Particle Physics

CERN Lectures

David Tong

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OBA, UK

http://www.damtp.cam.ac.uk/user/tong/particle.html
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

There are a number of textbooks and semi-popular books that hit roughly the right level. First the semi-popular:

- Tini Veltman, "*Facts and Mysteries in Elementary Particle Physics*"

A straightforward book describing the essentials of the particle physics with some historical anecdotes thrown in for colour. Veltman won the Nobel prize with Gerard 't Hooft for demonstrating the renormalisability of the Standard Model.

- Abraham Pais, "*Inward Bound*"

A spectacularly detailed book, covering discoveries in particle physics throughout the 20^{th} century by a scientist who had a ringside seat for many of the key developments. Pais was no slouch as a physicist, providing important classification schemes for particles, naming both "baryons" and "leptons" and, with Gell-Mann, the first to understand the mixing of neutral kaons. On the flip side, he also coined the wildly inappropriately modest name "The Standard Model". (And he didn't even capitalise it.)

- Leon Lederman, "*The God Particle*"

To get anything out of this book you first have to get past the appalling title and then past the grating voice in which every story is buried neck deep in wisecracks, delivered with the all the subtlety of a dinner party bore after a line of cocaine. Still, given that Lederman is one of the great experimental particle physicists of the past century, it is sometimes worth the effort.

If you want more mathematical meat, then there are a few books that require a knowledge of quantum mechanics but fall short of using the full machinery of quantum field theory. Two good ones are:

- Halzen and Martin, "*Quarks and Leptons*"

- David Griffiths, "*Introduction to Elementary Particles*"

Contents

Acknowledgements

Every summer, CERN plays host to a cohort of university students from around the world. The students come from a range of backgrounds, from theoretical and experimental physics, to computing, engineering and mathematics. They spend the summer working on some of the CERN experiments, while taking a number of crash courses designed to get them up to speed with the CERN mission.

These lecture notes form the introduction to the CERN course. They cover the basics of particle physics and are designed to be accessible to students with any scientific background. This means that, despite the advanced topics, the notes require significantly less mathematical sophistication than my other lecture notes. Sadly there is a price to be paid, and this comes in the form of facts. Lots and lots of facts. But particle physics is a subject where it helps to know what you're getting yourself into before you meet the somewhat daunting mathematics. These lectures should hopefully give you a flavour of what awaits in more detailed courses.

1 Introduction

The purpose of these lectures is to address one of the oldest questions in science: what are we made of? What are the fundamental building blocks of the universe from which you, me, and everything else are constructed?

In the twentieth century, progress in addressing this question was nothing short of spectacular. By the time the dust had settled, we were left with a remarkably simple picture: every experiment that we've ever performed can be explained in terms of a collection of particles interacting through a handful of forces. The theory which ties all of this together is the pinnacle of 350 years of scientific endeavour. It is, by any measure, the most successful scientific theory of all time. Yet we give it a rubbish name: it is called the *Standard Model*.

These lectures have two, intertwined narratives. The main thread describes the contents and structure of the Standard Model. The language in which the Standard Model is written is known as *quantum field theory*, and much of our initial focus will be on describing this framework, and the way it forces upon us certain inescapable facts about the universe. As we proceed, the emphasis will be on the key theoretical ideas that underpin the Standard Model and more detailed descriptions of the particles and the forces that make up our world.

Many of the ideas that we will meet are abstract and counterintuitive and they took physicists many decades to understand. It is striking that, at nearly every step, what ultimately lead physicists in the right direction was experiment. Sometimes these experiments simply confirmed that theorists were on the right path, but more often they came as a surprise, forcing physicists in entirely new directions.

The second thread of these lectures is to describe these experimental advances, and attempt to connect them to the more theoretical ideas. With this goal in mind, each chapter ends with an accompanying "Interlude" in which we take a more historical tour through the subject, describing some of the key experimental results, along with some of the confusions that plagued the physicists of the time.

Finally, it is clear that the Standard Model is not the last word, and we will see what questions it fails to answer. In last Section we raise some of the outstanding open problems and speculate a little on what lies beyond.

In the remainder of this extended introduction, we will give a whirlwind tour of the Standard Model, painting the big picture by omitting many many details. As these

lectures progress, we will fill in the gaps. Much of the theory sits on an excellent footing, in the sense that we now know that some aspects of the laws of physics could not be any different: many details are forced upon us by mathematical consistency alone. In contrast, other parts of the theory remain mysterious and it is unclear why the world appears one way, rather than another.

The Structure of the Atom

Before our forefathers understood atoms, or nuclei, or elementary particles, they understood chemistry. And this is where our story starts.

Our modern, scientific understanding of the structure of matter begins with the chemist John Dalton and his "law of multiple proportions". This is the observation that, when mixing elements together to form different compounds, you should do so in integer amounts. So, for example, you can combine carbon and oxygen to form one type of gas, but it you want to make a different type of gas then you need exactly double the amount of oxygen.

That's surprising. It's not, for example, what happens when you bake. If you've discovered the perfect recipe for sourdough, then you don't double the amount of flour and suddenly find that you've got bread for a bagel. That's not the way things work. Dalton understood the importance of his observation which he interpreted, correctly, as evidence for the old ideas of Leucippus and Democritus who argued that matter is made of indivisible objects called atoms.

These days, the idea of atoms is sewn into the names that we give to the gases. Add carbon and oxygen in equal measure and you get carbon monoxide CO. Double the oxygen and you get carbon dioxide CO_2. But there's no such thing as $CO_{\sqrt{2}}$ because you can't have $\sqrt{2}$ oxygen atoms attached to each carbon atom.

A fuller picture came with Mendelev's arrangement of the elements in the pattern known as the periodic table. In 1867, he placed the elements in (roughly) order of their mass, grouped together based on their observed properties. Mendelev realised that the gaps in his table were opportunities, rather than flaws: they were elements that were yet to be discovered. In this way, he predicted the existence of germanium, gallium and scandium. It would not be the last time that a theorist was able to predict the existence of a new, seemingly fundamental, particle.

From the perspective of a chemist, the periodic table is important because it places elements in groups with similar behaviour. Those elements on the left of the table go fizz when you put them in water. Those on the right don't. However, from the perspective

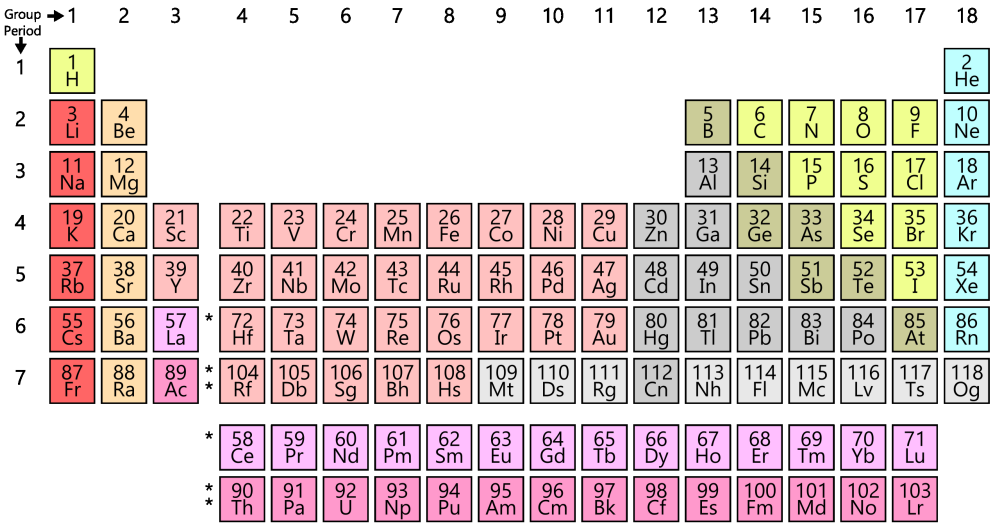


Figure 1. The periodic table of elements. (Image from Wikipedia.) We can now do better.

of fundamental physics, the importance of the periodic table can be found in the clues it gives us for what lies beneath. By Mendelev's time, it had long been understood that elements are made of atoms. The name atom was optimistically derived from the Greek "atomos", meaning "indivisible". The order in the periodic table suggests a structure to the atoms. The elements are labelled by two numbers. The atomic number Z is an integer and tells where the atom sits in the table. The atomic weight A tells us the mass of the atom and, for the first few elements in the table, is very close to an integer. We now know that both of these numbers have their origin in the fact that atoms are very much divisible.

Concrete progress came in 1897 when JJ Thomson discovered the particle that we now call the electron. He announced the discovery to a stunned lecture room at the Royal Institution in London. Thomson later recalled that one of the distinguished scientists in the audience told him that he thought the whole thing was a hoax.

It took another 35 years to unravel the full structure of the atom, with much of the work done by Ernest Rutherford and his colleagues. By the time the dust had settled, it was clear that each atom consists of a nucleus, surrounded by a somewhat blurry cloud of electrons. The nucleus itself is comprised of two further particles, the proton and neutron. The atomic number Z counts the number of protons in the nucleus; the atomic weight A counts (roughly) the combined number of protons and neutrons.

The electrons carry electric charge. By convention, the charge is taken to be negative, an annoying choice that you can blame on Benjamin Franklin, one of the founding fathers of the US. The protons carry positive charge. The neutrons are, as their name suggests, neutral. Remarkably, but importantly, the magnitude of the charge of the proton is exactly the same as the electron; they differ only in sign. Atoms contain an equal number of electrons and protons and so are themselves neutral.

The nucleus sits at the heart of the atom, but is tiny in comparison. Atoms have a typical size of 10^{-10} m, while the nuclei have a typical size of 10^{-15} to 10^{-14} m. Rutherford himself used the analogy of a fly in the centre of a cathedral. Despite its small size, the nucleus contains nearly all the mass of the atom. This is because the protons and neutrons are much heavier than the electron. We will learn more about the properties of particles later, but for now we'll just mention that the mass of the electron is

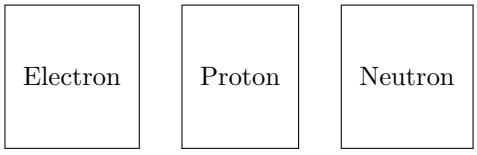
$$m_{\text{electron}} \approx 9.1 \times 10^{-31} \text{ kg}$$

(The kilogram is a useful unit when weighing humans. Less so for elementary particles. We'll meet a better unit shortly.) Both the proton and neutron are roughly 2000 times heavier. Or, more accurately,

$$\begin{aligned} m_{\text{proton}} &\approx 1837\, m_{\text{electron}} \\ m_{\text{neutron}} &\approx 1839\, m_{\text{electron}} \end{aligned}$$

The fact that the masses of the proton and neutron are so close remains something of a mystery. As we will later see, it is related to the fact that two smaller particles called quarks have almost negligible masses but this, in turn, is not something that we can explain from more fundamental arguments. Nonetheless, the approximate equality $m_{\text{proton}} \approx m_{\text{neutron}}$ is important and is the reason that the atomic weights A are so close to integers for the light elements. For now, these numbers simply tell us that the electrons contribute less than 0.1% of the mass of an atom.

The story above paints a much simpler picture of the structure of matter than that proposed by Mendelev. At the fundamental level, the complicated periodic table can be replaced by something significantly simpler. It would appear that we need just three particles to explain the elements. They are:



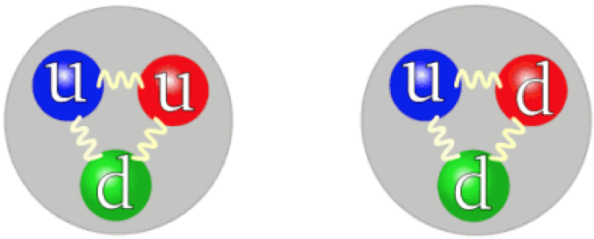


Figure 2. The proton, shown on the left, contains two up quarks and a down quark. The neutron, shown on the right, contains two down and an up.

Sadly physicists had less than 100 days to enjoy this simple picture! The neutron was the last of the three particles to be discovered. That happened in May 1932. In August of the same year, anti-matter was found and subsequent discoveries then came thick and fast. We now know that the sub-atomic world contains many riches beyond the three obvious particles that can be found in atoms.

The Structure of the Structure of the Atom

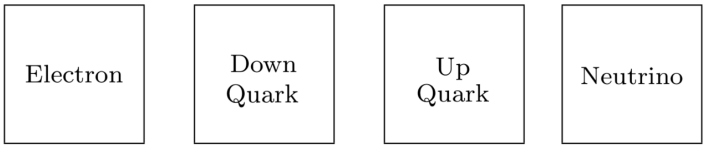
The history of how we understood the structure of matter is long, complicated and confusing and will be told in part as these lectures progress. But, at first blush, the end result seems to be not much more complicated than that of the rosy picture that scientists had in early 1932. For all we can tell, the electron remains as a fundamental particle. In contrast, neither the proton nor the neutron are fundamental. Each contains smaller particles known as quarks. A cartoon version (we'll do better later) says that the proton and neutron each consist of three quarks that, in turn, come in two different varieties. They are called the *up quark* and the *down quark*. The names are not particularly evocative: there is nothing "up" nor "down" about either of the quarks.

The proton contains two up quarks and a down quark, while the neutron contains two down quarks and an up. Both quarks have fractional electric charge. In units in which the electron has charge -1, the up quark has charge $+\frac{2}{3}$ and the down quark charge $-\frac{1}{3}$. This then gives the familiar charges of the proton ($\frac{2}{3}+\frac{2}{3}-\frac{1}{3}=+1$) and the neutron ($-\frac{1}{3}-\frac{1}{3}+\frac{2}{3}=0$).

In addition, there is a further, neutral particle called the *neutrino*. This is not one of the building blocks from which we're made but, as we shall see, has an important role to play in the universe. The neutrino carries no electric charge and is much lighter than all the other particles. It is usually introduced with the epithet "elusive": it barely

interacts and can travel through a lightyear of lead with only a 50% chance of hitting anything.

This, then, leaves us with our new periodic table, a world with just four particles:



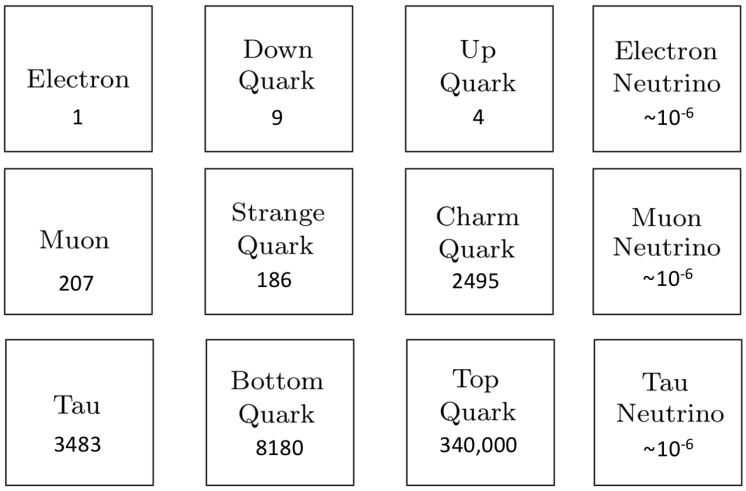
One might have thought this was a good place to stop. However, at this stage, something strange happens. For reasons that we don't understand, Nature chose to take this pattern of four particles and repeat it twice over. The total number of particles of this type that we know about in the Universe isn't 4, but 12.

In addition to the electron, there are two further particles. They behave like the electron in every possible way. For example, they have the same electric charge and, as we will see later, the same interactions with other forces. The only way in which they differ from the electron is that they're heavier. They are called the *muon* and the *tau* (pronounced to rhyme with "now"). They have masses

$$m_{\text{muon}} \approx 207\, m_{\text{electron}}$$
$$m_{\text{tau}} \approx 3483\, m_{\text{electron}}$$

Similarly, there are two extra neutrinos and four extra quarks. The neutrinos inherit their name from the corresponding electron-like particle: we talk of "electron neutrinos" and "muon neutrinos" and "tau neutrinos". The quarks in the second group are called *strange* and *charm*. There was a brief time when physicists toyed with the idea of naming the third group of quarks *beauty* and *truth*. The latter was subsequently rejected out of a well-placed sense of embarrassment, but the names that remain in their place, *bottom* and *top*, are astonishingly dull[1]. This, then, is the final pattern of particles that we find ourselves with:

[1]Nearly everyone now refers to the b quark as bottom. One exception is LHCb, an important experiment at CERN devoted to the study of b quarks. They prefer the older name, presumably because they would rather be LHC-beauty than LHC-bottom. The obvious suggestion that they embrace both names and rebrand themselves LHCbb has gone sadly unheeded.



The numbers in the table are the masses of the particles, written as multiples of the electron mass. (Hence the electron itself is assigned mass 1.) The masses of the neutrinos are known to be very small but, otherwise are only constrained within a window and not yet established individually.

Each horizontal line of this diagram is called a *generation*. Hence, each generation consists of an electron-like particle, two quarks, and a neutrino. The statement that each generation behaves the same means that, among other things, the electric charges of all electron-like particles in the first column are -1 (in appropriate units); the electric charges of all quarks in the second column are $-\frac{1}{3}$ and all those in the third column $+\frac{2}{3}$. All neutrinos are electrically neutral.

We understand aspects of this horizontal pattern very well. In particular, various mathematical consistency conditions tell us that the particles must come in a collective of four particles, and their properties are largely fixed. In particular, we understand why the particles have the electric charges that they do: this is forced upon us by the mathematics and they simply can't be anything else. Moreover if, one day, we were to find a fourth species of electron-like particle, then we can be sure that there are also two further quarks and a neutrino to discover as well. We'll describe this more in Section 4.

We don't, however, understand the observed pattern of masses. More importantly, we don't understand the vertical direction in the pattern at all. We don't understand why there are 3 generations in the world and not, say, 17. Nonetheless, we know from both particle physics and from cosmological observations that there are no more than 3

light neutrino species. If there is an undiscovered fourth generation then the neutrino must be much much heavier (by a factor of about 10^{10}!) than the neutrinos in the existing generations. This is highly suggestive that the story stops at three.

The 12 particles listed above are all the "matter particles" that we have so-far discovered in the universe. Each has some fairly intricate properties that we will learn as these lectures progress. In particular, each particle has a corresponding anti-particle, and both the particles and anti-particles decompose further into "left-handed" and "right-handed" pieces. We will describe all this in Section 2.

Forces

All the particles that we described above interact through a handful of forces. It's usually said that there are four fundamental forces at play in the universe. In fact, by any logical count, we should say that there are five forces, with the interaction of the Higgs boson providing the fifth.

The traditional four forces of Nature are:

- Gravity: This was the first force to be discovered and, in many ways, the one we understand least. The effects of gravity are very familiar: it's the reason why apples fall from trees and tides wash in and out. It's the reason why planets orbit stars, and stars form galaxies, and the reason why these are all dragged inexorably apart by the expansion of the universe. Our best theory of gravity was given to us by Einstein: it is the theory of General Relativity.

