) disc of the Milky Way as well as in a vertical direction perpendicular to the disc. The Sun now moves with a velocity ~ 20 km s^{-1} with respect to the local ISM. One revolution around the Milky Way centre takes around 250 Myr, at a tenfold speed. Thus the Sun travels ~ 1 kpc through the ISM per each 10 Myr.

Currently, the star closest to us, Proxima Centauri is still 1.3 pc away from us. But at times, a star will have come closer to the Solar System, and if it approaches to well within a parsec it might disturb the Oort Cloud where most icy minor bodies are loosely gravitationally bound to the Sun. This may sent more comets than usual on a path towards Earth, increasing the chances of a direct hit. But we will not be hit by a star.

Not surprisingly, the Solar System currently finds itself in a warm, tenuous portion of the ISM. In fact, it traverses the 'Local Bubble' of a few hundred pc across. It was most likely created by a few supernovæ. On a much longer timescale of dozens of Myr, the Sun moves in and out of spiral arms, which is where most molecular clouds are concentrated and most supernovæ go off.

17.2.2 A Galactic Habitable Zone?

Diversity of the species as deduced from fossil records correlates with the typical timescales associated with the motion of the Sun through the Milky Way. The likely reason for such correlation is that supernovaæ are more likely to occur close to Earth when it passes through a molecular cloud, or spiral arms in general. It may be the vertical motion of the Sun with respect to the Milky Way disc which is most important. The equation of motion in the vertical direction is given by the prescription for its acceleration under the influence of the gravitational attraction by the Milky Way disc (where the density is highest):

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} \;=\; -\,\frac{\mathrm{d}\Phi}{\mathrm{d}z}. \tag{17.1}$$

Solving this equation is simplified by performing a Taylor expansion of $\Phi(z)$ around $z = 0$:

$$\Phi \sim \Phi(0) + z\left|\frac{\mathrm{d}\Phi}{\mathrm{d}z}\right|_{z=0} + \frac{z^2}{2}\left|\frac{\mathrm{d}^2\Phi}{\mathrm{d}z^2}\right|_{z=0} + ... \tag{17.2}$$

Ignoring quadratic terms and realising the symmetry around the mid-plane:

$$\left|\frac{\mathrm{d}\Phi}{\mathrm{d}z}\right|_{z=0} \;=\; 0, \tag{17.3}$$

hence we obtain:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} \;=\; -\left|\frac{\mathrm{d}^2\Phi}{\mathrm{d}z^2}\right|_{z=0} z \;\equiv\; -\omega^2 z. \tag{17.4}$$

The solution of this equation is that of a harmonic oscillator:

$$z \;=\; Z\;\cos(\omega t + \theta). \tag{17.5}$$

Using the Poisson equation to relate the gravitational potential to the mass density that causes it, we obtain:

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}z^2} \;=\; 4\pi G\rho(z), \tag{17.6}$$

and thus:

$$\omega \;=\; \sqrt{4\pi G\rho(0)} \;\equiv\; \frac{2\pi}{P}. \tag{17.7}$$

A rough estimate of the mass density in the Solar Neighbourhood would be 1 $\mathrm{M}_\odot\ \mathrm{pc}^{-3}$. This yields $P \sim 28$ Myr, which is remarkably close to the ~ 30 Myr periodicity seen in the species' diversity.

To take matters further, the increased rate of supernovæ in the inner portions of the Milky Way galaxy may imply a Galactic Habitable Zone at some distance from the Milky Way centre. We may well find ourselves orbiting the Milky Way centre at a safe distance of 8 kpc for good reason.

Chapter 18

Travel into space

In 1957, Sputnik 1 was the first man-made object reaching space, orbiting the Earth without the need for further propulsion. In the same year, the dog Laika was the first animal to orbit Earth, in Sputnik 2, followed in 1961 by the first human, Yuri Gagarin. After Luna 2 made the first hard landing on the Moon, the first soft landing was made by Luna 9 in 1966, and the first people set foot on the Moon in 1969 with the Apollo 11 mission. Since then, the Earth is orbited by a large number of artificial satellites and a permanently manned space station, and all the planets of the Solar system have been visited by robotic missions, with soft landings on Mars, Venus, Saturn's moon Titan and comet Churyumov–Gerasimenko.

The first challenge for space travel is to escape from the gravitational grasp of the Earth, or even from the Sun. The escape velocity is (cf. Equations 6.19 and 8.10):

$$v_{\rm esc} \;=\; \left(\frac{2GM}{r}\right)^{1/2} \equiv\; 2^{1/2}\; v_{\rm orbit}, \tag{18.1}$$

for a circular orbit. The escape velocity from Earth's surface is 11 km s^{-1}, but that from Earth's orbit around the Sun is 42 km s^{-1}.

18.1 Basics of rocket propulsion

Traditional rocket propulsion is based on the law of conservation of momentum:

$$m\;{\rm d}v + v_{\rm ex}\;{\rm d}m \;=\; 0, \tag{18.2}$$

or:

$$ {\rm d}v \;=\; -v_{\rm ex}\;\frac{{\rm d}m}{m}. \tag{18.3}$$

If mass is expelled in one direction at speed $v_{\rm ex}$, the rocket will gain a velocity increment dv in the opposite direction. Integration of this differential equation yields the final increment in velocity Δv:

$$\Delta v \;=\; v_{\rm ex}\;\ln\left(\frac{m_{\rm i}}{m_{\rm f}}\right), \tag{18.4}$$

where $m_{\rm i}$ is the initial mass of the spacecraft (including fuel) and $m_{\rm f}$ is the final mass of the spacecraft (after having expelled all its fuel). This can be re-arranged to obtain an expression for the required initial-to-final mass ratio:

$$R \equiv \frac{m_{\rm i}}{m_{\rm f}} = \exp\left(\frac{\Delta v}{v_{\rm ex}}\right). \tag{18.5}$$

18.1.1 Improving efficiency of rocket propulsion

Multi-stage rockets

Multi-stage rockets are essentially composed of several rockets, each of which is detached once it has burnt up its fuel. It is most efficient if each of n stages has the same initial-to-final mass ratio, in which case:

$$\Delta v = n\, v_{\rm ex}\, \ln R, \tag{18.6}$$

or:

$$R^n = \exp\left(\frac{\Delta v}{v_{\rm ex}}\right). \tag{18.7}$$

A large velocity is attainable, but at the expense of a large overall mass ratio at launch.

Another simple procedure to maximise the efficiency of propulsion is the use of a nozzle:

- ⋆ It increases $v_{\rm ex}$. Following Bernoulli's law of fluid dynamics (Equation 3.2), in order to move matter at the same rate through a narrower channel, the speed of the flow must increase. Thus, a small nozzle opening results in a high exhaust speed.
- ⋆ It maximises the forward thrust. If exhaust is expelled in a scattered way, its velocity component opposite the direction of travel will be smaller than if the exhaust is expelled in a collimated beam.

Types of fuel

Chemical rockets use reactive components, such as hydrogen and oxygen. The newly formed chemicals have negative binding energy; this energy is released and goes into heating the gas mixture. The hotter the gas, the faster the atoms move – it is this thermal motion which is directed through the nozzle and becomes $v_{\rm ex}$:

$$v_{\rm ex} = \left(\frac{kT}{\mu}\right)^{1/2}. \tag{18.8}$$

So, the thrust is increased for higher temperature T and/or lower molecular weight μ.

In an ion engine, charged particles (ions) are injected into and accelerated by an electric field. They gain kinetic energy out of the potential energy they start of with:

$$\frac{1}{2} m v_{\rm ex}^2 = qV, \tag{18.9}$$

where q is the electrical charge of the ions and V the potential difference over which the ions are accelerated.

Huge quantities of energy are released in nuclear reactions, the fission (breaking up) of a nucleus or the fusion of two nuclei. This is because of the great strength of the nuclear force, which keeps the nucleus together and which is a consequence of the strong force (which is primarily responsible for keeping nucleons in one piece). The energy is so huge that the difference in mass between the nuclei before and after the process is remarkable. Fusion is more difficult to achieve than fission, as it requires a lot of effort to overcome the electromagnetic barrier between two nuclei (which are always both positively charged).

A problem with nuclear propulsion is that the resulting temperature is too high for any known material to endure. Thus the spacecraft would simply melt (if it does not explode due to the immense pressure). A solution would be to expel nuclear pellets (micro-bombs) and have them explode outside the spacecraft, at some distance behind it. The pressure on the spacecraft as a result of the blast would propel it forward, whilst the level of heating would be acceptable. This was the concept behind a design study made by the British Interplanetary Society in the 1970s – Project Dædalus.

18.2 Gravity assist manœuvres

Energy may be extracted from the orbital energy of a planet moving around the Sun, by passing close to it. For instance, if a spacecraft approaches a planet frontally, with velocity v (in the inertial frame of the Solar System) and the planet is able to gravitationally sling the spacecraft 180° around it, then the spacecraft will leave the planet behind at a gain in velocity of $2v_{\rm orbit}$. To see this, consider the inertial frame with the planet at rest. The spacecraft approaches the planet with a velocity $v + v_{\rm orbit}$ and leaves it with the same $v + v_{\rm orbit}$ (but in opposite direction). Now transform back to the inertial frame of the Solar System, by adding $v_{\rm orbit}$. Hence, the spacecraft leaves the planet with a velocity of $v_{\rm orbit} + (v + v_{\rm orbit}) = v + 2v_{\rm orbit}$. Such gravity-assist manœuvres have been used to accelerate robotic probes on a path towards the outer planets.

Alternatively, a modest velocity boost during a flyby past the Sun can result in a relatively high speed after the encounter. To see how this works, imagine a spacecraft coming from far at low speed, so the initial energy – the sum of the gravitational potential energy and kinetic energy – $E_1 \simeq 0$. As it approaches the Sun, it is accelerated by gravity. At closest approach, the spacecraft will have gained a velocity which is equal to the Sun's escape velocity at that distance. The total energy of the spacecraft remains the same:

$$E_2 \;=\; \frac{1}{2}mv_{\rm esc}^2 - \frac{GMm}{r} \;\equiv\; E_1 \;\simeq\; 0. \tag{18.10}$$

Now imagine applying a boost (for instance by rocket propulsion), resulting in a velocity increment Δv. The energy of the spacecraft thus increases:

$$E_3 \;=\; \frac{1}{2}m(v_{\rm esc} + \Delta v)^2 - \frac{GMm}{r} \;>\; 0. \tag{18.11}$$

The spacecraft's energy after the encounter, far away from the Sun, is:

$$E_4 \simeq \frac{1}{2}mv_\infty^2 \equiv E_3. \tag{18.12}$$

Realising that (cf. Equation 18.10):

$$\frac{GMm}{r} \equiv \frac{1}{2}mv_{\rm esc}^2, \tag{18.13}$$

we obtain an expression for the spacecraft's velocity with which it exits the Solar System:

$$v_\infty = \left((v_{\rm esc} + \Delta v)^2 - v_{\rm esc}^2\right)^{1/2}. \tag{18.14}$$

To demonstrate the power of this procedure, for $v_{\rm esc} = 400$ km s^{-1} (corresponding to $r = 2.4$ R$_\odot$) and $\Delta v = 1$ km s^{-1} only, the spacecraft leaves the Solar System with $v_\infty = 89$ km s^{-1} – eight times the escape velocity from Earth's surface!