- Electromagnetism: Like gravity, this force is familiar because it manifests itself in the macroscopic world, where it is harnessed for much of modern technology. On the atomic level, it is the electromagnetic force, acting between the electrons in an atom, that give rise to the chemical properties of the elements.

- Strong Nuclear Force: This force has no counterpart in classical physics. It is responsible for binding quarks together inside protons and neutrons and, subsequently, for binding protons and neutrons together as nuclei.

- The Weak Nuclear Force: Another force that manifests itself only on very small scales. Its primary role is to allow certain particles to decay into other particles. For example, beta radiation, in which a neutron decays into a proton, electron and anti-neutrino occurs because of the weak force.

The story above contains a little bit of a lie. At the fundamental level (meaning at the shortest distance scales), the force of electromagnetism should be replaced by

something called *hypercharge*. It's not dissimilar to electromagnetism, but it differs in details. What we observe as electromagnetism is some mix of hypercharge and the weak force.

Finally, the "fifth force" is:

- The Higgs Force: Again, a force which has no classical analog. Its role, however, is rather dramatic: it allows all the elementary particles described in the table above to get a mass.

Each of these five forces will be described in considerable detail as the lectures progress. Section 2 describes the matter particles and their interaction with electromagnetism; Section 3 describes the strong force; and Section 4 describes the weak force and the Higgs boson. Finally, we will turn to gravity in Section 5.

1.1 Quantum Fields

Ironically, the theory of particle physics is not a theory of particles. It's a theory of *fields*. A field is a fluid-like object which is spread throughout all of space. A field can take a different value at every point in space, and that value can change with time.

The most familiar examples are the electric and magnetic field which are associated to the electromagnetic force. These comprise of a pair of vectors, which exist at every point in the universe. Mathematically, the electric and magnetic field are functions $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$, which can take different values at different points $\mathbf{x}$ in space and points t in time. Like all fluids, the electromagnetic field can ripple. These ripples are what we call light waves.

Things get more interesting when we introduce quantum mechanics into the mix. In the 1920s, physicists understood that, on the smallest distance scales, the universe doesn't follow the common sense laws that Newton gave us. Instead, it's much more mysterious and counter-intuitive, and follows the rules of quantum mechanics. One of the key consequences of quantum mechanics is that energy isn't something smooth and continuous. Instead, energy can only be parcelled in discrete lumps. That's what the word "quantum" means: discrete, or lumpy.

The real fun happens when we try to combine the ideas of quantum mechanics with fields. One implication is that the electromagnetic waves that make up light are not continuous. Instead light is made up of particles called *photons*. The photons are ripples of the electromagnetic field tied into little parcels of energy due to quantum mechanics.

Figure 3. This is not what physicists mean by a field. It's what a farmer means by a field. Or a normal person.

The surprise is that the paradigm above holds for all other particles too. First, each of the forces described above has a field associated to it. And, when quantum mechanics is taken into account, ripples of the field become particles. The names of the particles associated to each of the forces are:

- Electromagnetism: As described above, the particle is the photon. In particle physics, photons are denoted by the greek letter γ (gamma). This comes originally from high-energy photons known as "gamma rays", but is now used to describe photons of any energy.
- Strong Nuclear Force: The field associated to this force is called the *Yang-Mills field*. The corresponding particles are *gluons*.
- Weak Nuclear Force: The field is another variant of the Yang-Mills type. The corresponding particles are the *W and Z bosons*.
- The Higgs Force: The associated particle is called, unsurprisingly, the *Higgs boson*. It was discovered at CERN in 2012, the last of the Standard Model particles to be found experimentally.
- Gravity: The force of gravity is rather special. Einstein's theory of general relativity teaches us that the gravitational field is actually space and time itself. Ripples of space and time are called *gravitational waves* and were first observed by the LIGO detector in 2015. The associated quantum particles, known as *gravitons*, have not been observed experimentally. Given the weakness of gravitational interactions, it seems unlikely that this situation will change any time soon.

So each force is associated to a field and an associated quantum particle. But, so too, are the matter particles. For example, spread throughout the room you're sitting in, and in fact throughout the entire universe, there exists something called the electron field. The ripples in this fluid get tied into little knots, or little bundles of energy, by the rules of quantum mechanics. And those bundles of energy are the particles that we call electrons. Every electron is a ripple of the same underlying field, like waves are ripples of the same underlying ocean.

There is also a muon field, and a tau field, together with six different kinds of quark fields and three kinds of neutrino fields. The Standard Model of particle physics is a theory describing how 12 matter fields interact with 5 fields of force. If one field — say the electron field – starts to move and sway, then it causes the gravitational field and the electromagnetic field to move. These, in turn kickstart the quark fields, and so on. All of these fields are engaged in an intricate, harmonious dance, swaying backwards and forwards, to a music that we call the laws of physics.

This can be contrasted with our view of classical physics. There we have two very different objects: particles and fields. At times, they make fairly awkward bedfellows. But there is a beautiful unification in the quantum world: everything is field. The particles are emergent objects.

The Implications of Quantum Fields

The world of quantum fields can, at times, be difficult to get our heads around. It is often easier to resort to the language of particles and, for the most part, this is what we will do in these lectures. Nonetheless, there are times when the particle picture breaks down and it is only when we think in terms of fields that things make sense. Here we describe a number of implications of the field theoretic picture. We'll see many more as the lectures progress.

Implication 1: First, and most importantly, field theories allow us to write laws of physics consistent with locality. If you shake an electron, it doesn't immediately affect a second electron sitting elsewhere. Instead the shaking electron produces a perturbation in the neighbouring electromagnetic field. This then propagates outwards, until it reaches the second electron. In this manner, there is no "action-at-a-distance", and causality is ingrained in the very structure of field theory.

Implication 2: The second consequence of quantum fields is the following: all particles of a given type are the same.

For example, two electrons are identical in every way, regardless of where they came from and what they've been through. The same is true of every other fundamental particle. Suppose, for example, that we capture a proton from a cosmic ray which we identify as coming from a supernova lying 8 billion lightyears away. We compare this proton with one freshly minted in a particle accelerator here on Earth. And the two are exactly the same! How is this possible? Why aren't there errors in proton production? How can two objects, manufactured so far apart in space and time, be identical in all respects? The answer is that there's a sea of proton "stuff" that fills the universe. This is the proton field or, if you look closely enough, the quark field. When we make a proton we dip our hand into this sea and mould a particle. It's not surprising that protons forged in different parts of the universe are identical: they're made of the same stuff.

Implication 3: The field perspective allows us to simply interpret situations where the number of particles changes. For example, beta decay is a process in which the neutron decays,

$$n \;\rightarrow\; p + e + \bar{\nu}_e$$

The decay products are a proton p, and electron e and an electron neutrino $\bar{\nu}_e$. (The bar on the $\bar{\nu}_e$ tell us that it's actually an anti-neutrino; we'll describe anti-matter in Section 2.) In terms of the underlying quarks, the down quark decays into an up quark

$$d \;\rightarrow\; u + e + \bar{\nu}_e$$

From a particle perspective, this is somewhat confusing. You might be tempted to view the decay above by thinking that a proton, electron and anti-neutrino are sitting inside the neutron, perhaps bound together by some mysterious force and just waiting to get out. (Or, equivalently, that the down quark is made of an up quark, electron and anti-neutrino.) But that's *not* the right way to think about the neutron (or the down quark.) Indeed, we've stated above that the down quark is a fundamental constituent of matter, and that would be difficult to believe if it contained other objects.

Happily, these confusions evaporate when we think in terms of fields. The particle that we call the down quark is a ripple of the down quark field. But there's no reason to think that this ripple can last forever. Instead, it can decay, but only at the cost of exciting three other fields: those of the up quark, electron and neutrino. The reason why these three particular fields get excited, and no others, is due to the detailed interactions of the various fields. These rules will be described as these lectures progress.

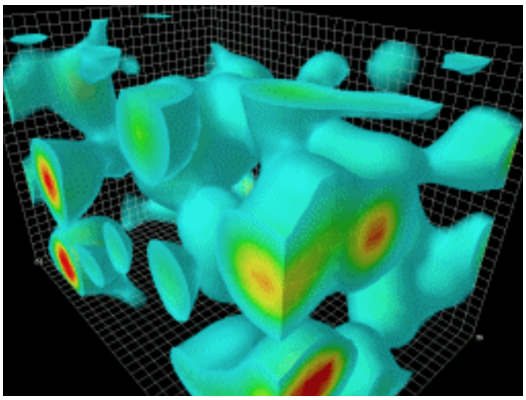


Figure 4. The vacuum of space is an interesting and complicated place. This picture is taken from the QCD simulation of Derek Leinweber

Implication 4: There is one last idea that will be useful to have in the back of your mind: the existence of quantum fields means that empty space, also known as the vacuum, is not a dull place. It is filled with quantum fields which, even when left alone, are not necessarily calm. An example is shown in Figure 4, depicting a computer simulation of empty space. What's shown is a typical configuration of the gluon field in the vacuum. The true vacuum is, in fact, much more complicated even than that shown in the picture. The vacuum doesn't have just a single field configuration but is something more murky: a quantum superposition of infinitely many different field configurations, each appearing with some probability. In quantum field theory, the vacuum of space is an interesting place. It froths with quantum uncertainty.

The take-home message for these lectures is that the vacuum of space is not some inert, boring substance. The bubbling fields breathe life into the vacuum and mean that it is able to respond to things happening within it. This phenomenon, as we shall see, lies at the heart of some of the more subtle effects of quantum fields.

1.2 Natural Units

Before we get going, we should introduce the units that we use to quantitatively describe the sub-atomic world. Usually in physics, we introduce different units for length, time and mass. For example, the SI units are meters, seconds and kilograms respectively. However, at the fundamental level these concepts are not as different as they first appear and the laws of physics provide a way to translate between them.

For example, there is a speed limit in place in the universe. No particle can travel faster than

$$c = 299792458 \text{ ms}^{-1} \approx 3 \times 10^8 \text{ ms}^{-1} \tag{1.1}$$

All particles with mass are obliged to travel slower than this speed, while all massless particles are obliged to travel at exactly this speed. The most familiar massless particle is the photon and for this reason c is referred to as the speed of light.

The speed of light allows us to translate freely between units of length and units of time. Given a length scale L, there is a natural time scale $T = L/c$, which is the time it takes light to cross the distance L.

In fact, the existence of a constant speed in the universe is hinting at something deeper: the concepts of space and time are not as distinct as we first thought. To put this in perspective, here's an analogy. I might decide that I'm going to measure all horizontal distances in centimeters, and all vertical distances in inches. I then proudly reveal a new fundamental constant of nature C which translates between the two,

$$C \approx 0.394 \text{ Inches cm}^{-1}$$

This is clearly a dumb thing to do. If I have a ruler marked in cm to measure horizontal distances, then I can always rotate it and use it to measure vertical distances in cm as well. The rotational symmetry in the world means that there is no fundamental difference between distances in the horizontal and vertical directions.

But exactly the same story holds for the speed of light. The theory of special relativity tells us that there is a symmetry between space and time, albeit one that only becomes apparent when you travel fast. If you move close to the speed of light you experience strange effects like time dilation and length contraction, which can be explained by a rotation of time into space and vice versa. What this means is that, at the fundamental level, we should measure time and space using the same units. If we choose to measure time in seconds, then we should measure length in light-seconds, the distance that light travels in a second. With this choice, the speed of light is simply

$$c = 1$$

A corollary of this is that mass is now measured in the same units as energy. This is because Einstein's famous formula $E = mc^2$ becomes simply $E = m$. In these lectures, we will specify the masses of elementary particles in units of energy. If, for some reason, you want to get the mass in, say, kilograms, then you simply need to reinstate the factor of the speed of light in $m = E/c^2$ and use the value $c \approx 3 \times 10^8$ ms^{-1}.

A similar story arises when we consider quantum mechanics. The fundamental constant in quantum mechanics is Planck's constant,

$$\hbar = 1.054571817 \times 10^{-34} \text{ Js}$$

It has units of energy $\times$ time. What this constant is really telling is us that, at the fundamental level there is a close connection between energy and time. A process with energy E will typically take place in a time $T = \hbar/E$. In this way, we can translate between units of energy and units of time, and these concepts are not as distinct as our ancestors believed. To highlight this, we choose units so that

$$\hbar = 1$$

The choice $c = \hbar = 1$ is referred to as *natural units*. It means that there's only one dimensionful quantity left, which we usually take to be energy. Any measurement — whether it's of length, time or mass — can be expressed in terms of energy.

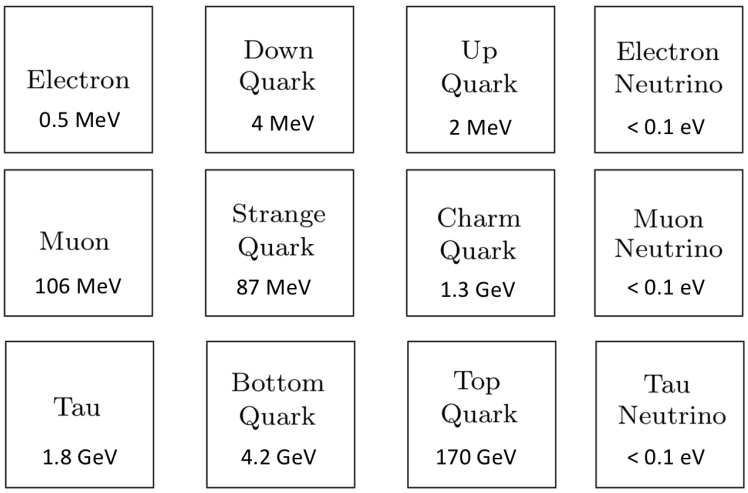
A Sense of Scale

The SI unit of energy is a Joule and is not particularly appropriate for the sub-atomic world. Instead, we use the electronvolt (eV) which is the energy an electron picks up when accelerated across 1 volt. This is

$$1 \text{ eV} \approx 1.6 \times 10^{-19} \text{ J}$$

The electronvolt is the energy scale appropriate for atomic physics. For example, the energy that binds an electron to a proton to form a hydrogen atom is $E \approx 13.6$ eV. For elementary particles, we will need a somewhat larger unit of energy. We typically use MeV $= 10^6$ eV or GeV $= 10^9$ eV. The LHC, our best current collider, runs at an energy scale measured in TeV $= 10^{12}$ eV.

The masses of the 12 matter particles cover a range from eV to GeV. They are:



The entries for neutrinos are upper bounds on the mass. Meanwhile, the masses of the 5 force-carrying particles are

For each of these particles, there is an associated length scale. We get this by transforming energy E into a length using the fundamental constants of Nature c and $\hbar$,

$$\lambda = \frac{\hbar c}{E} \tag{1.2}$$

This is known as the *Compton wavelength*. Roughly speaking, it can be viewed as the size of the particle. For example, for the electron the Compton wavelength is $\lambda_e \approx 2 \times 10^{-12}$ m. Perhaps somewhat surprisingly, the heavier a particle is, the smaller its Compton wavelength.

The Biggest and the Smallest

There are two further length scales that we should mention before delving into details of the subatomic world. One is associated to the strength of the gravitational force. Newton's constant is given by

$$G_N \approx 6.67 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}s^{-2}}$$

In natural units, Newton's constant has dimensions of [Energy]$^{-2}$. Putting back the factors of $\hbar$ and c, we can derive an energy scale known as the Planck mass

$$M_{\rm pl} = \sqrt{\frac{\hbar c}{8\pi G}} \approx 2 \times 10^{18} \text{ GeV}$$

(The factor of 8π is conventional, and is sometimes dropped.) This is an enormous energy scale, fifteen orders of magnitude larger than the scales that appear in the Standard Model. The corresponding length scale is the Planck length

$$L_{\rm pl} = \sqrt{\frac{8\pi\hbar G}{c^3}} \approx 8 \times 10^{-35} \text{ m}$$

This, in turn, is a tiny length scale, 15 orders of magnitude smaller than the scale that we have explored at our best particle colliders. The Planck scale is where both quantum mechanics and gravity become important. It seems likely that space and time cease to make sense at these scales, although it's not clear exactly what this means.

On the other end of the spectrum, the largest size that we can talk about is the entire observable universe,

$$L_{\rm universe} \approx 9 \times 10^{26} \text{ m}$$

The corresponding energy scale is

$$H \approx 2 \times 10^{-33} \text{ eV}$$

This energy scale is closely related to the so-called "Hubble constant" that measures how fast the universe is expanding, albeit written in the unusual units of electronvolts. Clearly, it's a tiny energy scale. A particle with the mass of H would have the same size as the entire universe.

This, then, is the playground of physics. The goal of physics is to understand everything that can happen at lengths in the range

$$10^{-34} \text{ m} < L < 10^{26} \text{ m}$$

or, equivalently, at energies in the range

$$10^{-33} \text{ eV} < E < 10^{27} \text{ eV}$$

It turns out that there is quite a lot of interesting things in this window! And, underlying many of them, is the Standard Model.

A Interlude: The Road to Discovery

Throughout these lectures, our focus will be on explaining the key ideas and concepts that underlie the subatomic world. This means that we will describe our current understanding viewed through the lens of the Standard Model. Yet this theory took many decades to build, years in which physicists were mostly bewildered and confused. It's important to ask: how did we arrive at this point?

The answer to this question is almost entirely: experiment. While there were many groundbreaking theoretical observations, at each step the important progress was only made in response to a novel experimental discovery. After each section of these lectures, we will have a short interlude in which we describe this experimental progress. This will also provide an opportunity to present a more historical approach to the subject of particle physics. The hope is that these interludes will serve to ground some of the ideas that we meet in the main thread, and place them in more concrete context.

In this first interlude, we explain the beginnings of particle physics. We describe the discoveries that resulted in the comforting and familiar picture of matter as made of just three particles: the electron, proton and neutron.

A.1 Ray Physics

Before particle physics was the study of fields, it was the study of rays. By the end of the 1800s, several kinds of "rays" had been found, each seemingly exhibiting different properties. It took several decades to distill the properties of these rays (do they undergo diffraction? can they be bent by electric or magnetic fields?) and, ultimately, to understand their particle constituents.

Cathode Rays

Take a glass tube, and remove most of the air. If you drop a large voltage across the tube, a faint glow can be seen at one end, a result of rays emitted by the cathode (the negatively charged electrode) hitting the glass wall. In many ways the discovery of these cathode rays, first observed in 1869 by the German physicists Hittorf and Plücker, marks the beginning of modern day particle physics.

A number of properties of cathode rays were soon established, including the fact that they travel in straight lines, as revealed by the shadow cast by objects placed in their path, but their trajectory can be deflected by both electric and magnetic fields. We now know that cathode rays are beams of accelerated electrons.

With hindsight, these early discharge tubes can be viewed as the world's first particle accelerators. The electrons were accelerated to energies of around 10^5 eV. In contrast, in the last electron-positron accelerator at CERN, known as LEP, electrons were accelerated to energies of around 2×10^{11}eV. Its successor, the LHC, accelerates protons to energies of 6×10^{12} eV. Evidently, the increase of energy by 7 orders of magnitude took 150 years of hard work. There are reasons to believe that the next 7 orders of magnitude will be even more challenging.