18.3 Entering orbit and descent

The orbital speed is a factor $\sqrt{2}$ smaller than the escape speed (Equation 18.1), hence it is easier to launch a satellite into low-Earth orbit than send it to the Moon. If the altitude is negligible (compared to the Earth's radius) then the orbital speed would be 7.7 km s^{-1} and one orbit around the Earth would take 91 minutes.

If the satellite needs to stay stationary above the same point on the Earth's surface, then it will need to orbit at the same rate as the Earth spins:

$$P_{\rm orbit} = \frac{2\pi r_{\rm geostationary}}{v_{\rm orbit}} \equiv P_\oplus \Rightarrow r_{\rm geostationary} = \left(\frac{P_\oplus}{2\pi}\right)^{2/3} (GM_\oplus)^{1/3}, \tag{18.15}$$

which works out at $\approx 36,000$ km altitude above Earth's surface (high, but much closer than the Moon). Similar equations apply to orbiting other planetary bodies (or stars).

To enter orbit upon launch requires gaining kinetic energy; to enter orbit after approach requires losing some. Descent towards, and landing upon, the surface requires further deceleration. The presence of an atmosphere brings two advantages. Firstly, use can then be made of the frictional energy losses as the lander rubs against atmospheric gas particles, super-sonically (so a shock wave will precede it). Precautions must be taken to ensure the heat does not end up burning the lander, but is dissipated through heating atmospheric gas instead, and that the ram pressure does not cause structural damage. The shields of single-use spacecraft deal with the heat by ablation; the Space Shuttle's light-weight aluminium frame instead was protected by insulating ceramic material (fibrous silica) and reinforced carbon – a bit of the latter on *Columbia*'s wing was punctured when struck by foam that had come loose from the tank during launch, causing internal overheating upon re-entry which led to the disintegration of the spacecraft. Secondly, once the speed is down to sub-sonic levels a parachute may be deployed. The atmospheric pressure underneath the parachute will exert an upward restoring force instantiated by the force with which the falling parachute presses upon the air.

It becomes immediately clear that landing on the Earth is relatively easy because of the atmosphere; landing on the Moon because of the low gravity; but landing on Mars is challenging because of the thin air but substantial gravity. The use of rockets may be needed to provide additional deceleration; this is used to land on the Moon, but also by the private venture SpaceX to return its Falcon rockets and rocket boosters back to Earth for renewed use. To *return* from the surface of Mars is also harder than from the Moon, requiring a much heavier vehicle. Landing on a very small object also poses challenges: the feeble gravity of, for instance, a small asteroid means that the slightest excess speed (on landing or when trying to move around across its surface) could mean a return into space.

As landing on sharp or inclined structures, on thick layers of thin dust, active regions (geysers, volcanoes) or on either solid or liquid surfaces, may be undesirable the descent may need to be controlled. Given the long distances over which communication with Earth takes place this must be done *in situ*, either by a person or autonomously via sensory recognition (imaging or other methods for detection of structure and composition) and manœuvering protocols in case of a robotic mission. The latter is also desirable during further exploration after landing. An example is the landing, in 2019, of the Chinese Chang'e 4 on the far side of the Moon, and its rover Yutu 2. Communication (relay of data) takes place via a satellite placed in the second Lagrangian point of the Earth–Moon system.

Chapter 19

Interstellar travel and settlement

Science fiction? Not really. Four man-made robotic spacecraft, the Pioneers 10 and 11 and the Voyagers 1 and 2 were launched in the 1970s and are now leaving the confines of the Solar System. Voyager 1, still sending scientific data, is the farthest at nearly 140 au from the Sun, having crossed the heliopause and venturing into interstellar space at a pace of about 3 au per year. Yet even if it were heading in the right direction (which it isn't), it would take it 100,000 years to arrive at the next nearest star. Thus interstellar travel poses challenges far beyond those encountered by interplanetary travel. But there is serious interest and activity in this arena, with examples such as the Tau Zero Foundation (endorsed by NASA), and the privately funded Breakthrough Starshot initiative to send a laser-sail powered micro-spacecraft to a nearby exoplanet within a human lifetime.

We shall discard the following popular tricks to fool common perceptions of space-time:

- ⋆ Wormholes. These are shortcuts in space, as a result of extreme space curvature. Imagine a folded sheet, and pressing the sheet from above and underneath at the same spot. The two sides of the sheet will touch, and you could imagine an opening emerging, allowing you to pass from one side of the sheet to the other, which would be much quicker than having to travel along the sheet, around the fold, until you reach the other spot. Current theory allows for wormholes to exist. The problem is that the extreme gravity would rip you and your spacecraft apart. The only way to prevent this from happening would be to use a protective anti-gravity shield. But that does not exist, as the agent of gravity, mass can only assume positive values. And it would defy the purpose of the wormhole, making it vanish.