X-Rays

One recurring theme of particle physics is how one discovery paves the way for the next. The first example occurred in November 1895, when Wilhelm Röntgen was playing with cathode rays. He covered the glass tube with thin black cardboard and noticed that, when the tube was turned on, a paper screen painted with a chemical called barium platinocyanide would give a faint green glow, even when placed up to a meter away from the apparatus. He concluded that the tube was emitting some invisible rays, which he dubbed *X-rays*.

Röntgen's subsequent investigations showed that X-rays were not the same as the cathode rays from which they originated since they travelled much further in air. He also realised that X-rays could be used to develop photographic plates, and his first paper includes the astonishing photograph of his wife's hand shown to the right. His wife's perfectly reasonable response: "I have seen my death".

Hand des Anatomen Geheimrath von Kölliker. Im Physikalischen Institut der Universität Würzburg mit X-Strahlen aufgenommen von Professor Dr. *W. C. Röntgen.*

Röntgen's paper resulted in enormous excitement, both among other scientists and the general public. Rather like Einstein in later years, everyone wanted a piece of Röntgen. However, it appears that he was less than impressed with the whole show and worked hard to stay out of the limelight. A sole newspaper interview from this time reveals a Dirac-like level of brevity:

> *Interviewer:* What did you think?
> *Röntgen:* I did not think; I investigated.
> *Interviewer:* What is it?
> *Röntgen:* I don't know.

The nature of X-rays remained mysterious for a long time. The suggestion that they may be short wavelength electromagnetic waves was not generally accepted since no one

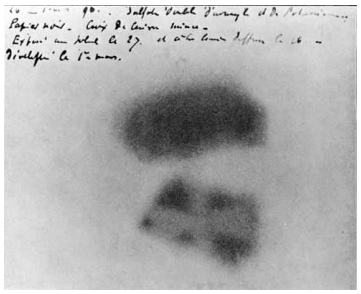
succeeded in observing X-ray diffraction. This changed in 1912, when Max Laue (later von Laue) realised that the crystal lattice of solids had the right separation needed to observe diffraction of X-rays, opening up the entire field of X-ray crystallography.

Uranic Rays

Among the many people to be inspired by Röntgen's discovery was a French physicist named Henri Becquerel. For some years, he had held a fascination with phosphorescent materials and decided to explore the connection to X-rays. Ironically, in the end his discovery had nothing to do with X-rays.

Uranium salts had long been known to have phosphorescent properties. Becquerel's experiment involved exposing uranium to sunlight for several hours. With the uranium suitably excited, he observed that it emitted rays which, rather like X-rays, created silhouetted images on photographic plates wrapped in thick, black paper.

Becquerel's breakthrough came because of the weather. The final days of February 1896 were dark and overcast in Paris. With no sunlight available, Becquerel stored his experiment in the drawer of his desk and got on with other things, like a spot of shopping. By March 1st the sky remained cloudy and, perhaps bored, perhaps inspired, Becquerel decided to develop the photographic plate anyway, expecting to find nothing. Instead, to his astonishment, there appeared the clear image of a copper cross that had been placed between the plate and the uranium source. Becquerel's photograph is shown to the right.



Becquerel's desk drawer discovery showed that all his careful preparation, exposing uranium to sunlight, had nothing to do with the uranic rays that it emitted. This was, to say the least, disconcerting. If the rays were emitted without any prompting from an external energy source like the Sun then where did their energy come from? It was tempting to think that the whole thing violated the conservation of energy.

α, β and γ Rays

Soon after Becquerel's breakthrough, three further radioactive elements – thorium, polonium and radium – were discovered by Marie Curie (née Skłodowska) and her husband Pierre. The next step was to try to characterise the rays that were emitted. This was done by one of the early heroes of particle physics: Ernest Rutherford.

Rutherford was the first to realise that uranium emitted not one, but two different types of radiation. He called these α-rays and β-rays:

- α-rays: Alpha radiation is easily absorbed. Indeed, we now know that all α-rays would have been absorbed by the black paper wrapping the photographic plate in Becquerel's original experiment and so were not detected. Over a period of 13 years, from 1895 to 1908, it slowly became clear that α particles are four times heavier than hydrogen, with charge +2. In other words, they are what we now know to be the nucleus of a helium atom.

- β-rays: Beta radiation is more penetrating. In time, it was identified with a stream of electrons. This will be described below.

Later, Villard discovered that radium emitted yet a different kind of radiation, one that he naturally called...

- γ-rays: Another, very penetrating form of radiation. This was later found to be highly energetic electromagnetic waves.

Around the turn of the century, the situation was rather confusing. There were cathode rays and x-rays and α and β and γ rays. not to mention a number of false discoveries that were purely illusory. Getting to the heart of these phenomena would, ultimately, require an understanding of electromagnetism and the strong and weak nuclear forces.

A.2 The Electron

Cathode rays were the first to be discovered and, subsequently, the first to be understood in terms of a constituent particle. This particle is, of course, the electron.

The claim that rays are composed of particles means that their properties can be understood using Newtonian mechanics in which the particles are endowed with a mass m and electric charge e. In 1885, Hendrik Lorentz wrote down the equation of motion for such a particle moving in the presence of an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$. A particle with velocity $\mathbf{v}$ will experience an acceleration $\mathbf{a}$ given by

$$m\mathbf{a} = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{A.1}$$

The first goal was therefore to measure the deflection of the cathode rays due to electric and magnetic fields. This measurement was performed by J.J. Thomson in 1897. Two other physicists, Wiechert and Kaufmann made similar measurements using only magnetic fields around the same time. However, as you can see from the equation of motion, the deflection of the rays cannot tell us about m and e individually: it only tells us about the ratio e/m.

Fortuitously, the same ratio e/m had arisen in an entirely different context the year before. This came from the discovery that the atomic spectral lines can be split in the presence of a magnetic field, a phenomenon known as the *Zeeman effect*. Lorentz himself analysed this effect using the equation (A.1) and, happily, the value of e/m that was needed to explain the observed splitting was close to that later found in cathode rays.

But the ratio e/m for cathode rays came with a surprise: it was significantly larger than the value for other known ions. (Thomson's measurement of e/m gave a value that was 770 times larger than that of a hydrogen ion – the particle that would later be rebranded as a proton. We now know that the correct value of the ratio is around 1836.) But the question remained: is this because the electric charge e of the particle is very big, or because the mass m is very small. To resolve this issue, one had to find a way to measure either e or m individually.

The prevailing viewpoint at the time was the right one: that the mass of the particle was unusually small. Indeed, there was already some indication of how small it had to be, since the electric charge on a hydrogen ion had been estimated to reasonable accuracy. This in itself was no mean feat. It was fairly straightforward to measure the electric charge on, say, a mole of ions. The difficulty is in figuring out how many atoms are in a mole. Or, in other words, in figuring out Avogadro's number $N \approx 6 \times 10^{23}$. (See the final question on this Statistical Mechanics Example Sheet if you want to challenge yourself.) A number of ingenious ways to determine N were proposed, resulting in a ballpark figure for e, the minimum unit of electric charge carried by what we now call the proton. One of the best estimates (out by a factor of 20 or so) was proposed by the Irish physicist Stoney, who also coined the name *electron*, for this "atom of electricity".

A more direct measurement of the electric charge was first achieved by J.J. Thomson, and this is the reason that he, rather than Wiechert or Kauffman, is primarily remembered as the discover of the electron. He didn't study cathode rays, but he turned instead to the *photoelectric effect*. This occurs when UV light is shone on a material, causing electrons to be emitted, but at much slower velocities than those in cathode rays. Thomson measured the ratio e/m of these emitted particles and found that it agreed with his earlier measurements of cathode rays. But this time he could go further, employing a preliminary version of a detector known as the *cloud chamber*. This chamber contains supersaturated vapour which condenses into little droplets, or clouds, as a ray passes through and ionises the atoms. In this way, the path of the emitted object can be tracked.

In 1899, with his new cloud chamber toy in hand, Thomson was able to determine the number of negatively charged ions that formed due to photoelectric emission simply by counting the droplets along the path. He was also able to determine the overall electric charge by counterbalancing the effect of gravity with an electric field. In this way, he measured the charge on each individual droplet, getting within 30% of the electric charge that we know today. This was the first time that the cloud chamber gave rise to a major breakthrough in physics. It would not be the last.

A more precise measurement, employing a similar technique but using oil droplets rather than a cloud chamber, was performed in 1909 by Millikan and Fletcher. They again balanced gravitational and electric forces, but were able to observe individual droplets, deducing that the charge was always quantised in units of 1.6×10^{-19} Coulombs. Their measurement was within less that 1% of the modern value. Famously, Millikan struck a dubious bargain with his student Fletcher and the resulting paper was published under Millikan's name alone. No doubt he felt bad as he collected his Nobel prize twelve years later.

A.3 The Proton

The discovery of the proton went hand in hand with the discovery of the nucleus itself. The key experiment is one of the most famous in physics. Working in Manchester in 1909, Hans Geiger and his undergraduate assistant Ernest Marsden were firing alpha particles at thin sheets of metal and measuring how they bounced off. One day Rutherford entered the room and, according to Marsden, suggested "see if you can get some effect of α-particles directly reflected".

The results were startling and entirely unexpected. About 1 alpha particle in 8000 was reflected back in the direction from which it came. In later years, Rutherford recounted his surprise in a well known quote:

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

But what did it mean?

The prevailing theories for the structure of the atom could not account for these experiments. Of course, physicists knew that atoms contained electrons, and there was acceptance that there had to be a compensating positive charge, but theories of the structure of the atom – whether based on plum puddings, planetary systems, or vortices – were put forward with little evidence.

The Gieger-Marsden experiment held the key. That it was Rutherford himself who understood its consequences is, in some ways, rather surprising. Rutherford was not known to hold much love for theoretical physics. He is reported to have said of relativity, "Oh, that stuff. We never bother with that in our work." and he was never a strong proponent of Bohr's founding work on quantum theory. My favourite Rutherford quote, capturing both his attitude to theorists, and his personality, is

> "How can a fellow sit down at a table and calculate something that would take me – *me* – six months to measure in a laboratory?"

Nonetheless, when needed, Rutherford showed himself to be no mean theorist. In 1911, he postulated that each atom contains a heavy, almost point-like object at the centre, carrying positive charge Q. He computed that an α-particle, repelled by the electrostatic Coulomb force, would be deflected by an angle θ with probability

$$\sigma(\theta) = \frac{Q^2 q^2}{4m^2 v^2 \sin^4(\theta/2)} \tag{A.2}$$

Here q, m and v are the charge, mass and velocity of the α-particle. This formula is now known as the Rutherford cross-section. You can find a derivation in the lecture notes on Dynamics and Relativity. The formula agrees with the experimental data with impressive accuracy.

As with many great discoveries in science, there was no small amount of luck involved. On the experimental side, the alpha particles used in the experiment were fast enough to blast through the electrons of the atom without care, but slow enough to be deflected from the nucleus before they experienced the strong nuclear force. On the theoretical side, Rutherford deduced his formula using Newtonian classical mechanics. But the correct calculation of the cross-section requires quantum mechanics and in nearly all cases this differs from the classical result. The Coulomb force law turns out to be the exception – it is the one force where classical and quantum results for scattering agree! (You can learn more about this in the scattering theory section of the lectures on Topics in Quantum Mechanics.)

With Rutherford's explanation of the nucleus, there was still work to be done. At the time, all elements were labelled by their atomic weight A which, at least for light elements, was close to an integer. Listing all the elements in order then gives two numbers: A, their weight and Z, their place in the list. The first few elements are shown in Table 1.

	H	He	Li	Be	B	C	N	O
Z	1	2	3	4	5	6	7	8
A	1	4	7	9	11	12	14	16

Table 1. The first few elements

But the question remained: what is the physical meaning of Z? In particular, if you're just given the list of known elements, how do you know you've got them all? For example, it might be tempting to think that A is really telling us the rightful place in the list, with two missing elements with $A = 2$ and 3 that are just waiting to be found.

The suggestion that Z should be equated with the positive electric charge of the nucleus did not have an auspicious start. It was first made by van den Broek, a Dutch real estate lawyer who was pushing some new and improved 3d version of Mendelev's periodic table. His accounting of the elements never took off, but his suggestion that Z measures the electric charge was rapidly accepted. Moreover, Rutherford's scattering formula (A.2) gives a clear way to measure the charge Q of the nucleus. These, and other experiments, ultimately led to the complete periodic table that we know and love today[2].

From here it was a short step to the idea that the hydrogen nucleus – originally called the H-particle – was a building block of all nuclei. The issue was finally put to rest some years later when Rutherford demonstrated that H-particles were emitted by other nuclei – specifically nitrogen – when bombarded by α-particles. The name "proton" was coined by Rutherford at this time.

A.4 The Neutron

In 1911, the nucleus was discovered. By 1920, there was no doubt about the existence of the proton. But it took until 1932 for the other constituent of the nucleus, the neutron, to be found.

[2]A physicist called Henry Moseley deserves special credit here. When Mendelev originally proposed his periodic table he found gaps, allowing him to successfully predict the existence of three new elements. Moseley, using X-ray techniques to determine the atomic number Z, successfully predicted seven more!

Moseley, like many physicists of that time, was drafted into the Great War. Geiger and Marsden, for example, found themselves both posted to the Western front, fighting on opposite sides. Both survived. Moseley was not so lucky and the man widely recognised as England's most promising young physicist was killed in the battle of Gallipoli by a bullet in the head. He was 27 years old.

In the decade between the discoveries of the proton and neutron, physics changed beyond anyone's wildest imagination. Quantum mechanics was formulated and the basics of quantum field theory were laid down. These ideas provide the foundation for nearly everything that we discuss in these lectures – both experimental and theoretical. Yet the discovery of the neutron owed only little to these new developments. The neutron took so long to find simply because it's hard to see. For this reason, the neutron still carries an air of that pre-quantum world, less exotic than, say, anti-matter whose origin story is so closely tied to developments in quantum theory. Yet, astonishingly, there was less than a 100 days between the discovery of the two particles!

You might wonder why physicists didn't stay up at night, puzzled by the difference between the atomic number Z and the mass A of the nucleus. It's because they had a very convincing explanation. They thought that the nucleus must consist of A protons with $A-Z$ electrons to cancel the charge. Moreover, there was an *extremely* good reason to think that the nucleus contained electrons. This was beta decay. The electrons emitted in beta decay are far more energetic than the orbiting electrons in the atom, and this meant that they had to have their origin in the nucleus. But if electrons were being emitted from the nucleus, then obviously they must have been there all along. That's simply common sense. Of course, we now know that common sense isn't always the best guide when it comes to the sub-atomic world.

If you knew where to look, the advent of quantum mechanics did make it increasingly difficult to believe in electrons in the nucleus. Trapped inside a cell the size of a nucleus, the Heisenberg uncertainty relation means that the electron necessarily has energy greater than 40 MeV, significantly larger than nuclear binding energies and making it untenable that the electron could remain in place. Further trouble came with the discovery of spin. (We describe spin in more detail in Section 2.1.) If both the proton and electron have spin 1/2 then, regardless of whether spins add or subtract, a nucleus with A protons and $A-Z$ electrons should have integer spin when Z is even and half-integer spin when Z is odd. But that's not what's seen. Nitrogen, for example, has $Z=7$ but was long known to have integer spin. Opinions differed on what to make of this. So ingrained was the idea that the nucleus contains protons and electrons that Fermi and Rasetti even wrote a paper suggesting that the mismatch should cast doubt on the idea of spin.

Still, when the breakthrough came it owed essentially nothing to new-fangled quantum ideas and everything experiment. The first hint that something new was afoot came in 1930 in Berlin. Walther Bothe and Herbert Becker took alpha rays from a polonium source and directed them on beryllium. They found that the beryllium gen-

erated a new radiation of great penetrating power which they concluded, incorrectly, must be gamma rays. Over the next couple of years the experiment was repeated and improved, notably by Iréne Curie who was sitting on the world's most powerful source of polonium, a gift from her mother. Together with her husband Fréderic Joliot (by that time both doubled-barrelled Curie-Joliot's), they directed this beam of supposed gamma rays at parafin and found that it could eject protons at huge velocities. But still they stuck with the gamma ray interpretation.

The Curie-Joliot experiment was the watershed moment. Their interpretation was not, it's fair to say, universally embraced. Apparently the Italian physicist Ettore Majorana responded to the news with the exclamation

> "What fools! They have discovered the neutral proton, and they do not recognise it!"

In Cambridge, Rutherford and James Chadwick, his second-in-command, held similar sentiments. There was too much amiss for the radiation to be gamma rays. Less than three weeks later, Chadwick discovered the neutron.

In fairness, Chadwick had been searching for something like a neutron for over a decade. He didn't originally envisage a new elementary particle, but instead a closely knit bound state of a proton and electron, much smaller than a hydrogen atom so that it could fit inside the nucleus. That meant he was well prepared when the Curie-Joliot result came in. His short paper studies the penetrating power of the radiation. The Bothe-Becker-Curie-Joliot interpretation was that the original alpha rays react as

$$^{9}\mathrm{Be} + \alpha \longrightarrow {}^{13}C + \gamma$$

But the properties of the carbon nucleus were known well enough to put an upper bound on the energy of the emitted gamma ray. Whatever was coming out of this reaction was much more powerful. Chadwick found the correct conclusion: he was seeing something entirely new

$$^{9}\mathrm{Be} + \alpha \longrightarrow {}^{12}C + n$$

Chadwick had found the neutron.

2 A First Look at Quantum Fields

In this section we look in more detail at some of the key features of quantum fields and their interactions. We illustrate these properties with the simplest quantum force, electromagnetism. Or, to give it its fancy name, *quantum electrodynamics*.

2.1 Matter Fields and Force Fields

We'll meet a bewildering number of names in these lectures, each of them classifying particles according to various properties. But one classification is more important than all others: every type of particle falls into one of two classes called

- Bosons
- Fermions

The distinction between these two kinds of particles lies in the quantum world. Fermions have the property that no two particles can occupy the same quantum state. Roughly speaking, this means that you can't put two fermions on top of each other. This property is known as the *Pauli exclusion principle*. In contrast, there is no such restriction on bosons. You can pile up as many of them as you like, one on top of the other.

(A mathematical aside: if you've done a little quantum mechanics then it's very easy to describe the difference between bosons and fermions. Two identical particles are described by a wavefunction $\psi(\mathbf{x}_1, \mathbf{x}_2)$ which tells you the probability amplitude to find the two particles at positions $\mathbf{x}_1$ and $\mathbf{x}_2$. If the particles are bosons then, when you swap their positions, the wavefunction remains unchanged: $\psi(\mathbf{x}_2, \mathbf{x}_1) = \psi(\mathbf{x}_1, \mathbf{x}_2)$. In contrast, if the particles are fermions then the wavefunction picks up a minus sign when you swap them: $\psi(\mathbf{x}_2, \mathbf{x}_1) = -\psi(\mathbf{x}_1, \mathbf{x}_2)$. This means, in particular, that if you try to bring the two particles together at some point $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$, then $\psi(\mathbf{x}, \mathbf{x}) = 0$ for fermions, so there is vanishing probability that the two particles sit on top of each other.)