- ⋆ Time travel. Although allowed in most theories (e.g., General Relativity, Quantum Mechanics), running time in the opposite sense would violate the laws of causality. There are thermodynamic arguments based on the concept of entropy, for time to run in one sense only. Note that the exclusive direction or sense is a property shared between time and mass, but not space (in which you can travel back and forth).

- ⋆ Tele-portation. This is also sometimes considered as a means to circumvent the limitations set by the speed of light. Even if tele-portation were possible, coding then transmitting and reconstructing information still cannot be done faster than the speed of light.

19.1 Limits set by relativity theory

It is impossible to travel with the speed of light, and it becomes extremely difficult to approach it very closely. Traveling with a speed very close to that of light has two benefits:

- ⋆ It will take less time to cross space, as clocked by an observer on Earth. A spacecraft traveling with a speed close to that of light would be seen by an observer on Earth to reach a star at 10 light years from Earth in not much more than 10 years.
- ⋆ It will take dramatically less time to cross space, as clocked by the traveler onboard the spacecraft, because of the relativistic effect of time dilation. Thus, a traveler might reach a star at 10 light years from Earth in much less than 10 years on their own clock. This also holds for the "biological" clock, i.e. the traveler would really age much less than 10 years.

The time dilation effect is quantified by the common relativistic factor:

$$\gamma \;=\; \frac{1}{(1-\beta^2)}, \tag{19.1}$$

where $\beta = v/c$. The problem is that the momentum of the spacecraft, too, increases by this factor when compared to classical description:

$$p \;=\; \gamma m_0 v, \tag{19.2}$$

where m_0 is the mass of the spacecraft at "rest" (say, just before launch). Once the spacecraft starts approaching the speed of light, γ increases significantly, and relatively more momentum needs to be supplied in order to accelerate it any further. In the limit of $v \to c$, the required momentum $p \to \infty$. It is thus energetically impossible to exactly reach the speed of light. Whilst there is no real gain in saving time by traveling at 99.9% of the speed of light as opposed to 99%, *in the restframe of the observer on Earth*, the time dilation effect does become much more dramatic so the traveler would greatly benefit.

19.2 Alternative propulsion methods

To reach relativistic speeds ($> 0.1c$), other propulsion methods have been considered, which differ from the rocket:

- ⋆ Solar sail. The radiation pressure upon a highly reflective sail is (cf. Section 10.1.5):

$$P_{\rm rad} \;=\; \frac{L}{2\pi d^2 c}, \tag{19.3}$$

  resulting in a force on the spacecraft of magnitude:

$$F_{\rm rad} \;=\; P_{\rm rad}\; A, \tag{19.4}$$

  and an acceleration of the spacecraft, with total mass (including sail) m, of:

$$\frac{\mathrm{d}v}{\mathrm{d}t} \;=\; P_{\mathrm{rad}}\,\frac{A}{m}. \tag{19.5}$$

Note the same dependence on distance to the Sun as the gravitational acceleration.

⋆ Beam-powered solar sail. To mitigate the dilution of the Sun's radiation, the solar sail can be powered by a collimated beam. This could be a (very) powerful laser, or collimated sunlight. The latter can be achieved by placing an appropriate optical element with the Sun in one of its focal points. The emerging rays are then parallel and the intensity of the beam is independent of distance. The most practical type of lens would be a Fresnel zone plate, which uses the principle of diffraction – as opposed to refraction upon which lenses are based (but which makes them heavy). The zone plate can be made of light-weight opaque rings, with gaps at separations such that the pathlengths l from adjacent gaps to the focal point satisfy the condition for constructive interference of light with wavelength λ:

$$\Delta l \;=\; n\,\lambda. \tag{19.6}$$

⋆ One could also use the Sun's wind. The electric solar wind sail, or 'E-sail', employs electrically charged tethers (long, thin strings) with which the solar wind plasma interacts via Coulomb scattering, thus applying a ram pressure onto the structure.

⋆ The interstellar ram jet engine. In aircraft ram jet engines, ambient air entering the engine with sufficient speed is compressed and then heated (by continuous ignition of fuel). It leaves the engine at high speed. Such engines work at supersonic speed.

⋆ The magnetic scoop aided interstellar ram jet engine. Also called the Bussard engine, it uses a magnetic field with open field lines in front of the spacecraft's engine, to direct interstellar charged particles into the engine. Otherwise, the density in the diffuse interstellar medium is too low for a ram jet engine to operate.

⋆ An electro–magnetic gun. Giving the spacecraft electric charge, it can be accelerated in an electro-magnetic field (just as in particle accelerators) and then ejected as if it were a bullet, much like in Jules Verne's *Journey to the Moon*.

⋆ The scissors slide. The point at which scissors blades touch (i.e. where they are used to cut) moves towards the tips of the scissors much faster than the speed with which the scissors handles are being brought together. Placing a spacecraft in between the blades, and closing the scissors at subluminal speed, the spacecraft can in principle be propelled to superluminal speed – but in practice this is impossible as it would take an infinite amount of energy to close the scissors if there is a spacecraft in between the blades that has (nearly) reached the speed of light.

19.2.1 Beyond initial speed

Besides accelerating the spacecraft to attain high speed, there are other issues of flight:

- ⋆ Manœuvering. The spacecraft may need to change direction. One way of doing this would be to utilise the magnetic field of the interstellar medium, by applying charge to the spacecraft. The Lorentz force is then:

$$\vec{F}_{\rm B} \;=\; q\vec{v} \times \vec{B}. \tag{19.7}$$

  Only curves perpendicular to the magnetic field lines can be made, in one sense or (by applying the opposite charge) the other.