All the matter particles in the universe – electrons, quarks and neutrinos – are fermions. All the force carrying particles are bosons. In fact, this is more or less a definition of what we mean by a "matter" particle vs a "force" particle. The matter particles obey the Pauli exclusion principle; the force particles do not.

The distinction between bosons and fermions has a couple of familiar consequences. Electrons are fermions and therefore obey the Pauli exclusion principle. This is ultimately responsible for the structure of the periodic table. The electrons can't all sit

close to the nucleus, but fill up successive atomic shells, with the electrons in the outer shell — known as valence electrons — largely responsible for the chemical properties of the element.

Photons are bosons, and this means that the Pauli exclusion principle does not apply. A laser is an example of a system in which many photons sit in the same quantum state.

2.1.1 Spin

Particles are endowed with a number of other properties. The most familiar of these is the mass of the particle, but it is not the only one.

Particles also have an inherent angular momentum that we call *spin*. It's not a bad analogy to think of elementary particles as spinning about some axis, much like the Earth spins. But spin is a quantum mechanical property and if you push the spinning-Earth analogy too far then it breaks down. For example, if you ask questions like "how fast is the surface of the particle moving" then you'll get nonsensical answers. Furthermore, there's no way to spin up a particle like a basketball; the magnitude of the spin is something that is fixed and unchanging. You can, however, change the orientation of the axis along which the particle spins.

Like many phenomena in the atomic world, spin is quantised. That means that the spin can't take arbitrary values, but comes in discrete amounts. These are[3]

$$s = 0,\ \frac{1}{2}\hbar,\ \hbar,\ \frac{3}{2}\hbar,\ 2\hbar, \ldots \tag{2.1}$$

In natural units, we just say that a particle has spin 0, or spin $\frac{1}{2}$, and so on. Each particle in nature has a spin with a value taken from this list.

Particles that have a half-integer spin come with a rather strange property. If you rotate them by 360° then they don't quite come back the same as they were before! Instead, their quantum wavefunction comes back to minus itself. This means that you have to rotate the particle by 720° before it comes back to the same state. This is one of the more surprising facts about elementary particles and is a clear departure from our every day experience with classical objects.

[3]I've been a little bit sloppy here. Strictly speaking, the total spin of the particle is $\sqrt{s(s+\hbar)}$ with s taking one of the values listed in (2.1).

There is a deep theorem, originally framed by Pauli, which states that the spin determines whether a particle is a boson or fermion:

The Spin-Statistics Theorem: Particles with integer spin are bosons. Particles with half-integer spin are fermions.

This theorem follows when you combine the laws of quantum mechanics with the rules of special relativity. (The word "statistics" in the name of the theorem is not particularly helpful. Its origin lies in the fact that you get different answers when you count the number of possible states in which bosons or fermions can sit, and this counting is referred to as the "statistics" of the particle. We won't need this interpretation in these lectures. You can learn more in the lectures on Statistical Physics.)

The spins of all the known elementary particles in Nature are:

- Spin 0: The Higgs Boson.
- Spin $\frac{1}{2}$: All matter particles, i.e. the electron, muon and tau, together with the six types of quarks and three neutrinos.
- Spin 1: The photon, gluon and W and Z bosons. In other words, the particles associated to electromagnetism and the weak and strong nuclear forces.
- Spin 2: The graviton.

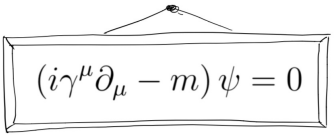
The remaining properties of particles mostly specify their interactions under the various forces. A familiar example is the electric charge, which determines the strength of a particle's interaction with electromagnetism. We'll describe the electric charge of all particles in Section 2.2, and the interactions with other forces in subsequent sections.

Finally, all of the properties described above, including the fermionic/bosonic nature of the particle, are really properties of the underlying field, which are subsequently inherited by the particle.

2.1.2 The Dirac Equation

All fields with spin $\frac{1}{2}$ — which, as we've just seen, means all fields associated to matter particles — are described by the *Dirac equation*.

We won't explain the mathematics behind the Dirac equation in these lectures, but it's so beautiful that it would be a shame not to show it to you. I've put it in a picture frame to highlight that it's here for decoration as much as anything else.



Here ψ is the quantum field; it depends on space and time. It also has four components, so it's similar to a vector but differs in a subtle way. It is known as a *spinor*. For what it's worth, the parameter m is the mass of the particle, while ∂_μ denotes derivatives and γ^μ are a bunch of 4×4 matrices. If you want to understand what the Dirac equation really means, you can find details in the lectures on Quantum Field Theory.

Dirac originally wrote his equation to describe the electron. But, rather wonderfully, it turns out that this same equation describes muons, taus, quarks and neutrinos. This is part of the rigid structure of quantum field theory. Any particle with spin $\frac{1}{2}$ must be described by the Dirac equation: there is no other choice. It is the unique equation consistent with the principles of quantum mechanics and special relativity.

The Dirac equation encodes all the properties of particles with spin $\frac{1}{2}$. Given such a particle, once you fix the orientation of the spin there are two possible states in which the particle can sit. Roughly speaking, it can spin clockwise or it can spin anti-clockwise. We call these two states "spin up" and "spin down".

The Pauli exclusion principle states that no two fermions can sit in the same quantum state. But the quantum state is determined by both the position and the spin of the electron. This means that an electron with spin down can be in the same place as an electron with spin up, because their spins differ. If you've done some basic chemistry, this should be familiar: both the hydrogen and helium atoms have electrons sitting in the orbit that sits closest to the nucleus. The two electrons in helium necessarily have different spins to satisfy the exclusion principle. But, by the time you get to the third element in the periodic table, lithium, there is no longer room for an additional electron in the closest orbit and the third electron is forced to sit in the next one out.

2.1.3 Anti-Matter

The real pay-off from the Dirac equation comes when you solve it. The most general solution has an interesting property: there is a part which describes the original particles, like the electron. But there is a second part that describes particles with the same

mass but with the opposite electric charge. The electron has negative charge, so these other particles must have positive charge. These positively charged electrons are called *positrons*: they are examples of *anti-matter*.

If a particle and anti-particle collide, both are annihilated. Typically, the energy is released in high energy photons. We denote the electron as e^- and the positron as e^+. Their annihilation usually results in the emission of two photons. (The emission of a single photon is not consistent with the conservation of both energy and momentum. For example, in the centre of mass frame, conservation of momentum would mean that the emitted photon would have nowhere to go.) The annihilation process is described by the reaction

$$e^- + e^+ \to \gamma + \gamma$$

The end result of all this is that the Dirac equation actually describes four different types of single particle states: a particle with either spin up or spin down, and an anti-particle with either spin up or spin down. The fact that there are four such states is related to the fact that the field ψ is a vector-like object with four components.

Dirac wrote down his equation in 1928. After a few years of confusion, Dirac himself suggested that these novel solutions should be interpreted as anti-matter. In 1931, he wrote

> "A hole, if there were one, would be a new kind of particle, unknown to experimental physics, having the same mass and opposite charge to an electron."

This bold proposal, was confirmed experimentally just one year later, a development that we will describe in more detail in Section B.) The prediction of anti-matter remains one of the great triumphs of theoretical physics. We now know that all the matter particles in Nature have corresponding anti-particles. In all cases, the conserved charges of the anti-particles are equal and opposite to those of the particles.

The Fallacy of the Dirac Sea

Although Dirac's genius led him to predict anti-particles, the argument that got him there was somewhat flawed.

Dirac's mistake was to misinterpret the meaning of ψ in his equation! He originally wrote down the Dirac equation as a relativistic generalisation of the Schrödinger equation, with ψ viewed as the wavefunction of a single particle. We now know that this is not the right interpretation: ψ should be viewed as a quantum field, whose excitations describe many particles, rather than the wavefunction for a single particle.

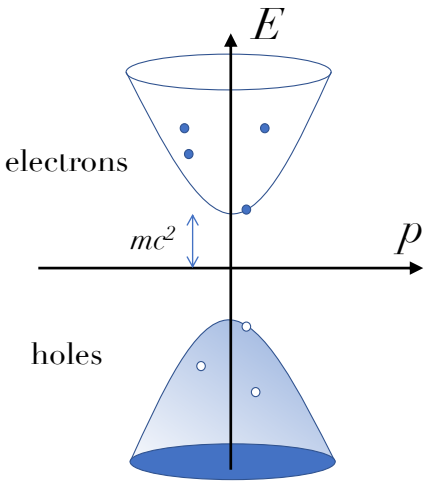


Figure 5. The filled Dirac sea, shown in blue, with holes interpreted as anti-matter.

We can explain what's going on here in a little more detail. The famous Einstein equation $E = mc^2$ tells us the energy of a particle of mass m when the particle is at rest. If the particle is moving with momentum p, then the correct formula is

$$E = \sqrt{p^2c^2 + m^2c^4} \tag{2.2}$$

where, for once, we've left the factors of c in the equation rather than setting $c = 1$. If you solve the Dirac equation, with ψ viewed as a wavefunction, then you find the two sets of solutions, but with energy

$$E = \pm\sqrt{p^2c^2 + m^2c^4}$$

The positive energy solutions are identified as, say, electrons. But what to do with the negative energy solutions? Note that as the particle moves faster, so p increases, and the negative energy solutions become more and more negative. This is problematic. If something can lower its energy then it usually does. But clearly we don't observe any particles getting faster and faster.

Dirac found a clever, but not altogether convincing, trick to escape from this conclusion. He suggested that the negative energy states were already filled by electrons. Because electrons are fermions, the Pauli exclusion principle means that no other electron is allowed to sit in these states, blocking the possibility for electrons to lose energy by tumbling to ever-lower states. This situation is shown in Figure 5.

In this picture, the vacuum of the universe consists of an infinite number of electrons. This is called the *Dirac sea*. One might worry about why we don't feel the infinite electric charge, but since this situation represents the ground state of the universe, Dirac argued that we reset the clock and say that this is what we mean by neutral. Any charge is now measured relative to this ground state.

Dirac's picture suggests that something novel may happen. We could excite an electron out of the Dirac sea, and into the positive energy states. This is shown in Figure 5. To do this, we need to inject a minimum of $E = 2mc^2$ energy into the system, to bridge the gap between the lower and upper bands in the figure. But if we can somehow achieve this, then we have created an electron out of the vacuum. But we have also created a *hole*, an absence of an electron, in the sea of negative energy states. In a zen-like manoeuvre, we now attribute properties to this absence. Like the excited electron, it can freely move around. Because there's an absence of charge, relative to the vacuum it will appear to have positive charge. Finally, if the electron and the hole come into contact, the electron can drop back down into the negative energy state. It will appear as if the electron and hole have annihilated, releasing at least energy $E = 2mc^2$ in the process.

Dirac's picture of anti-particles is ingenious. But, ultimately, it's not the right way to think about things. If you view the object ψ not as a single-particle wavefunction, but rather as a quantum field, then the energy of both particles and anti-particles turns out to be positive. There is no need to invoke an infinity of electrons, disappearing to the bottom of the sea. Instead, there are no negative energy states: simply particles and anti-particles.

Moreover, it turns out that bosons also have anti-particles. But now there is no counterpart to the Dirac sea argument because bosons don't obey the Pauli exclusion principle. Meanwhile, bosonic anti-particles arise just as straightforwardly in quantum field theory as fermionic anti-particles.

Although Dirac's clever argument is not the right one for fundamental physics, it does turn out to have its uses elsewhere because it's a good description of what happens in solid materials. Any solid is made of atoms, and some number of electrons typically disassociate themselves from the nuclei and wander around which, in the quantum world, means that they fill up the lowest energy levels provided by the surrounding solid. In this context, this is called the *Fermi sea* but it conceptually identical to Dirac's sea. When an electron is excited out of this sea, it leaves behind a hole. This hole – which is the absence of an electron – behaves in many ways like a particle with

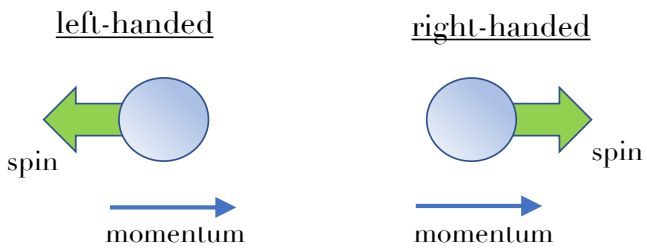


Figure 6. The handedness of a massless particle is determined by the relative direction of its spin and momentum.

positive electric charge. Indeed, there are some materials in which electricity appears to be conducted by positively charged particles. These aren't protons! They are holes.

There is a lesson here which is repeated over and over again in the history of science: a good idea tends to find a place in the world, even if it's not where it was originally intended.

2.1.4 Massless Particles

For massless particles, the story of spin needs to be slightly retold. Before we jump into the details, it's natural to ask: why bother? What spin $\frac{1}{2}$ particles in Nature are massless?

The answer to this question is shocking: all of them! One of the most striking features of the Standard Model is that, at the fundamental level, all the matter particles are massless. In fact, more than that, it turns out that it's not possible to incorporate masses into the theory without first doing some damage to some aspects of the weak force. This damage is achieved by the Higgs boson which, ultimately, is why the fundamental particles appear to have mass. We will describe all of this in Section 4. But, in preparation for that, it will be useful to explain here what becomes of spin when particles are massless.

Solving the Dirac equation, one finds that as a particle gets faster, its spin necessarily becomes oriented along the direction of motion. For massless particles, which travel at the speed of light, there are two options: either the spin points in the same direction as the particle is travelling, or it points in the opposite direction. When the spin points in the same direction, the particle is said to be *right-handed*. When it points in the opposite direction, it is said to be *left-handed*. This is shown in Figure 6.

This distinction is quantified using something called *helicity*. If the particle moves with momentum $\mathbf{p}$ and the spin points in the direction $\mathbf{s}$ then the helicity is defined to be

$$h = \frac{\mathbf{s}\cdot\mathbf{p}}{|\mathbf{p}|}$$

Right-handed particles have helicity $+\frac{1}{2}$; left-handed particles have helicity $-\frac{1}{2}$.

Such a distinction doesn't make sense for massive particles. One can simply overtake the particle and look back to see it moving in the opposite direction, but with the spin remaining the same, so its helicity appears to be flipped. However, you can never overtake a massless particle because it travels at the speed of light, and this means that everyone agrees on the helicity of a massless particle.

In fact, one can go further. It turns out that, for massless particles, it is possible to have "half a Dirac fermion". This is a particle where only, say, the left-handed helicity exists. There is no particle at all with right-handed helicity. The anti-particle would then exhibit the opposite behaviour, only existing in the right-handed state, never left-handed. A particle with these properties is known as a *Weyl fermion*. This idea will play a key role when we discuss the weak force.

I should stress that such Weyl fermions, with fixed helicity, are only possible for massless particles. The particles that we observe, such as electrons, do ultimately have a mass and they achieve this by gluing together two Weyl fermions to form a complete Dirac fermion, with both kinds of spin. We'll learn more about how this happens in Section 4.

2.2 Quantum Electrodynamics

The Dirac equation described in the previous section tells us that matter particles necessarily come with anti-particles. But for these particles to subsequently do something, they must interact. Those interactions happen through forces.

The simplest force in particle physics is electromagnetism. In large part, it is simplest because we have some classical intuition for this force: it is the same force understood many centuries ago by Coulomb, Ampére, Faraday and Maxwell, albeit dressed by some quantum bells and whistles.

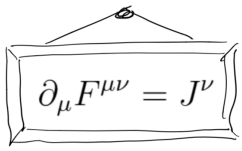
The force is mediated by two fields, the electric field $\mathbf{E}(\mathbf{x},t)$ and the magnetic field $\mathbf{B}(\mathbf{x},t)$. Each of these is a *vector field*, meaning that at every point in space $\mathbf{x}$ and for every time t, the field is specified by both a magnitude and a direction.

The equations that describe the dynamics of the electric and magnetic fields are known as the *Maxwell equations*. We won't need them in these lectures but, for completeness, here they are:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad , \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$
$$\nabla \cdot \mathbf{B} = 0 \quad , \quad \nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \tag{2.3}$$

The fields $\mathbf{E}$ and $\mathbf{B}$ react to the presence of electric charge density ρ and electric currents $\mathbf{J}$, while ϵ_0 and μ_0 are two constants that characterise the strength of the electric and magnetic forces in a way that we will describe more below.

The equations, as written above, hide the full beauty of the Maxwell equations. A better formulation encodes both the electric and magnetic fields in a 4×4 anti-symmetric matrix called the field strength, which takes the form $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Only then do the Maxwell equations reveal their true simplicity, in a way that deserves hanging in a frame,



You can learn more about the Maxwell equations and their classical solutions in the lectures on Electromagnetism. Famously, among the solutions to these equations are electromagnetic waves, including visible light. When you look at these solutions through the lens of quantum mechanics, you find that they decompose into particles, known as *photons*.

A photon can come in two different states that we call polarisation. These are entirely analogous to the "spin up" and "spin down" states of the electron. (The fact that both spin 1/2 and spin 1 particles have two internal states is something of a coincidence. For example, it's only true in three spatial dimensions; the counting is different in other dimensions.)

The theory describing the electromagnetic field interacting with the electron field is known as *quantum electrodynamics*, or *QED* for short. It is the theory describing light interacting with matter, and ultimately underpins large swathes of science, including condensed matter physics and chemistry. Happily, it is also the simplest component of the Standard Model.

The Strength of the Interaction

Take two particles carrying electric charge Q_1 and Q_2 and hold them some distance r apart. The relevant solution to the Maxwell equations tells us that the particles experience a force given by

$$F = \frac{Q_1 Q_2}{4\pi\epsilon_0 r^2} \tag{2.4}$$

This is called the *Coulomb force*. The force is repulsive if the two particles carry charge of the same sign; it is attractive if they carry charges of opposite sign.

This formula shows us that the constant ϵ_0 characterises the strength of the Coulomb force. If the value of ϵ_0 was smaller, then Coulomb force would be more powerful. If you look up ϵ_0, you'll find some unhelpful number quoted with unit of Farads per metre. A more useful measure of the strength of the electric force comes from the dimensionless quantity known as the *fine structure constant*,

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \tag{2.5}$$

where e is the electric charge of the electron. It turns out that the value of the fine structure constant is roughly

$$\alpha \approx \frac{1}{137}$$

This is the cleanest way to characterise the strength of the electric force. In particular, in natural units with $\hbar = c = 1$, two electrons held a distance r apart experience a force given by

$$F = \frac{\alpha}{r^2}$$

Maxwell's equations also contain a second constant, μ_0, which characterises the strength of the magnetic interaction. It turns out that this is not independent from ϵ_0. One of the great discoveries of Maxwell is that the two constants are related by

$$\epsilon_0\mu_0 = \frac{1}{c^2}$$

with c the speed of light. As a side remark, note that if the strength of the electric force $1/\epsilon_0$ were weaker, then the strength of the magnetic force $1/\mu_0$ would necessarily be stronger.