- ⋆ Deceleration. If the spacecraft needs to be placed in orbit around a star or planet, or land on it, it needs to be decelerated. It may pay off to time the arrival according to the anticipated relative speed of spacecraft and planet.

- ⋆ Return or onward flight. In this case, it needs to be propelled again to high speed. This cannot rely on a Solar System bound launching mechanism.

19.3 Challenges for interstellar travel

There are many issues which make interstellar travel difficult, especially manned flight:

- ⋆ Construction difficulties. Spacecraft may need to be massive and large, which poses challenges for building it (probably in space). It may also be limited by the available resources, such as raw materials or energy.

- ⋆ Heating. This is a problem when passing close to the Sun, for gravity assist and/or when using a solar sail (wich is never 100% reflective). A large temperature gradient between the sun-lit and shadow side can be mitigated by rotating the spacecraft.

- ⋆ Strong acceleration. This causes stress and may break the spacecraft. An especially precarious situation occurs in the presence of differential acceleration. This might happen for very large sails, for instance. Humans, too, can only endure accelerations of a few times that of Earth's gravity, for a short time; strong acceleration cuts the oxygen flow to the brain and pulls down the lung diaphragm.

- ⋆ Weightless-ness. Zero (effective) gravity occurs in orbit or away from stars and planets. In humans, this leads to decalcification of the skeleton. It is reversible for flights up to a year, but may not be so for longer flights. Weightless-ness also causes fluid and hormonal imbalances, blood production to drop and the heart to shrink. Moving around while weightless is hampered by linear and angular momentum conservation, and ears and eyes will disagree about orientation and motion.

- ⋆ One way to get around weightless-ness is to apply rotation to the spacecraft, creating the illusion of a centripetal force in the absence of a real attractive force responsible for the rotation. This allows passengers to walk on the inside of the outer wall of the spacecraft. If this "wheel" is too small then differential weight is noticeable between head and feet. When walking around, the Coriolis force makes itself noticeable.

- ⋆ Descent. We already discussed this in Section 18.3. And you may wish to be able to lift off again!

- ⋆ Radiation. This poses a threat already in the Van Allen belts and in interplanetary space, but even more so outside the protective heliosphere and on long flights in general (even in aircraft!). To protect the spacecraft and its inhabitants, a solid shield (heavy) or a magnetic field shield (energy-expensive) can be considered.

- ⋆ Impacts. While close encounters with large rocks are unlikely, micro-meteoroids up to mm size can do plenty of damage and over time can erode surfaces ('sputtering'). Cruising through a molecular cloud or supernova remnant is not advisable!

- ⋆ Energy generation. For keeping the spacecraft functioning, and everything upon which the inhabitants rely, energy needs to be produced continuously. Far from stars, solar panels would be useless. A nuclear reactor might be needed.

- ⋆ Nourishment. Inhabitants need to eat and drink. To ensure a continuous supply of food and drink, an ecosystem needs to be devised.

- ⋆ Waste. Rather than disposing, a near-100% recycling of waste products may be needed. On Earth, recycling does not usually mean that the same items are produced that generated the waste. In space it would be desirable if this were the case, though.

- ⋆ Social issues. Locking up people in a confined space can cause serious tension. Being away from home (possibly never to return, or even to reach a new "home"), lacking sunlight and boredom are factors which tend to cause negative emotions in people.

- ⋆ Suspended animation. Stopping and reviving metabolism and/or aging in humans is not within the remotest reach of our capabilities. For instance, freezing or freeze-drying causes irreparable damage to cell tissue. It it thus more likely that space colonies are established, where it is accepted that only a (distant) future generation may arrive at the destination.

- ⋆ Navigation. Lacking a fixed reference frame on Earth, navigation will need to rely on an astronomical reference frame. Stars can be used, but they all have peculiar velocities which will need to be taken into account – especially if it is the one aimed to reach. Furthermore, if traveling at relativistic speed, the Lorentz contraction of length scales by the factor γ needs to be taken into account.

- ⋆ The Twin Paradox. People traveling near the speed of light and returning to Earth, will see their family and friends having aged more than they themselves. Travel to and from the nearest star, Proxima Centauri at 4.2 light-years, could have taken mere days for the traveller, but by the time the traveller has returned to Earth life on Earth will have moved on by 8.4 years. Possibly, when visiting more distant stars, many generations will have passed, and Earth's human civilization has changed beyond recognition (or ceased to exist).

19.4 Challenges for extra-terrestrial settlement

NASA has serious plans for establishing permanently inhabited stations on the Moon and Mars, and experiments have been carried out that attempted to create a closed, self-reliant ecosystem from scratch (Biosphere 2) – which turns out to be problematic.

Some of the challenges to settlement on other planetary bodies include:

- ⋆ Atmospheric composition. A hospitable atmosphere must include oxygen to breathe (but in moderation, or it will oxidize too much), ozone to protect against ultraviolet radiation, and an inert gas such as nitrogen to make up the rest. Toxic trace elements need to be avoided. Low pressure will cause blood to boil, and eyes and ears to bust.
- ⋆ Terraforming. To recreate Earth-like conditions, the existing ecosystem may need to be changed by setting off a runaway process (e.g., by a Greenhouse effect) possibly by creating a basic biosphere which does the rest for us.
- ⋆ Radiation. Unless there is a dense atmosphere and strong magnetic field, protection against radiation is critical. The (linear) 'stopping power', $S(E)$ quantifies how much energy E the radiation (with initial energy E_0) loses as it travels a certain distance x through material, with the mean range, Δx found from

$$\Delta x \;=\; \int_0^{E_0} \frac{1}{S(E)}\,\mathrm{d}E. \tag{19.8}$$

  Particles lose energy first through electronic interaction, followed by nuclear interaction, resulting in $S(x)$ ('Bragg curve') to reach a peak before a sudden drop.
- ⋆ Gravity. Weak(er) gravity may pose a health hazard (see above). Stronger gravity might also cause problems – it will make the heart work harder, for one thing.
- ⋆ Raw materials. For construction and consumption, these may need to be mined or synthesized, for instance water, or hydrocarbons to make plastics).
- ⋆ Communication with home. This is not a serious problem within the Solar System (although it may take up to 40 minutes for a message to make it to Mars and back), but it is impractical between stars where messages will take many years to arrive.

19.5 Ethical considerations

Landing on an extra-terrestrial object brings with it the serious risk of contaminating it, either with biotic material or something that could trigger processes that would otherwise not have occurred (such as life instantiating or delibitating mechanisms). Despite some intentional crash-landings by NASA on comets, the space community is by and large united in its concerns about spreading life. We have stark examples on Earth where invasive species have upset ecosystems, including the introduction of smallpox in South America by European explorers which decimated indigenous peoples. Contaminating, say, Mars could make it difficult to determine whether life was present already.

But introducing life to other worlds might be done on purpose. Settling is an obvious example, along with 'Terraforming'. But it might also be done purely as an experiment, to study whether life could establish itself and how it would evolve. Some would advocate that, with so many exoplanets and the possibility that the emergence of simple life is actually quite common and panspermia might be a naturally occurring phenomenon, fertilizing exoplanets with just prokaryotic life might be acceptable.

Chapter 20

Extra-Terrestrial Intelligence

Aside from Gauß's proposition in 1826 to cut a humongous triangle in the Siberian forest to draw the attention of lunar inhabitants, serious attempts to search for, and communicate with extra-terrestrial intelligent civilizations started not long after the dawn of radio astronomy, in 1959/60. Although first met with skepticism, most radio observatories are now sympathetic towards the Search for Extra-Terrestrial Intelligence (SETI).

It has also been realized that technologically advanced civilizations are characterised by a high energy demand and thus a likely energetic appearance. To quantify this, the Kardashev scale was invented in 1964 – and modified by Carl Sagan in 1973 – and can be formulated as

$$K \;=\; 2 \;+\; 0.1 \;\log L[\mathrm{L}_\odot]. \tag{20.1}$$

Thus, a Kardashev 2 civilization has a luminosity equal to that of the Sun; a Kardashev 3 civilization that of an entire galaxy. That seems a promising premise to search for them.

20.1 SETI

Many radio frequencies can be searched for simultaneously, and often "for free" whilst other astronomical programmes are being carried out. SETI@home taps the tremendous pool of private personal computers in the world, by offering a screen saver which crunches on SETI data whilst the computer owner isn't using their machine. These screensavers perform Fast Fourier Transform operations, identifying periodicities in the variation of signal strength with time. This way, millions of CPU *years* are being used. The privately funded Breakthrough Listen Initiative also uses radio telescopes for SETI (they listened to interstellar asteroid 'Oumuamua, in case it were a spaceship, setting a limit of 0.08 W).

One reason for not having found ETI, is that we have supposed that they communicate in ways which we can imagine based on our current technological heritage. Searches are often conducted around the 1420 MHz radio transition of atomic hydrogen (due to electron–nucleus spin–spin coupling), as this is a very strong line in the sky of any civilization because hydrogen is omnipresent throughout the Milky Way galaxy. Many astronomical programmes observe this transition, allowing commensal SETI programmes to be conducted. It is also in the broad range of radio frequencies for which some of our own communication methods were designed (this has resulted in a continuous battle between astronomers and industry over protection and commercial exploitation of the radio sky).

20.1.1 What would one be looking for?

To distinguish an artifical signal from ETI, from a natural signal, one needs to identify structure in the data that cannot be explained by a natural phenomenon. This can be structure in time or in radio frequency, and is likely to contain periodic components (just like in our communication use is made of a periodic carrier wave which is modulated with the message of interest). Nature produces many periodic phenomena, though, on all timescales down to a millisecond (as in some pulsars) or perhaps even faster. To *find* artificial intelligence may require *employing* artificial intelligence: 'machine learning'.

Interstellar communication is impractical for species with life expectancies of the order of a century (like us). Perhaps extra-terrestrial species exist that live many times longer, maybe with a much slower metabolic rate (and for whom time appears to run slower in a biological and perceptional sense). They might not mind waiting a dozen years for an answer to their message. Or they may have other reasons for trans-generational communication. Still, it seems likely that much more communication takes place *within* societies, and mostly in a lateral direction to reach others elsewhere on the surface of a planet. These broadcasts might seep out, and we might overhear them. If their planet was inhabited in a very inhomogeneous way, like major urban centres on Earth, then one would expect the signal to peak sharply around the rising and setting times of these centres as the planet rotates (assuming we don't view the planet pole-on). That being said, the use of directed low-power masts and optical fibres have rendered Earth much more radio silent. But light pollution has increased, betraying our presence.