2.2.1 Feynman Diagrams

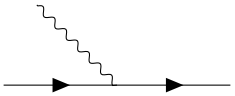
There are some simple cartoons that allow us to figure out what processes are allowed in quantum electrodynamics (and, indeed, in the other forces). These cartoons are called Feynman diagrams.

We will take time to run horizontally, from left to right[4]. We then draw electrons as solid lines with a forward pointing arrow, like this .

Positrons are depicted as solid lines with a backwards pointing arrow, like this . We'll see the utility of the backwards-arrow notation below. It suggests that it may be possible to think of anti-particles as particles that move backwards in time. There is mathematical sense in which this statement is correct, but it shouldn't be taken too literally.

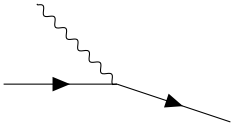
Finally, photons are depicted as wavy lines like this .

There is just a single interaction between the electron and photon, from which all other processes can be built. This can be viewed as an electron absorbing a photon, and scattering off in a different direction. It looks like this



The point where the photon hits the electron is referred to as a *vertex*.

Conservation of momentum means that the electron necessarily moves off in a different direction after absorbing the photon. So you might have thought that it would be better to draw the Feynman diagram like this



Indeed, sometimes we'll draw diagrams like this. However, the Feynman diagrams should not be read too literally: the paths aren't the actual paths of particles in spacetime. They should be viewed in a more topological fashion, like the London underground map. We'll say more below about what Feynman diagrams are, and what they aren't.

[4]This is the convention used in the Quantum Field Theory lectures, but it's not universal. Some authors prefer time to flow upwards.

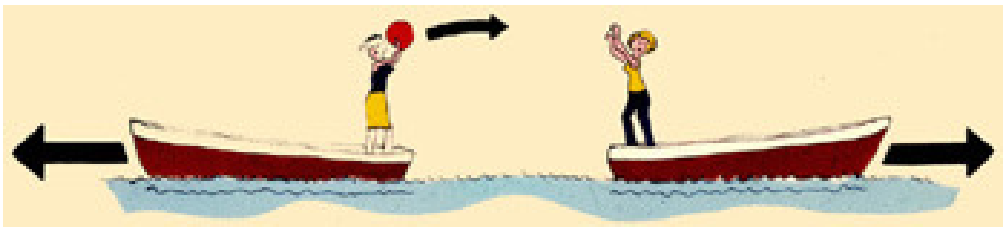
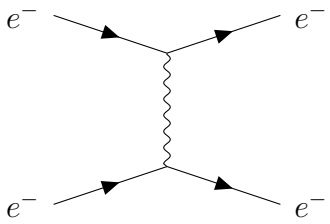


Figure 7. This is not a good analogy for virtual particles.

Now the game is as follows: you can describe any process by stitching together the Feynman diagram building blocks above. You can orient the different legs of the diagrams in any way you wish. You just have to make sure that the arrows on the solid lines follow each other. Any process that you can draw can happen, *provided that* it is allowed on grounds of energy and momentum conservation.

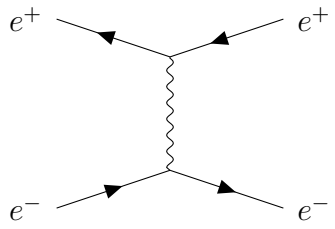
Let's look at some examples. Here is a Feynman diagram describing one electron scattering off another



I've added the name of the particle to the external legs, a practice that will prove useful as we progress. Note that the electrons don't just bounce off each other; there is no direct contact between them. Instead, the electrons scatter by exchanging a photon. Particles that appear only in internal legs of Feynman diagrams, like the photons above, are referred to as *virtual particles*. This is a lesson that we'll see repeated later: all forces can be understood by the exchange of virtual bosonic particles.

In some ways, Feynman diagrams are a little too evocative, and we should be careful not to interpret the diagram above too literally. For example, you shouldn't think of one electron as recoiling as it emits a virtual photon, which is then absorbed by the second, resulting in a repulsive force from Newton's third law, like two people in boats throwing a ball back and forth between themselves. This will then leave you puzzled about how such particle exchange can possibly lead to an attractive force.

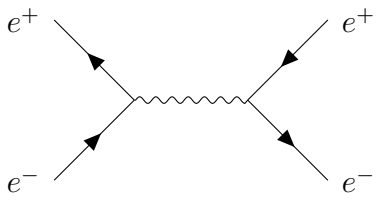
Yet in quantum field theory, there is no problem with virtual particles describing an attractive force. Indeed, the Feynman diagram for the scattering of an electron off a positron is almost identical to the one above:



Translating this diagram into mathematics gives the attractive Coulomb force, but this isn't easily captured by the people-in-the-boat analogy. (If you get to the point where you start thinking about people in boats throwing boomerangs backwards and forwards then you might realise that the analogy has clearly been stretched too far.)

In fact, what's really going on here is that both of the scattering diagrams above are a reformulation of the familiar result from classical physics in which one electron experiences a force due to the electric field of another. If you translate the diagram above into mathematics, you will find that it is simply a rewriting of the Coulomb force law (2.4). Viewed this way, the "virtual particles" are merely a handy device to capture the behaviour of the underlying field. If we were to think in terms of fields, then we have no need to discuss virtual particles. Moreover, there are situations — like for the strong force — where the concept of virtual particles is not useful, while the fields remain.

For the scattering of an electron off a positron, there is a second, qualitatively different diagram that also contributes.

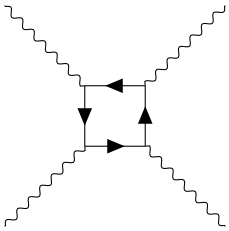


This has the interpretation of the electron-positron pair annihilating into a virtual photon, which then turns back into a pair of particles. It turns out that this diagram doesn't contribute to the Coulomb force (2.4), which holds only in the non-relativistic limit where velocities of all particles are low, but does change the scattering behaviour at higher energies. For our purposes, however, it is useful simply to illustrate the utility

of the forwards/backwards arrow notation for particles and anti-particles: they capture the conservation of electric charge.

The total electric charge of an electron-positron pair is zero, which allows them to annihilate into a (virtual) photon. In contrast, such a process isn't possible for the scattering of two electrons, because their charge is non-zero. In the diagrammatic language, we see this because the corresponding Feynman diagram  doesn't have the arrows matching up, and so is illegal.

There are also interesting processes that we can construct with photons on the external legs. For example, here is a diagram that corresponds to one photon scattering off another



Famously, light doesn't scatter off itself in the classical world. This is important, for it allows us to see! But it's no longer true in the quantum world. The diagram above can be viewed as a light scattering off a particle-anti-particle pair which briefly appear as a vacuum fluctuation. The probability for such a process is small, which is why we don't notice this process every day. But, although small, it is non-zero, and light-by-light scattering has been observed in particle colliders.

2.2.2 What is a Feynman Diagram Really?

All quantum processes have an element of randomness. Particle physics is no different. If you collide two particles together at high energies, there are many possibilities for what may emerge. Quantum field theory allows us to assign probabilities — or, more precisely, quantum amplitudes — to all of these possibilities.

However, there's a hitch. Quantum field theory is hard, and the expressions for these probabilities are ridiculously complicated. In many situations, we have no idea how to compute them. However, for QED we can make progress based on the observation that the interaction strength, as captured by the fine structure constant $\alpha \approx 1/137$, is small. This means that we can expand the complicated probabilities in a perturbative expansion, rather like Taylor expanding a function.

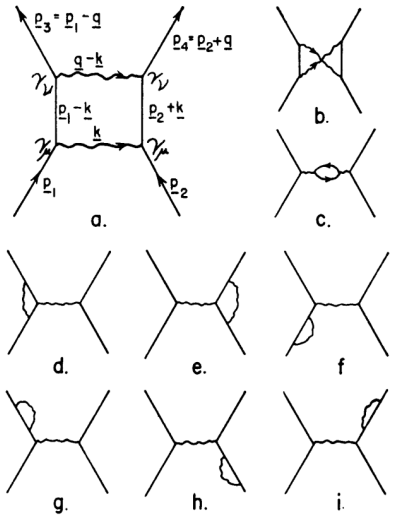
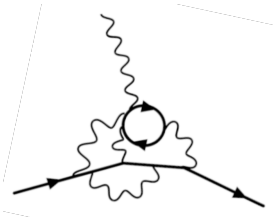


Figure 8. A handful of Feynman diagrams, taken from Feynman's original paper.

The Feynman diagrams are a pictorial way of capturing this perturbative expansion. Suppose that you want to compute the probability for some process to happen, for example the electron-positron scattering described above. The process itself is defined by the external legs of the diagram — these are what tell you, for example, that you start with two particles and end up with two particles. Given this data, you should now write down all possible Feynman diagrams. The diagrams that we drew above are the simplest diagrams, but there are an infinite number of diagrams contributing to any process with an increasingly complicated structure of internal lines. For example, the original vertex can be dressed with all sorts of other lines, to give things that look like this:



Some examples of Feynman diagrams for $e^- + e^+$ scattering are shown in Figure 8

So far, this procedure doesn't sound very helpful. We have to write down an infinite number of ever more elaborate diagrams to describe any process. Moreover, there are rules which translate each diagram into a mathematical expression, usually involving

some very complicated and challenging integrals. What saves us is the fact that the more complicated a diagram, the less important it is in some process. To compute the importance of any diagram, you need simply to count the number of vertices. While the exact contribution of any diagram may be very difficult to compute, the happy news is that it is proportional to

$$\alpha^{\#\text{ of vertices}}$$

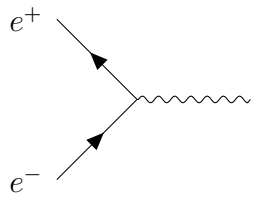
So, for example, a diagram with a single vertex will mean that the probability is something in the ballpark of 1/137. Meanwhile, the the elaborately dressed vertex shown above contributes something proportional to α^9. With $\alpha \approx 1/137$, this kind of diagram barely changes the answer, and so can be safely neglected. The light-by-light scattering diagram that we showed earlier comes in at α^4, explaining why we don't observe this phenomena in every day experience.

There is one important lesson to take from this: the utility of Feynman diagrams is intimately connected to the weakness of the electromagnetic interaction. In situations where the interactions between fields are strong, Feynman diagrams are not the right way to think about the physics.

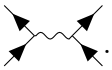
If you want to learn more about how to unmask Feynman diagrams, and turn them back into the underlying equations, then you can find details in lectures on Quantum Field Theory.

Some Examples

To illustrate these ideas, we can compute the relative probabilities that an electron and positron will annihilate to a bunch of photons. First, we need to address a subtlety. It's possible to draw a diagram representing an annihilation to a single photon:



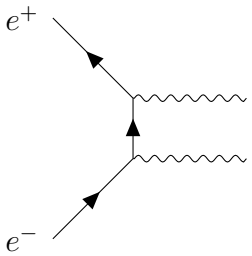
This would appear to be proportional to α. However, if you calculate this diagram, you'll find that it's vanishing. This is because although it is consistent with charge conservation, it's not consistent with the conservation of energy and momentum. To see this, consider the frame in which the electron and positron have equal and opposite momenta. The outgoing photon must then have vanishing momentum, but non-vanishing energy, and this isn't possible.

It's worth pointing out that this diagram can occur as a sub-diagram in other processes. For example, we already used it in electron-positron scattering  . In this case, the intermediate photon is virtual and it turns out that there's no requirement for virtual particles to obey the usual energy-momentum relations. You can handwave this away by saying that virtual particles can borrow energy for some period of time by virtue of (the slightly dodgy version of) the Heisenberg uncertainty relation: $\Delta E \Delta t \sim \hbar$.

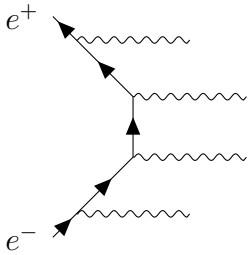
Back to the annihilation of an electron and positron, the simplest process results in two photons:

$$e^+ + e^- \to \gamma + \gamma$$

This is described by the following Feynman diagram:



From our discussion above, we know that this is proportional to α^2. But other processes are possible. For example, the pair could annihilate to any number $n > 1$ of photons. For example, the diagram for annihilation to four photons is



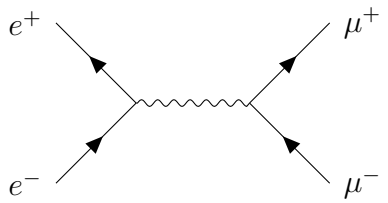
This probability for such a diagram is proportional to α^4. This means that the probability of getting four photons our of a collision is suppressed by a factor of α^2 relative to the probability of getting two photons.

2.2.3 New Particles From Old

Electrons are not the only particles that experience the force of electromagnetism. Any particle that carries electric charge also interacts with the photon. In particular this means that all fermions, except for the neutrinos, feel the force of electromagnetism.

We can easily extend our Feynman diagrams to include these extra particles. By convention, we depict any fermion with a forward arrow —▶— and any anti-fermion with a backward arrow —◀—, but now we must label these lines with the particle name to show what particle we're talking about. Every particle with electric charge will have an interaction vertex of the form —▶~▶—. When evaluating Feynman diagrams, these vertices contribute a factor of $Q^2\alpha$ to the probabilities, where Q is the charge of the particle, in units where the electron has $Q = -1$.

This brings new opportunities. For example, we can collide an electron and positron, but now produce new particles such as a muon-anti-muon pair as shown in the diagram below.



We still have to worry about energy and momentum conservation when evaluating this diagram. If, in the centre of mass frame, the energy of the incoming electron-positron pair is less than $2mc^2$, the rest mass of the muon-anti-muon pair, then the muons cannot be produced. However, when the incoming energy exceeds this threshold, then we can start to produce new particles from old ones. The same kind of process allows us to produce any charged particles from the collision of electrons and positrons. This, of course, is the basis for particle colliders.

2.3 Renormalisation

At the fundamental level our world is built not from particles, but from fields. Moreover, as we stressed in the introduction, these fields froth and foam in the uncertain quantum world. This gives rise to an important phenomenon known as *renormalisation*.

Let's consider a single electron. It gives rise to an electric field which, like the force, varies as an inverse square law,

$$\mathbf{E} = \frac{e}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}$$

where $\hat{\mathbf{r}}$ is the unit vector pointing radially outwards. Clearly the electric field gets bigger and bigger as we approach the position $r = 0$ of the particle. But what is happening to the electron field near this point? It turns out, that both electric and electron fields start thrashing wildly as we get near to $r = 0$.

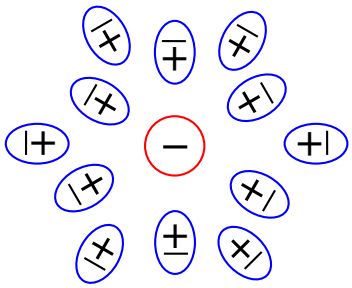
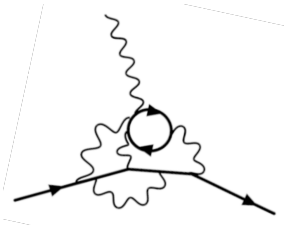


Figure 9. The renormalisation of electric charge.

While it's challenging to talk about quantum fields, we can build some intuition by reverting to the language of particles. As we get closer to the electron, the electric field gets stronger and, as a result, the energy density stored in the field gets larger and larger. At some point — a distance of around 10^{-12} m — the energy density is so large that an electron-positron pair can be produced from the vacuum.

There is a general rule in quantum field theory that anything that can happen does, in fact, happen. This means a single electron is surrounded by a swarm of particle-anti-particle pairs, continually popping in and out of the vacuum. As we get closer still, muons, taus and even quarks will also appear in the mix. We learn that any simple picture you may have of a single particle giving rise to the electric field is really very far from the truth: it is impossible to enforce any kind of social distancing in the quantum world.

This story should really be viewed in terms of quantum fields. When we talk of a swarm of particle-anti-particle pairs, it is really a metaphor for the quantum field being excited in a tangled and elaborate fashion. Just as the vacuum is something complicated in quantum field theory, so too is the notion of a single particle. In the language of Feynman diagrams, these particle-anti-particle pairs are captured by the diagrams that include loops of particles, like the one shown on the right.



The excited quantum field has an important consequence for the strength of the electromagnetic interaction. Again, we can understand this in the language of particles. The swarm of particle-anti-particle pairs will not be oriented randomly around the electron. Instead, the positrons, which carry positive charge, will be attracted to the

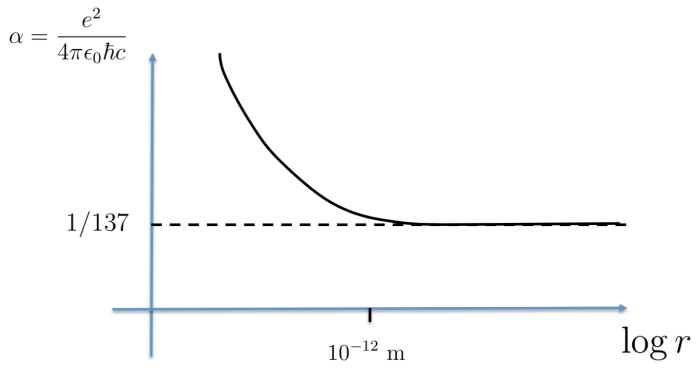


Figure 10. The renormalisation of fine structure constant.

electron in the centre, while the electrons will be repelled, as shown in Figure 9. The net effect is that as you get closer to the electron, you find more and more negative charges outside you. Which must mean that the original charge must have been bigger than we see. This is an effect known as *screening*. This swarm of particle-anti-particle pairs is actually hiding the true charge of the electron in the centre.

In fact, an excellent analogy of this phenomenon arises in metals. Take a positive charge and place it in a metal. The mobile electrons will enthusiastically cluster around, screening the positive charge so that it can't be detected at large distances. This is very similar to what happens with the electron in the vacuum and is one of many situations in which ideas in particle physics are mirrored in condensed matter physics.

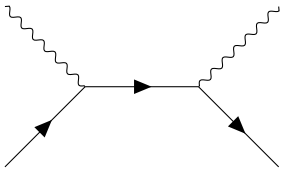
Because the effective charge of the electron gets bigger at shorter distances, so too does the interaction strength as captured by the fine structure constant (2.5). In fact, we learn that the fine structure constant is very badly named, since it's not in fact constant at all. At distances larger than $r \gtrsim 10^{-12}$ m, it plateaus to the usually quoted value of $\alpha \approx 1/137$. But as you go to smaller scales, the strength of the electromagnetic interaction increases logarithmically. For example, the strength of the interaction has been well measured at the scale of the weak force, which is roughly $r \approx 10^{-17}$ m, where it is found to be $\alpha \approx 1/127$. A sketch of the variation of the fine structure constant — often referred to as *running* — is shown in Figure 10.

The lesson of renormalisation as described above is a general one. It turns out that none of the dimensionless physical constants of nature are, in fact, constant. All of them depend on the distance scale you're looking at.