The potential is being realised of the use of lasers in transmitting messages over large distances, as lasers are highly collimated beams of electro–magnetic radiation and thus remain intense far away from the transmitter. It is highly directional, for the same reasons, and thus unlikely to point our way – unless intended for us! Lasers by convention operate at optical wavelengths ("light"), but the amplified stimulated emission which defines it can also operate at other wavelengths. For instance, micro-wave sources of such intense emission exist in nature, called 'masers' (in particular abundant molecules such as water, OH, SiO and methanol, but for the mechanism to work special conditions need to be met to create a population inversion of the energy levels with respect to their natural occupation, and to maintain the coherency of the electro–magnetic wave). Interstellar electrons, however, scatter electro–magnetic waves, smearing out signals over typically

$$\Delta\nu[\mathrm{Hz}] \;\sim\; 0.1\;\nu[\mathrm{GHz}]^{-6/5}. \tag{20.2}$$

This favours optical (high-frequency) transmission over radio-wave transmission.

False alarms include PSR B1919 +21, the first pulsar (extremely regular pulses of radio emission from a neutron star) discovered in the UK in 1967 by Jocelyn Bell Burnell, and GCRT J1745−3009, a source of transient bursts of radio emission seen in the general direction of the Galactic Centre (but explained as a coherent *auroral* maser).

More sinister scenarios include detecting the signs of nuclear war, or other high-energetic and/or explosive manifestations of advanced technology. Less dramatic would be detecting lighting on the night side of a planet, or atmospheric pollution such as CFCs.

20.2 CETI

Communication with Extra-Terrestrial Intelligence (CETI) has not been pursued with rigour. Frank Drake's message broadcasted in 1974 with the most powerful radio dish on Earth, at Arecibo, has not been followed up upon. This is partly because the signal will be extremely weak by the time it arrives at any planet inhabited by a civilization with the technological capability (and the interest) to intercept it. It is also because we do not yet know what are good targets to aim at. Drake's message was sent to the globular cluster M 13. The idea being that with hundreds of thousands of stars, there might be a fair chance that some of them would have a planet inhabited by an intelligent species. However, stars in globular clusters are tightly packed together and notice each other's individual gravitational potential much more regularly than the Solar System notices that of its neighbours. It is therefore quite likely that planetary systems in globular clusters are disrupted, if they formed at all. Searches for planetary transits in globular clusters indeed seem to suggest that Hot Jupiters may be less common in globular clusters than in the field. In any case, Drake's message will only arrive around about the year 27,000.

Communication by "snail mail" has also been initiated. The Pioneer probes carry a gold disc with engravings of our place in the Solar System and the place of the Sun with respect to nearby stars, so nearby civilizations can figure out where we are (which may well have been a dangerous thing to do!). It also bears pictures of the hydrogen atom with the spin of electron and nucleus parallel and anti-parallel, notifying the relevance of the 1420 MHz transition. It also shows examples of both gender of our species, with a friendly expression (as perceived in our culture). The Voyager probes also carry such disc, but this time engraved with audio recordings; some additional instructions were engraved to help aliens play the disc. Already in our own civilization, engraved records ('vinyl') are largely obsolete.

20.2.1 What would a message be like?

Apart from the not unimportant questions with regard to what we might want aliens to know about us (and what not), and what would they be able to understand (it is not easy for us to understand exactly what monkeys are trying to tell us), how would we write it?

With regard to the format of the message, it is clear that the "language" in which the message is written should not rely on humanoid conventions, such as units or the awkward decimal system. The message should be based upon universal standards that are likely to be recognised by any intelligent species. Examples include:

- ⋆ Mathematical fundamental numbers. For instance, π is not a concoction of our own making, but quantifies how radius, circumference, surface area and volume of spherical entities are related in the spatial framework that makes up the Universe we live in. Another example, e is not defined by us but by the way in which waves can be described in complex notation.

- ⋆ Prime numbers. Drake's message was a stream of digits, but it was made up of a prime number of rows with each a prime number of "characters".

- ⋆ Atomic structure. Although the second and thus the Hertz are units of our own choice, the transition between energy levels within an atom really does not care about how we quantify it – it is a universal truth.

- ⋆ Binary system. On–off, true–false, a notation in terms of zero and one makes sense. Drake's message was simply a stream of switching the transmitter on and off. Whether aliens would read binary notation from the right to left as in our convention has to be seen, but at least they might be expected to try both ways (let's hope reading our message from right to left does not cause offense).

One further consideration when communicating with aliens is he possibility that their messages are infected with viri, which could intrude, upset and perhaps even take over our technological infrastructure. One should be careful when decoding such messages.

20.3 Dyson spheres

It has been speculated, first by Freeman Dyson, that intelligent, technologically advanced civilizations may need to find an alternative to their planetary habitat, for reasons of population expansion, energy generation, environmental degradation, or other. One in principle simple idea would be to enclose a star by a rotating solid sphere. The advances would include:

- ⋆ The enormous surface area could accommodate a huge population.

- ⋆ All energy from the star would be captured and made available.

- ⋆ Waste would be disposed of by throwing it into the star.

But this is not as simple as it sounds. Construction would cost a gigantic amount of raw materials, energy, and money (if the aliens have an equivalent concept). Stability of the structure would be another concern. It is more likely that a ring would be fabricated, rather than a closed sphere.

A Dyson sphere would look from afar like an object with the luminosity of the star, but a temperature around 300 K or so. It would shine predominantly at infrared wavelengths. It would be much larger than normal for a star of that luminosity or any object of that temperature, typically an au. Nearby Dyson spheres might be resolved with large telescopes. But analyses of various sensitive all-sky infrared surveys have revealed no Dyson spheres as bright as the Sun within 300 pc.

20.4 Colonization by intelligent civilizations

It seems not unlikely that technologically advanced civilizations will be able to construct spacecraft that travel at moderately-relativistic speed. From our own experience on Earth, advanced civilizations tend to expand, either out of necessity, greed, or curiosity. Even if they do not travel (much) themselves, they may be sending robotic reconnaissance probes. Even just communication, using relativistic messenger probes would be nearly as fast as sending electro–magnetic signals and vastly less expensive.

If a civilization travels with 10% of the speed of light, and each century settles on a planet, builds a second spacecraft, and both spacecraft then set out in different directions, and this repeats itself century upon century, then the entire Milky Way galaxy (some 100,000 light years across) would be colonized within a million years. Even if the spacecraft would travel one or two orders of magnitude slower, and it would take a millenium for the second spacecraft to be built, then still the feat would be accomplished within a billion years. The Universe has an age of about 13.8 billion years.

The Fermi paradox thus states: if intelligent life can colonize the Galaxy, where are they?

Perhaps the above scenario is too naïve. One can think of all sorts of reasons why the failure rate of colonization attempts may exceed their success rate, and why there would be much longer delays before any subsequent colonization attempt. But also the spread of a civilization may not be as dramatic. If only one spacecraft is sent onto a next mission, then either the spread occurs similar to the diffusion process – which results in a fairly uniform and well populated region of space but which does not expand very rapidly – or it resembles a wave, resulting in a linear expansion of the range but at an increasingly sparse coverage. The speed at which a galaxy is colonized also depends on the distances that civilizations are willing to bridge: it is natural to think that they would try to remain within communicating distance (i.e. within a generation lifetime) of their home, which for humans would be of the order of 10 light years. Thus, proliferation of life throughout the cosmos may occur through percolation, rather than by avalanche.

The possibility that there are no other intelligent civilizations seems hard to accept as it contradicts the now widely (in science) accepted Copernican principle: we do not take a special place within the cosmos. If we are here, there should be others, too. But perhaps not in our galaxy – there are many billions of other galaxies out there, where intelligent civilizations could still exist. In a galaxy far, far away...

Fundamental limiting factors to colonization may exist such as the lifetime of civilizations. Colonization can be likened to the concept of virulence. Parasitic civilizations may destroy their hosts and thereby limit their own longevity, unless they are able to find a new host quickly. Mutualism between civilization and host, on the other hand, leads to longevity of the civilization but probably slower population of space. If civilizations do not last, then a spreading civilization will leave a decaying civilization in its wake. If the pressure on society, resources and environment resulting from expontential population growth (say, at a sustained 1% per year) were the driver behind colonization, then colonization can never relieve this pressure – only start over somewhere else again. The colonization may spread like a wave but leave large swaths devoid of civilizations, or attempt to repopulate extinguished civilizations and thus spread less quickly. Also, when we do intercept a message from a civilization perhaps a thousand light years away, or a probe travelling at much less than the speed of light, that civilization might no longer exist.

20.5 Closing remarks

Perhaps there are other factors that restrict the proliferation of life throughout the Universe. Perhaps life is very rare indeed, especially complex life, more so intelligent life. After all, we have not found life in other places in the Solar System despite indications

that some simple life might be able to survive or even have existed on Mars, on Europa, perhaps Titan. Life on Earth did not take off in earnest until 600 million years ago, after a period of glaciation we still don't know what caused it. And advanced technology has only been in existence for mere millenia, centuries, or decennia, depending what one would consider "advanced". We contemplate our own existence and the hypothetical existence of other civilizations because we can – we did not do this a million years ago, and it is hard to imagine other species doing this today. Step changes in evolution may require special conditions, perhaps catastrophic events, and even luck[1], and may not always work out. These potential barriers are referred to as 'Great Filters'. With us having passed through more than one such filter, we may find ourselves in an unusual, not common situation.

Likewise, of the many kinds of Universes that might (perhaps) be possible, we live in one that makes life possible. If the fundamental constants had different values – for instance the fine-structure constant or the gravitational constant – the Universe would look very different. Thus, typically we would expect to find ourselves in an atypical situation; this is an important bias in our philosophy.

It has been argued that we live at the pinnacle of our civilization, because population growth has meant that we are much more likely to have been born in the past few decennia than a couple of centuries ago. While this is true based on the past, known history of population growth, obviously people have lived a thousand years ago and it could have been us living then, and pondering the possibility that civilization is coming to an end soon. There is no predictive power in this; Gott's reasoning that led to the above Doomsday Argument is based on an unrestricted prior distribution and the absence of characteristic timescales, whereas in our case the prior distribution is known and the lifetime of a human individual is limited.

What is certain, though, is that life and the conditions for life cannot be taken for granted. They are special, but not guaranteed. While we search for lie elsewhere in the Universe, and take strides in human exploration of space, we must not forget and neglect the only place we call home. Because we have the capacity to change it, and we are conscious of the effects of our actions, we have a duty to preserve it. We don't want to fall victim of the next 'Great Filter'...

[1]Ice ages may have reduced the human population to just a few hundred individuals living in South Africa some 150,000 years ago.