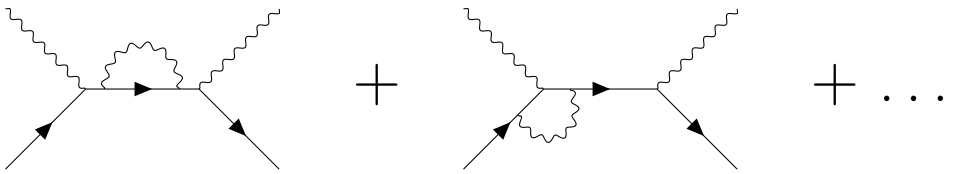
2.3.1 The Long, Confusing History of Renormalisation

While the description of renormalisation described above is fairly straightforward, the mathematics underlying it is not. For this reason, our forefathers had to travel a long and tortuous road to make sense of quantum field theory in general, and the issue of renormalisation in particular.

The story starts in the late 1920s, soon after the original development of quantum mechanics. The quantum pioneers — Heisenberg, Dirac, Pauli and others — tried to apply their ideas to the interaction of light and matter. They tried to ask very simple questions, like the probability for a photon to scatter off an electron. While they didn't have the diagrammatic tools later introduced by Feynman, they did understand that we could approach the problem in a perturbative expansion, starting with a process which we now draw like this:



They found that this calculation gave pretty good agreement with the experiments. But then they tried to do better, and compute the leading corrections. In diagrammatic language, this means evaluating diagrams like this



However, here they ran into a problem. Each of these subsequent diagrams was proportional to α^4, as expected. But the proportionality constant was infinity. That made it very hard to argue that the contributions from these diagrams was smaller than the first.

The quantum heroes worked on this problem for well over a decade, but made little progress. Looking back, many of their ideas were simply too crazy. Having forged one revolution they were, like Che Guevara, all keen for the next. Bohr wanted to get rid of energy conservation. Heisenberg wanted to make spacetime non-commutative. Pauli wanted to invade Bolivia. Yet the answer they were seeking did not, ultimately, require an overhaul of the foundations of physics. It needed a different approach.

The war intervened, and as life returned to normal a new generation of physicists took up the problem, people like Tomonaga in Japan, and Schwinger, Feynman and Dyson in the US. They were helped in no small way by a new experimental result, discovered by Willis Lamb. The eponymous Lamb shift is a tiny, but detectable change to the energy levels of hydrogen due to the problematic one-loop Feynman diagrams. Their solution was a slow-burn revolution, one that took many decades to play out as the power of quantum field theory became clear. However, at the time it didn't feel like a revolution. It felt like a con. Their solution was this:

$$\infty - \infty = \text{finite}$$

In other words, they found a mathematical procedure that allowed you to subtract one infinity from another, leaving an unambiguous finite answer. They called this process *renormalisation*. The results were nothing short of spectacular.

The poster boy for renormalisation is a quantity known as the magnetic moment of the electron. If you place an electron in a magnetic field, then the spin of the particle will precess, as shown in figure. The speed at which the spin precesses is characterised by a dimensionless number g known as the electron's magnetic moment.

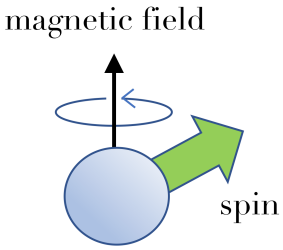


Figure 11. The precession of a spin

In the grand scheme of things, this number is not particularly important. However, it has played a key role in the development of quantum field theory because it is a quantity that we can determine with some accuracy, both experimentally and theoretically. After many decades of painstaking work, the experimental result for the electron magnetic moment is

$$g_{\text{expt}} = 2.0023193043617 \pm 3$$

Meanwhile, after many decades of extraordinarily challenging calculations, evaluating increasingly complex Feynman diagrams up to corrections of order α^5, the theoretical result is

$$g_{\text{theory}} = 2.00231930436\ldots$$

The agreement is awe inspiring. In most areas of science you jump up and down with joy if you get the first digit right. In economics you don't even need that. Yet here there is agreement between theory and experiment to 12 significant figures[5]!

Despite this runaway success, there was something a little disquieting about renormalisation. The idea that you can take one infinity away from another, to leave something finite was not mathematically legitimate. This was Feynman's take:

> "The shell game that we play to find [the answer] is technically called renormalization. But no matter how clever the word, it is what I would call a dippy process! Having to resort to such hocus-pocus has prevented us from proving the theory of quantum electrodynamics is mathematically self- consistent. .... I suspect that renormalization is not mathematically legitimate."

The physical meaning of renormalisation took several more decades to uncover. And it came from an unusual place: the attempt to understand boiling water! In particular, when water is close to the critical point, it turns out that the physics can be understood using very similar Feynman diagram techniques to those employed in particle physics. But this time there are no infinities. That's because when you get to the place in the calculation where infinities might rear their head, you need to remember that water isn't infinitely divisible, but is made of atoms. When you take this into account, the infinities in the diagrams simply aren't there.

But the same should also be true of quantum field theories in particle physics. There is no reason to think that our theories are valid to arbitrarily high energies, or arbitrarily short distance scales. A modern perspective on renormalisation absorbs this lesson. Just as Newtonian mechanics comes with a health warning, stating that you shouldn't trust it in extreme situations — when speeds become too large, masses become too heavy, or particles become too small — so too does quantum field theory. No quantum field theory should be trusted to arbitrarily small distance scales, or arbitrarily high energies. That would be hubris. Instead, we should admit that there is an energy scale beyond which our theory no longer applies. This energy is called the *cut-off*. For

[5]The same calculations for the magnetic moment of the muon give $g_{\rm theory} = 2.00233183602$ and $g_{\rm expt} = 2.00233184122$. Although the agreement is impressive, it fails at the $10^{\rm th}$ significant figure. This is one of the very few discrepancies between theory and experiment. If this disagreement is borne out, it may be due to the effects of new particles, beyond those of the Standard Model. If nothing else, this discrepancy should serve to show just how astonishingly good the Standard Model really is: there are surely no other areas of science where people have sleepless nights over a failure in the $10^{\rm th}$ digit.

the Standard Model, the cut-off is somewhere above 1 TeV. Simply injecting a little humility into the proceedings, and admitting that we don't yet understand everything, is sufficient to remove the infinities.

But now there is a new problem. We must make sure that no physical answer depends on our choice of the cut-off. After all, the cut-off is an expression of our ignorance, the limit of our current knowledge. It would be very unsatisfactory if something physical like, say, the electron mass depended on what we don't know about physics at high energies.

It took many scientists many years to figure out how to include a level of our ignorance in the theory, without anything depending on our ignorance. The different parts of the story were finally pieced together in the early 1970s by Kenneth G. Wilson. His work is surely the most influential piece of theoretical physics in the latter part of the 20^{th} century, and now has applications from particle physics to biological physics to gravitational physics. Wilson's insight was that nothing you can physically measure depends on the choice of cut-off providing that physical quantities are not constant: they must change depending on the scale at which you explore the world. This is way that we introduced renormalisation in the previous section. Moreover, the complicated and dubious calculations which seemingly gave $\infty - \infty =$ finite can be reinterpreted in a more palatable way as telling us how quantities change with scale.

This new approach is sometimes referred to as the *Wilsonian renormalisation group*, to distinguish it from the older approach of Feynman and others. You can read more about this in the lectures on Statistical Field Theory.

The upshot of this is that if you do quantum field theory carefully, there are no infinites. But neither are there (dimensionless) constants of Nature. Instead, one key lesson of quantum field theory is that the universe in which we live is organised by scale. If you want to write down a theory of our world, then you need to explicitly state the scale at which the theory holds. Change the scale, and you must change the theory. Or, at the very least, you must change the parameters of the theory.

There is a final twist to this. The Feynman quote on the previous page is from 1985, fifteen years after Wilson did his crucial work and three years after Wilson was awarded the Nobel prize. I don't know why Feynman still held that opinion at that time. It is conceivable that he was unaware of the importance of Wilson's work. After all, the detailed calculations are not wildly different from those Feynman himself did decades earlier, and perhaps he did not appreciate the all-important change of emphasis. Or perhaps he simply didn't like the truth getting in the way of a good story.

B Interlude: Looking to the Sky

Radioactivity was a gift to physicists. Beta decay provided the first insights into the weak force, a topic we will return to in Section 4, while beams of alpha particles — which can reach energies up to 5 MeV — were used as a microscope to peek into the nucleus for the first time. As we learned in Interlude A, both the proton and the neutron were discovered by bombarding other elements with alpha particles.

But 5 MeV can only get you so far. Looking inwards requires higher energies. The smaller the distance scale you want to explore, the higher the energy you need. Ultimately, progress was made by constructing particle accelerators, but physicists first made use of another of Nature's gifts: cosmic rays.

The Rise of the Balloon

Our world is constantly bombarded by charged particles from the cosmos. These particles – which are collectively known as *cosmic rays* – are mostly protons or helium nuclei, with the occasional electron and heavier nuclei thrown in for good measure. They travel enormous distances before reaching us, originating far outside our galaxy in supernova explosions or in the accretion discs which surround supermassive black holes in the centre of other galaxies.

When cosmic rays hit the upper atmosphere, they create a shower of new particles, many of which survive the journey down to Earth where they can be detected. Theodore Wulf was the first to realise that there was something interesting to be explored. In 1910 he built a simple electrometer to detect ionised particles in the atmosphere. At the time, it was thought that this ionising radiation was emitted by the Earth. Wulf had the simple but brilliant idea to test this by climbing the Eiffel tower to perform his experiment. He found that the amount of radiation did indeed drop, but nowhere near as quickly as one would expect. Something was afoot.

The challenge was accepted two years later by Victor Hess. Needing to get to greater heights, he took up ballooning. At a height of 1 km, he found that the radiation was more or less the same as on Earth. At a height of 5 km, he found the radiation was nine times greater. To test various hypotheses, he flew in the day and he flew at night. He even flew during a solar eclipse. He concluded that either there was some unknown substance hiding in the upper atmosphere, or the radiation had an extraterrestrial source. He called it *ultraradiation*.

The name didn't stick. In Caltech, Robert Millikan (of oil drop fame) turned his attention to this new phenomenon. He had a better name: cosmic rays. More importantly, he also had resources and a team of brilliant experimenters who could explore their implications.

With hindsight, it's very clear why cosmic ray showers provided such an opportunity for particle physics. Radioactivity offers alpha particles with energies up to 5 MeV. Cosmic rays have no such limitations. A plot of the energy vs the flux of cosmic rays is shown to the right. As you can see, cosmic rays with energies of 1 GeV are common place, but energies extend up to 10^{11} GeV, way beyond what we can create in colliders. To find interesting physics, we just need to have our detectors in the right place at the right time.

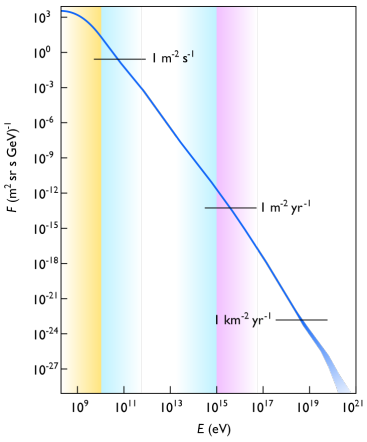


Figure 12. The distribution of cosmic ray energies. Image from Wikipedia.

B.1 The Positron

In May 1931, after struggling for some years with the meaning of the negative energy solutions of his equation, Dirac finally made the bold leap and predicted the existence of anti-electrons. In September 1932, Carl Anderson announced the discovery of a new particle, with the same mass as the electron, but opposite charge. He later named this particle the positron. Rather surprisingly, the discovery of anti-matter owed essentially nothing to its earlier theoretical prediction.

Anderson's interest was in cosmic rays. Unlike Hess, however, he had no intention of getting in a balloon. With a good detector, he needed to climb no higher than his third storey office to study the showers from cosmic rays.

Anderson's detector was the cloud chamber. We already briefly met a preliminary version of the cloud chamber in Section A.2 when describing J.J. Thomson's discovery of the electron. In the intervening years, it was perfected and was usually referred to as the *Wilson cloud chamber* after its inventor, C.T.R. Wilson. When a charged particle passes through the chamber, it leaves behind a path of ionised gas particles, around which droplets subsequently condense. The result is that an elementary particle leaves behind a misty trail, visible to the naked eye, like the contrails left by a plane in sky.

Working at Caltech under the guidance of Millikan, Anderson built a cloud chamber sitting within a magnetic field of 25,000 Gauss. The purpose of the magnetic field was

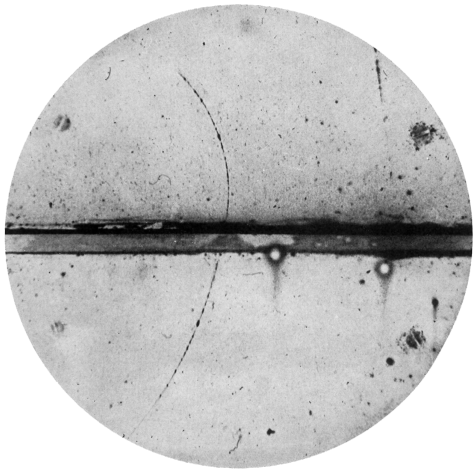


Figure 13. Anderson's first published picture of a positron. You can see the layer of lead running through the middle. The positron entered the cloud chamber at the bottom left, was slowed by the lead, with the trail visibly more curved after it exited. Somewhat unusually, this positron arose from a cosmic ray collision below the detector.

to bend the trajectory of a charged particle, allowing one to get a handle on the ratio e/m of charge to mass. Anderson found trajectories bending in both directions. Those that bent in one direction were clearly negatively charged electrons, coming down from the sky. But what about those that bent in the other direction? They were too light to be protons. However, they could have been electrons coming up from the ground. It seemed unlikely because, as Milikan pointed out: "Everyone knows that cosmic ray particles go down. They don't go up except in very rare circumstances".

To better understand what was going on, Anderson placed a thin, horizontal layer of lead in the the cloud chamber. This wasn't thick enough to stop the particles completely, but it did cause them to lose a significant amount of energy as they passed through. This meant that the particle would be travelling more slowly after it passed through the lead, and so its trail would bend in a tighter curve. In this way, Anderson was able to determine the direction of the particle and, hence, its charge. He found, in the words of his original paper, an "easily deflectable positive".

Anderson's results were soon confirmed by others, notably Patrick Blackett and Giuseppe Occhialini in Cambridge. Indeed, the week before Anderson dropped his bombshell, Blackett and Occhialini published a paper in Nature entitled "*Photography of Penetrating Corpuscular Radiation*" in which they boasted about their new toy, a cloud chamber in a high magnetic field, rigged up to work only when an accompanying

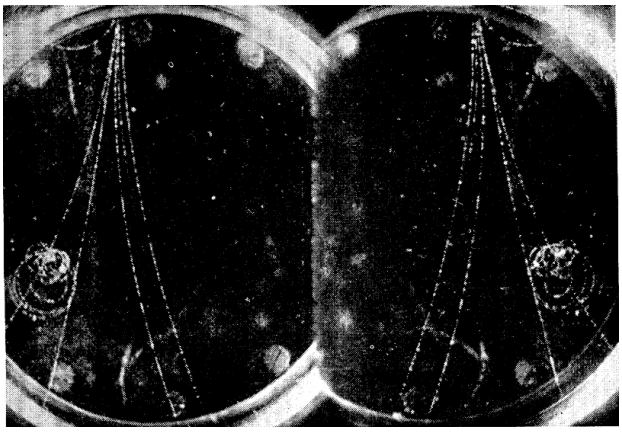


Figure 14. This cloud chamber track, showing electrons and positrons veering in opposite directions, was taken by Anderson and his student Seth Neddermeyer in 1936. They were still avoiding balloons, but climbed to Pike's Peak in the Rocky mountains, 4300 m above sea level, where, as they stated in the abstract, "the proportion of such tracks is considerably greater than at Pasadena."

Geiger counter triggered. They illustrated their paper with a few uninspiring photographs to show that the machine worked. It was only after they heard of Anderson's result that they realised they focussed on the wrong images: among their photographic plates were positrons in "great abundance". The pain of this oversight must have been felt even more acutely given that they were colleagues of Dirac. As a (very!) small compensation, they did quickly find something that Anderson had missed: the creation of electron-positron pairs within the chamber.

Later, when asked if they were aware of Dirac's theory when performing their experiment, Blackett replied that he ...

> "... could not recall but that it did not matter anyway because nobody took Dirac's theory seriously."

This seemed to be the prevailing attitude, at least among experimenters. Neither Anderson's original discovery paper, nor a longer paper published in 1933 in which he introduces the name *positron*, mentions the theory of Dirac[6]. Anderson later recalled:

[6] As a slightly bitchy aside, it appears that theory wasn't Anderson's strong point. His longer paper on the positron ends with the "the proton will then in all probability be represented as a complex particle consisting of a neutron and positron".

> "The Dirac work was not an important ingredient in deciding which way the experiments should be carried out or what should be done experimentally."

Rutherford, begrudging as always towards theorists, took things one grumpy step further:

> "It seems to me that in some way it is regrettable that we had a theory of the positive electron before the beginning of the experiments. Blackett did everything possible not to be influenced by theory, but the way of anticipating results must inevitably be influenced to some extent by theory. I would have liked it better if the theory had arrived after the experimental fact had been established."

B.2 Expecting a Meson

At this point our story of discovery gets somewhat out of sync with the main thread of the lecture notes. To understand what happened next, we must first make some comments on the strong nuclear force. We will describe the theory underlying this force in much more detail in Section 3.

The first hint that a new force was needed to explain the structure of the nucleus came from (who else?) Rutherford. In 1917, he started a series of experiments, following the set-up of Geiger and Marsden but replacing their sheets of metal with hydrogen.

The charge on a hydrogen nucleus is almost 80 times smaller than that of the gold used in the original Geiger-Marsden experiment. This means that the Coulomb repulsion is significantly smaller and the α particle can get much closer to the nucleus. Rutherford noted that the scattering of α-particles no longer agreed with his formula (A.2) that had worked so successfully in the past. Nor did it agree with a more detailed study by Darwin that assumes a Coulomb repulsion, but allows for scattering of the nucleus as well as the α-particle.

We now know that this is because the α-particle and hydrogen nucleus get close enough to experience the strong force. However, the world wasn't quite ready for a new force and Rutherford originally suggested that the effect could be explained by some deformation of the α-particle.

As time went on, this interpretation became increasingly untenable. At a meeting at the Royal Society in 1929, Rutherford stated clearly

> "The hydrogen and helium nucleus appears to be surrounded by a field of force of unknown origin"

But what are the properties of this force field?

The key insight was made in 1934 by the Japanese physicist Hideki Yukawa. He realised that if there were a new spin 0 particle[7] of mass m then it would give rise to a potential energy between particles which varies with the separation r as

$$V(r) \sim \frac{e^{-r/R}}{r}$$

This is now known as the *Yukawa potential*. It has the property that it quickly goes to zero for $r \gg R$, where the range R of the potential is inversely related to the mass as

$$R = \frac{\hbar}{mc}$$

This is the same relationship between energy and length that we met earlier in when describing the Compton wavelength (1.2). The data available at the time suggested that the strong force had a range of about $R \approx 2 \times 10^{-15}$ m. This meant that if you wanted to explain the strong force in terms of some field, you should be looking for a new particle approximately 200 times heavier than the electron, or

$$m \approx 100 \text{ MeV}$$

The idea languished until 1937, when just such a particle was found.

B.3 The Muon and the Pion

The discovery of the muon didn't happen overnight. There was no smoking gun event that people could point to and shout "Eureka". Instead it was more of a slow burn as, from 1934 to 1937, an increasing number of cosmic ray tracks had absorption properties that didn't seem to fit theoretical expectations.

The prime driver in these discoveries was, once again, Carl Anderson, now working with his recently-graduated student, Seth Neddermeyer. As they understood more about cosmic rays, they found tracks that didn't lose energy as quickly as theorists predicted. But it wasn't clear whether the theorists should be trusted, or whether the data contained something more interesting.

[7] This is draped in a little bit of hindsight. Yukawa's original paper suggests a massive spin 1 particle, but only looked at the contribution from the first component, analogous to focussing on the just the electrostatic potential and ignoring magnetic fields. Later, in 1937 with Sakata, he developed the theory with a massive spin 0 field.

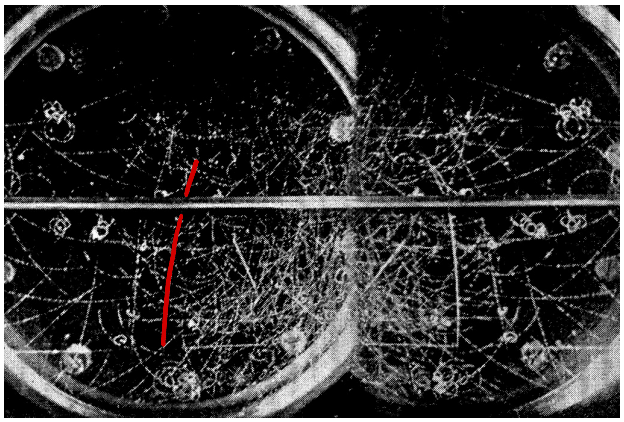


Figure 15. A cloud chamber picture from the same 1936 paper as Figure 14. The track daubed in red by my ipad pen was difficult to interpret as either an electron or a proton.

Already in 1936, Anderson and Neddermeyer published results which didn't conform to expectations. They pointed out that the red track shown in Figure 15 is too ionizing to be identified as an electron, but travels further than expected from a proton. In 1937, they finally bit the bullet, concluding that "there exist particles of unit charge but with a mass ... larger than that of a normal free electron and much smaller than that of a proton."

By 1939, the data seemed to suggest that this new particle had a mass 200 times heavier than the electron. The connection to Yukawa's proposed particle was obvious and a number of names were suggested for this new particle, including "yukon". Physical Review, pedantic as ever, insisted on "mesotron". Physicists ultimately converged on "meson", with the meso- from the Greek "mid".

However, as time went on, less and less about this new meson made sense. Models of nuclear binding worked much better with a particle that was slightly heavier and decayed significantly quicker. More brutally, experiments in 1946 showed that the interaction of the new meson with nuclei was around 10^{12} times weaker than that predicted by Yukawa's theory. That was too large a discrepancy to overlook!

Resolution

Happily the situation was resolved not long after. The discovery was made in 1947 in Bristol, England by a group of scientists led by Cecil Powell. Whenever Powell's collaborators are mentioned, people always include Giuseppe Occhialini (who recall, just missed out on the discovery of the positron) and sometimes César Lattes. But

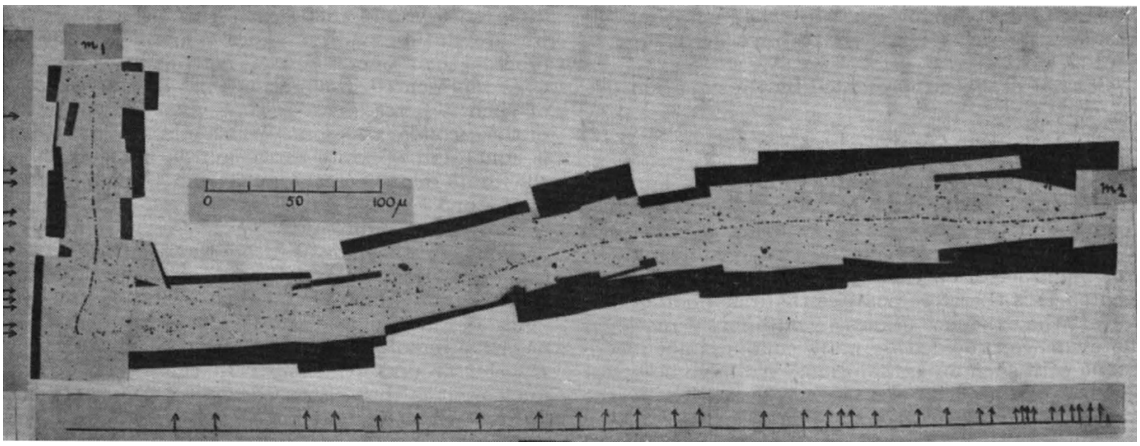


Figure 16. The discovery of the charged pion. It enters in the top left (labelled m_1), slows in the bromide and comes to rest, before decaying into a muon that flies off to the right (labelled m_2) and an anti-neutrino which is invisible in the picture. The caption in the paper starts with the comment "Observation by Mrs I. Roberts".

they rarely mention the fourth author on the paper, a poor graduate student by the name of Hughes Muirhead. And they certainly never mention the people behind the scenes who did the hard work: a team of women who painstakingly studied the images under microscopes to find the interesting events[8].

The discovery made by Powell's team was possible, as always, because they had a new piece of kit. Powell developed a new way of detecting particles by coating a glass plate with a photographic emulsion. When a charged particle passes through, it activates the emulsion, leaving behind a trail of silver bromide. They exposed their photographic plates to cosmic rays at high altitude, in balloons and on mountains, including Jungfraujoch and Kilimanjaro. When developed, the plates revealed a new meson, one both heavier and more short-lived, which decays quickly to the earlier meson. The V-shaped tracks were first found by Marietta Kurz, but these sat towards the edge of the emulsion and so were considered incomplete. A few days later, two clear L-shaped tracks were found by Irene Roberts. This was the long-sought meson.

Now, of course, we had two "mesons", rather than one. It quickly became clear that the particle discovered by Roberts had the properties expected of Yukawa's meson and, as we explain in the next section, can be viewed as the glue that binds together the proton and neutron in the nucleus. This became known as the π-meson, or pion. It is

[8]Actually, this statement was true when I wrote it, but then I decided to edit the Wikipedia page.

not an elementary particle but is, we now know, composed of quarks. Meanwhile, the particle discovered by Anderson and Neddermeyer is something new entirely, completely unrelated to the nuclear force. It became known as the μ-meson although, as time passed, the word meson was dropped. It is now simply the muon.

The muon-pion mix-up, which lasted a decade, is entirely due to a coincidence in their mass. We now know that the mass and lifetime of the two is

$$\begin{aligned} \pi^{\pm}: &\quad M \approx 140 \text{ MeV} \quad \text{and} \quad T \approx 2 \times 10^{-8} \text{ s} \\ \mu^{\pm}: &\quad M \approx 106 \text{ MeV} \quad \text{and} \quad T \approx 2 \times 10^{-6} \text{ s} \end{aligned}$$

The pion decays primarily as $\pi^- \longrightarrow \mu^- + \bar{\nu}_\mu$ and $\pi^+ \longrightarrow \mu^+ + \nu_\mu$. (You'll have to wait until we discuss the weak force in Section 4 to understand how this decay occurs.) Moreover, as we explain in Section 3.4, we have a good understanding of the pion in terms of its constituent quarks. We even understand why it has the mass that it does. In contrast, the muon remains a mystery, a repetition of the electron at a higher scale whose existence is as surprising to us today as in the 1940s.

B.4 The Beginning of the Deluge

Cosmic rays had still more surprises in store for physicists searching for elementary particles. The first came later in 1947, when George Rochester and Clifford Butler, working in Blackett's laboratory in Manchester, made a careful study of around 5000 photographs that they had taken over the previous year. Among them they discovered two with peculiar features.

The first, shown on the left of Figure 17, contains a forked track, seemingly appearing from nowhere. This is due to a neutral invisible particle which subsequently decays into two charged particles which are either muons or pions. The second, shown on the right of Figure 17, shows a marked kink, strongly suggesting that a charged particle decayed into a different charged particle (again, either a muon or pion), but also an invisible neutral particle which left without leaving a track. These new particles were estimated to have masses between 770 m_e and 1600 m_e. They were dubbed *V-particles* on account of the V-shaped tracks that they left.

It was not long before further V-particles were discovered, some lighter than the proton, some heavier. Indeed, at some point it seemed like there would be no end to these new particles. In collecting his Nobel prize in 1955, Willis Lamb quipped

> "The finder of a new elementary particle used to be rewarded by a Nobel Prize, but such a discovery now ought to be punished by a \$10,000 fine. "

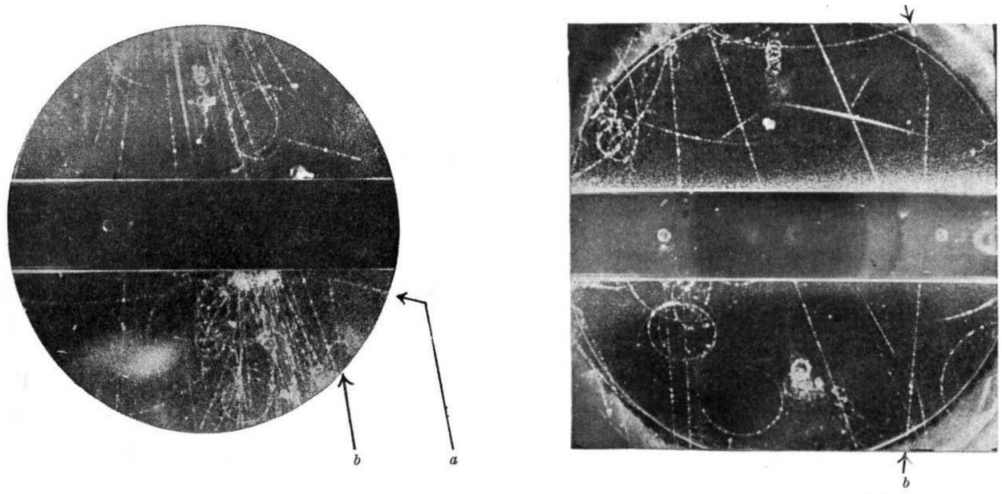


Figure 17. The discovery of kaons. A neutral kaon decays on the left, a charged kaon on the right.

For the next decade or so, particle physics entered a phase of taxonomy. The goal was to classify particles, first in terms of their mass and lifetimes, and then in terms of their decays, looking for patterns that would bring some order to the mess. It took many years before the things fell nicely into place and it was appreciated that all these particles could be understood in terms of yet smaller constituents called quarks. The V-particles shown in Figure 17 are now called *kaons* and were the first particles seen that contain a strange quark. We will tell the story of quarks in the next section.

The discovery of V-particles marks the beginning of a new era in theoretical physics. It also marks the end of an era in experimental physics. By the mid-1950's the energies and fluxes from accelerators were more than competitive with those in cosmic rays, and man-made muons, pions and V-particles were readily available. We will chart the rise of accelerator and detector technology in Interlude C.

3 The Strong Force

In the middle of the 20$^{\text{th}}$ century, the number of seemingly "fundamental" particles exploded. The electron, proton and neutron had long been known. These were joined, in 1947, by a new particle called the *pion*, whose role was to keep the proton and neutron bound inside the nucleus.

The discovery of the pion was welcomed: it had long been expected, and its role in the universe was understood. The discovery, the same year, of another particle called the kaon was more confusing. As was the subsequent discovery of the rho meson, the eta, the eta prime, the delta, the lambda and the xi. Before long, physicists were running out of Greek and Roman letters to name these particles. A glance through the current particle data book includes particles with enticing names like $Z_c(3900)$.

By the end of the 1960s, there were hundreds of seemingly elementary particles and the situation looked hopelessly complicated. It was clear that these particles could not all be fundamental, but it was difficult to see any simple underlying explanation. In despair, Freeman Dyson declared

> "I am acutely aware of the fact that the marriage between mathematics and physics , which was so enormously fruitful in past centuries, has recently ended in divorce."

Yet Dyson accepted defeat too soon. The answer was discovered in the early 1970s. In part, the answer lay in the existence of constituent particles called *quarks*. But, equally as important, were the peculiar and unprecedented properties of the force that binds these quarks together. This is the *strong nuclear force*.

3.1 Yang-Mills Theory

Both the strong and weak nuclear forces share a common property with electromagnetism: the force is carried by a field of spin 1. In the case of electromagnetism, this field is described by the Maxwell equations. For the two nuclear forces, it is described by a generalisation of the Maxwell equations known as the *Yang-Mills equations*.

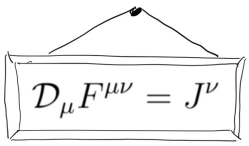
It's not so easy to write down a consistent generalisation of the Maxwell equations. In fact, it turns out that there's more or less a unique way to do it. This is based on a mathematical framework known as group theory. Here we'll give a baby version of this.

Yang-Mills theory is, like its predecessor, based on electric fields $\mathbf{E}(\mathbf{x}, t)$ and magnetic fields $\mathbf{B}(\mathbf{x}, t)$. Each of these is again a 3-dimensional vector,

$$\mathbf{E} = (E_x, E_y, E_z) \quad \text{and} \quad \mathbf{B} = (B_x, B_y, B_z)$$

This is the essence of what it means for a field to be spin 1. The novelty in Yang-Mills theory is that each component of these vectors is now a matrix at each point in space and time, rather than just a number. There are different versions of Yang-Mills theory for different kinds of matrices. There is a Yang-Mills theory based on 2×2 matrices, and one based on 3×3 matrices and so on for each integer N. However, once you decided on the size of the matrix, everything else is fixed.

To write the classical equations of motion, it's best to again bundle the "electric" and "magnetic" fields into a 4×4 matrix $F_{\mu\nu}$, each component of which is now itself an $N \times N$ matrix: i.e a matrix of matrices. The Yang-Mills equations of motion are one of the key equations of physics, and fully deserving of their place in a frame



Here $\mathcal{D}_\mu$ is something like a partial derivative with respect to space and time, but one that also includes some commutator of matrices. (It's known as a covariant derivative.) On the right-hand side sits J^ν, the analog of the electric current which, as we will see shortly, arises from quarks. The Yang-Mills equations are very similar to the Maxwell equations. Indeed, if you choose 1×1 matrices, which are just numbers, then the Yang-Mills equations reduce to the Maxwell equations.

To specify any force described by Yang-Mills theory, we just need to say how big the matrices are. Nature is kind to us: she has chosen to make use only of the simplest matrices.

- Electromagnetism: 1×1 matrices
- Weak Nuclear Force: 2×2 matrices
- Strong Nuclear Force: 3×3 matrices

Isn't that nice!

Before we go on, I should confess that the above discussion was a little imprecise in places. A correct statement is that there exists a version of Yang-Mills theory for every Lie group. Everywhere that I said "size of matrix", you should replace this with "choice of Lie group" where a "Lie group" is a fancy mathematical object. Matrices provide some simple examples of some Lie groups, but there are also others. Happily, all the groups that we need in particle physics can be reduced to matrices. A more grown-up version of the above list then characterises each force by a group. For what it's worth, they are:[9]

- Electromagnetism: $U(1)$
- Weak Nuclear Force: $SU(2)$
- Strong Nuclear Force: $SU(3)$

For the rest of this section, we will focus on $SU(3)$ Yang-Mills, relevant for the strong force. The theory of the strong force, interacting with quarks, is known as *Quantum Chromodynamics*, or QCD for short. The 3×3 matrix-valued electric and magnetic fields are sometimes called *chromoelectric* and *chromomagnetic* fields.

3.1.1 Gluons and Asymptotic Freedom

If you solve the Maxwell equations, you find waves propagating at the speed of light. These are light waves. As we have seen, in the quantum theory, these waves are comprised of massless spin 1 particles called photons. The massless nature of the photon is the reason light waves travel at the speed of light.

Similarly, if you solve the classical Yang-Mills equations, you again find waves travelling at the speed of light. In analogy with electromagnetism, we might expect that, in the quantum theory, there are massless particles associated to these waves. But no such massless particles are seen in the world. What's going on?

[9] A little matrix knowledge can be a confusing thing at this stage. The strong nuclear force is associated to the group $SU(3)$, which consists of 3×3 complex, unitary matrices of determinant 1. This means that the matrix U must obey $U^\dagger U = \mathbb{1}$ and $\det U = 1$ to be in $SU(3)$. But this isn't the kind of 3×3 matrix that make up the components of the chromoelectric and chromomagnetic fields. Instead, these are Hermitian matrices, namely 3×3 matrices which obey $E_x = E_x^\dagger$. There is, however, a relationship between these kinds of matrices: the exponential of a Hermitian matrix, like e^{iE_x}, is in the group $SU(3)$. Mathematically, this is the difference between a Lie group and a Lie algebra.

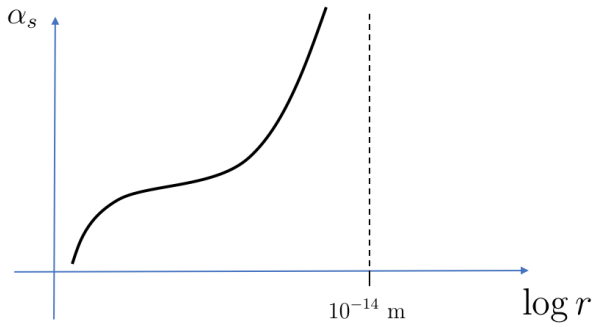
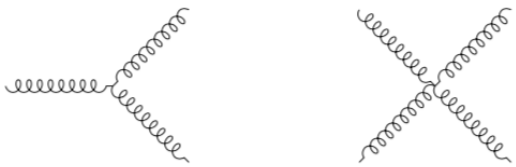


Figure 18. The renormalisation of coupling for the strong force.

The answer to this question lies in a subtle property of Yang-Mills theory, which means that the quantum theory looks very different from its classical counterpart. There are spin 1 particles associated to the $SU(3)$ Yang-Mills theory and these are called *gluons*. But, rather surprisingly, gluons turn out to be massive rather than massless.

To understand this, we first need to appreciate a key difference between Yang-Mills theory and Maxwell theory. The Maxwell equations are linear. One consequence of this is that light wave pass right through each other. In contrast, the Yang-Mills equations are non-linear. (This is not obvious in the framed equation on the previous page. It is hidden in the meaning of the covariant derivative $\mathcal{D}_\mu$.) This means that two classical waves in Yang-Mills will typically scatter off each other in some complicated fashion.

When we turn to the quantum theory, the non-linearity translates to an interaction vertex in the Feynman diagrams for gluons. We depict gluons using curly lines like this ~~~~ . There are, it turns out, two interaction vertices: one where three gluons interact, and another where four gluons interact:



Just as in QED, there is a dimensionless coupling constant that characterises the strength of this interaction. For QED, this was the fine structure constant α. For the strong force, the coupling is denoted α_s. And, importantly, just like for QED, the value of this coupling depends on the distance scale (or, equivalently, energy scale) at

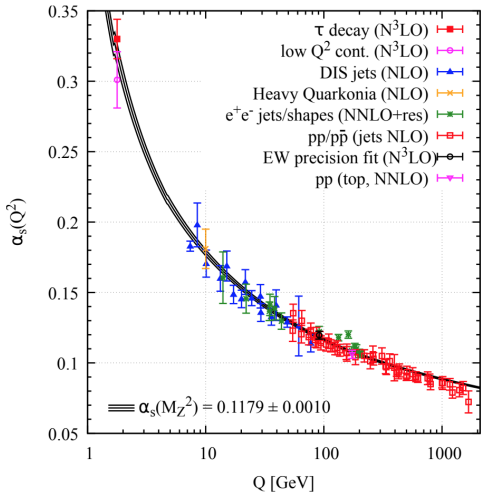


Figure 19. The running of coupling for the strong force, now plotted against energy scale. This figure is taken from the review of QCD by the Particle Data Group.

which it's measured. This is the story of renormalisation that we met previously in Section 2.3. But here is where there's a crucial difference between electromagnetism and the strong force: as we go to larger distances, the strong force gets *stronger*, not weaker. A sketch of the coupling is shown in Figure 18. (This should be contrasted with the running of the fine structure constant, as shown in Figure 10.) The experimental data for the running of the strong coupling is shown in Figure 19, now plotted against energy scale E, which is inversely related to length by $E = 1/r$.

How can we understand this intuitively? For electromagnetism, there was a simple physical picture in which the electric charge gets screened by particle-anti-particle pairs, and so appears smaller as we go to longer distances. For the strong force, the gluons themselves are doing the screening. Except they *anti-screen*, meaning that they cause the force to get stronger the further out we go!

In fact there is an intuitive way to understand this, although it's rather subtle. A clue can be found lurking back in the theory of electromagnetism. Recall that the Maxwell equations contain two parameters: $1/\epsilon_0$ characterises the strength of the electric force, while $1/\mu_0$ characterises the strength of the magnetic force. But these are not independent. They are related by

$$\epsilon_0 \mu_0 = \frac{1}{c^2} \tag{3.1}$$

with c the speed of light. This means that if the strength of the electric force gets weaker, then the magnetic force necessarily gets stronger, and vice versa.

There is a similar story for Yang-Mills. But now the gluons are doing the screening, and couple to both the chromoelectric and the chromomagnetic fields. It turns out that, because they are spin 1 particles, they screen the chromomagnetic fields more strongly than they screen the chromoelectric fields. In other words, the chromomagnetic part of the Yang-Mills interaction gets weaker as we go to larger distances. The relation (3.1) then tells us that the chromoelectric part of the interaction necessarily gets stronger. The upshot is that gluons anti-screen. If you want the gory details, the calculation can be found in Section 2.4 of the lectures on Gauge Theory.

So what is the strength of the strong force? At the energy scale $E \approx 100$ GeV, corresponding to a distance scale of $r \sim 10^{-17}$ m, we have

$$\alpha_s \approx 0.1 \quad \text{when} \quad E \approx 100 \text{ GeV}$$

Even at these fairly high energies, the strength of the force is an order of magnitude larger than QED. If we go to higher energies, or shorter distances, α_s decreases. In fact, as we go to arbitrarily high energies, the strength of the strong force vanishes, $\alpha_s \to 0$. This phenomenon is known as *asymptotic freedom*: it means that at high energies, or short distance scales, the strong force essentially disappears!

However, outside of particle colliders, everything that we observe in the world takes place at distance scales significantly larger than 10^{-17} m, and the strong force only gets stronger as we go to larger distances. But here there is another surprise. According to naive calculations, by the time you get to around $r \sim 10^{-15}$ m, or an energy scale of $E \sim 100$ MeV, the coupling constant appears to get infinitely large!

Above, I used the phrase "naive" calculations, because I have in mind the kind of perturbative Feynman diagram calculations that we described in the previous section. But, as we stressed, these diagrams only make sense when the coupling is small. As soon as the coupling is around $\alpha_s \approx 1$, the very complicated Feynman diagrams, involving lots of loops, are just as important as the simple Feynman diagrams and we have no control over the calculation. But this is exactly what happens in QCD! At very short distances, we're fine and we can do calculations. But at long distances, the theory becomes very challenging. The separation between "easy" and "hard" turns out to be around $r \sim 10^{-14}$ m to 10^{-15} m, but is usually expressed in terms of an energy scale known as the *strong coupling scale*, or *Lambda-QCD*,

$$\Lambda_{\rm QCD} \approx 200 \text{ MeV}$$

This is a characteristic energy scale of QCD. At energies $E \gg \Lambda_{\rm QCD}$, the strong force is not particularly strong and we can trust Feynman diagrams. But by the time we get to energies $\Lambda_{\rm QCD}$, the strong force lives up to its name. Most phenomena that are due to the strong force have an energy somewhere in the ballpark of $\Lambda_{\rm QCD}$

Before we describe some of these phenomena, it's worth pausing to mention that something unusual has happened here. The strength of the strong force, like QED, is characterised by a dimensionless coupling α_s. But the phenomena of renormalisation means that this coupling depends on scale, and the upshot of this is that we ultimately exchange a dimensionless number, α_s, for a dimensionful scale $\Lambda_{\rm QCD}$.

3.1.2 The Mass Gap

When we try to study the strong force on energy scales $E < \Lambda_{\rm QCD}$, corresponding to distance scales, $r > 10^{-14}$ m, we have a problem. The strong coupling means that the fields are wildly fluctuating on these scales, and our favourite method of Feynman diagrams is no longer useful. This also means that the classical equations of motion are no guide at all for what the quantum theory might look like.

The gluon is the first casualty of strong coupling. As we explained at the beginning of this section, the classical Yang-Mills equations suggest that the gluon should be massless. But the strong coupling effects change this. Instead, the gluon — which is a ripple of the Yang-Mills fields —has mass, given by

$$m_{\rm gluon} \approx \Lambda_{\rm QCD}$$

This is sometimes referred to as the *Yang-Mills mass gap*. The "gap" here is one between the ground state and the first excited state. For theories of massless particles, there is no gap because we can have particles of arbitrarily low energy. But for massive particles, the minimum amount of energy needed is $E = mc^2$.

To say that the Yang-Mills mass gap is difficult to prove would be something of an understatement. Demonstrating the mass gap is generally regarded as one of the major open problems in theoretical physics. Indeed, a million dollar Clay mathematics prize awaits anyone who succeeds. Although we do not have any rigorous (or even semi-rigorous) derivations of the mass gap, there is no doubt that it is a property of Yang-Mills. Our best theoretical evidence comes from computer simulations which, in this context, are called *lattice simulations*, reflecting the fact that spacetime is approximated by a grid, or lattice, or points. These simulations show unambiguously that the gluon is massive. You can read more about the lattice, and other approaches to Yang-Mills, in the lectures on Gauge Theory.

Finally, and most importantly, the existence of a mass gap is consistent with experiment, where no massless gluon is seen. Moreover, while the strong force is strong, it is also short-ranged. The characteristic energy scale $\Lambda_{\rm QCD}$ corresponds to a distance scale which in natural units (remember $\hbar = c = 1$) is

$$R_{\rm QCD} = \frac{1}{\Lambda_{\rm QCD}} \approx 5 \times 10^{-15} \text{ m}$$

To understand how this scale affects the world around us, we first need to throw in the final key ingredient: quarks.

3.2 Quarks

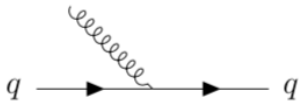
Usually in physics, we can get by without memorising lots of random names. The strong force is the exception, leaving us looking more like botanists than physicists. First, there are the many hundreds of names of different particles. But more important are names of groups of particles, each classifying a different property.

To kick things off, the fermions in the Standard Model are divided into two different types

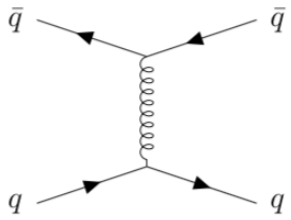
- Quarks: These are particles that feel the strong force.
- Leptons: These are particles that don't.

The leptons are the electron, muon, tau and three species of neutrino. (The name comes from the Greek $\lambda\epsilon\pi\tau\acute{o}\zeta$ meaning small.) Leptons don't interact with the $SU(3)$ Yang-Mills field, and we will ignore them for the rest of this section. In contrast, the six quarks — up, down, strange, charm, bottom and top — do feel the strong force.

In the language of Feynman diagrams, we denote the quarks as a solid line, with an arrow distinguishing quark from anti-quark. This is the same kind of line that we previously used to denote leptons, so we add a label q to show that it's a quark. The interaction between quarks and gluons is then described by the interaction vertex



When evaluating these Feynman diagrams, each vertex contributes a factor of the strong coupling, α_s. We can then use the Feynman diagrams to compute, say, the force between a quark and anti-quark. This comes from the following diagram:



If the quark and anti-quark are separated by a distance $r \ll R_{\rm QCD}$, then evaluating this diagram results in an attractive force that is very similar to the Coulomb force (2.4),

$$F(r) \sim \frac{\alpha_s}{r^2} \quad \text{when } r \ll R_{QCD} \tag{3.2}$$

Here α_s itself also depends on r, albeit logarithmically. We already saw a sketch of this dependence in Figure 18.

However, there's a catch: if the distance between the quarks is too big — bigger than R_{QCD} — then the language of Feynman diagrams stops working. As we increase the separation between quarks to distances greater than $R_{\rm QCD}$, the Coulomb-like expression (3.2) stops being the right one, and it instead changes to

$$F(r) \sim \text{constant} \quad \text{when } r \gg R_{QCD} \tag{3.3}$$

A constant force may not seem like much. But it gets exhausting. This is better seen if we look at the associated energy needed to separate a quark and anti-quark by distance R. For short distances, the energy takes the same form as in electrostatics,

$$V(r) \sim -\frac{\alpha_s}{r} + \text{constant} \quad \text{when } r \ll R_{QCD}$$

But when the quarks experience a constant force, the energy grows linearly

$$V(r) \sim \Lambda_{\rm QCD}^2\, r \quad \text{when } r \gg R_{QCD}$$

Clearly if you want to separate the quark-anti-quark pair by a long distance, then it costs an increasing amount of energy. In particular, it costs an infinite amount of energy to separate them an infinite distance. But taking, say, the anti-quark a long way away is tantamount to leaving the quark on its own. In other words, a solitary quark requires infinite energy! Quarks do not want to be alone: they only occur in bound states with other quarks or anti-quarks. This phenomenon is called *confinement*.

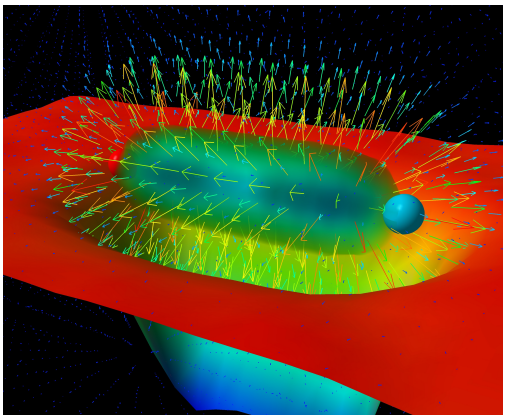
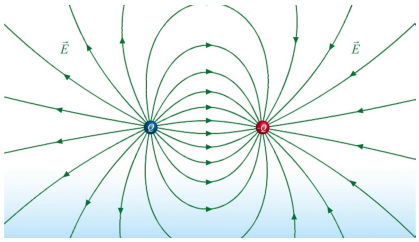


Figure 20. The chromoelectric flux tube between a quark and anti-quark in a meson state, from the QCD simulation of Derek Leinweber

Confinement, like the Yang-Mills mass gap, has so-far resisted a rigorous mathematical derivation. But we again have very clear evidence from numerical simulations, together with a handful of less-than-rigorous mathematical arguments, that confinement occurs. Moreover, this also gives us some intuition for what's going on.

First, let's recall what happens in electrostatics. If we separate a positive and negative electric charge by some distance r, then an electric field is set up between the two. The form of this electric field is shown on the right, and ultimately is responsible for the $F \sim 1/r^2$ Coulomb force law that the charges experience.



In Yang-Mills theory, the chomoelectric field takes a similar form if the quark and anti-quark are separated by a distance $r \ll R_{QCD}$. But as you increase the separation of the quark and anti-quark beyond the critical distance R_{QCD}, the form of the field changes. Instead of the field lines spreading out, the mass of the gluon forces them to bunch together into string-like configurations called *flux tubes*. This can be clearly seen in the computer simulation of QCD shown in Figure 20. It's as if the quark and anti-quark are joined by a piece of string. If you want to separate them further, you have to stretch the flux tube and this costs an energy $V(r) \sim r$ proportional to its length. This is responsible for the confinement of quarks.

3.2.1 Colour

The property that determines how particles experience the electromagnetic force is electric charge. The analogous property for the strong force is called *colour*[10]. Needless to say, this has nothing to do with the colour that we see. It is merely a label given by physicists who grew up without the classical education needed to name things in Greek or Latin.

While electric charge is just a number, colour charge is a little more involved: it is best thought of as a 3-dimensional vector $\boldsymbol{\omega}$ of fixed length. Here it's 3-dimensional because the Yang-Mills fields for the strong force are 3×3 matrices. This vector doesn't point in the three dimensions of space, but instead is something more abstract. If it points in different directions, we think of the quark as carrying a different colour. The exact translation between the vector and colour is pretty arbitrary, but we could take

$$\text{red: } \boldsymbol{\omega} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \text{green: } \boldsymbol{\omega} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \text{blue: } \boldsymbol{\omega} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

But the vector doesn't have to point exactly in one of these directions. For example, we could consider a vector $\boldsymbol{\omega} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ which should be thought of as a combination (or, in quantum language, a superposition) of red and green. Fortunately, we don't extend the colour analogy so far as to call this "muddy brown".

3.2.2 A First Look at Mesons and Baryons

Each quark carries a colour-vector $\boldsymbol{\omega}$, while the leptons do not. Confinement means that we do not see individual quarks, but they only appear in bound states so that the total colour charge vanishes. Any such composite particle, held together by the strong force, is referred to as a *hadron*. There are two ways in which hadrons can form,

- Mesons: These contain a quark and anti-quark. If the quark has colour vector $\boldsymbol{\omega}_1$ and the anti-quark colour vector $\boldsymbol{\omega}_2$, then they can combine as $\boldsymbol{\omega}_2^\dagger \cdot \boldsymbol{\omega}_1$ to form a colour-neutral state. Schematically, this looks like

  $$\text{meson} = \bar{r}r + \bar{g}g + \bar{b}b$$

  which should be read as "(anti-red)red + (anti-green)green + (anti-blue)blue". The flux tube for such a meson is shown in Figure 20.

[10]Scientists in the US work in units $u = 1$.

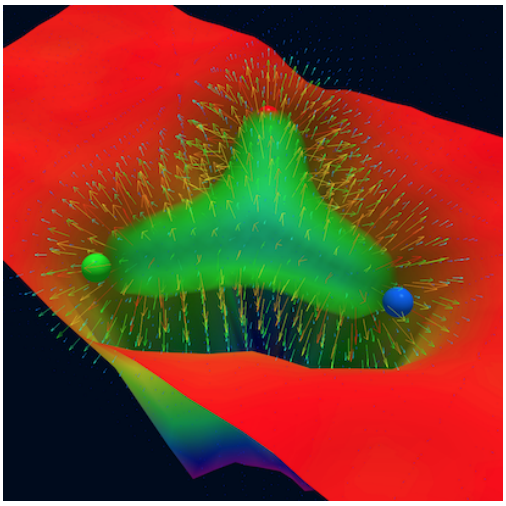


Figure 21. The flux tube between three quarks in a baryon state, from the QCD simulation of Derek Leinweber

- Baryons: These contain three quarks. If the three quarks have colour vectors $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$ then they combine as the triple product $\boldsymbol{\omega}_1 \cdot (\boldsymbol{\omega}_2 \times \boldsymbol{\omega}_3)$ to form a colour neutral state. Schematically, this looks like

$$\text{baryon} = rbg$$

  The fact that the baryon contains 3 quarks, rather than any other number, can be traced to the 3×3 matrices that describe the strong force. The flux tube for a baryon is shown in Figure 21.

To understand the collection of hadrons that emerges after confinement, we first need to look at the masses of quarks. In particular, we should compare the masses to the characteristic scale of the strong interactions, $\Lambda_{QCD} \approx 200$ MeV.

Three of the quarks have masses smaller than Λ_{QCD}.

$$\begin{aligned} m_{\text{down}} &= 5 \text{ MeV} \\ m_{\text{up}} &= 2 \text{ MeV} \\ m_{\text{strange}} &= 95 \text{ MeV} \end{aligned}$$

The up and down quark have masses significantly smaller than Λ_{QCD}, while the strange quark is only slightly smaller. Recall from our discussion in 1.2 that the Compton wavelength, $\lambda = \hbar/mc$, can be thought of as the size of a particle. Any particle with

$m < \Lambda_{QCD}$ necessarily has its size $\lambda > R_{QCD}$. That means that there's no way to bring two of these quarks closer than R_{QCD}, so these light quarks will only experience the confining force (3.3).

In contrast, the three heavier quarks have masses

$$\begin{aligned} m_{\text{charm}} &= 1.3 \text{ GeV} \\ m_{\text{bottom}} &= 4.2 \text{ GeV} \\ m_{\text{top}} &= 170 \text{ GeV} \end{aligned}$$

These are all much heavier than Λ_{QCD}. These three quarks all have Compton wavelength $\lambda \ll R_{QCD}$, so it makes sense for them to come close enough to experience the Coulomb-like force (3.2). This means that we might expect the spectrum of hadrons containing charm, bottom and top quarks to be a little different from those containing only the lighter quarks. Indeed, this turns out to be the case.

3.3 Baryons

We'll kick thing off with baryons, containing three quarks. To start, suppose that we have only the up and down quark to work with. There are various ways that we can combine these quarks. First, recall that the each quark has spin $\frac{1}{2}$. When combining quarks, we need to figure out what to do with their spins.

3.3.1 Protons and Neutrons

Suppose that we have two spins in one direction, and the third spin in the opposite direction. This will result in a baryon of spin $\frac{1}{2} + \frac{1}{2} - \frac{1}{2} = \frac{1}{2}$. There are two choices, which result in the two most familiar baryons: the proton (p) and the neutron (n). Their quark content and masses are

$$\begin{aligned} n\ (ddu) \quad & m_n \approx 939.57 \text{ MeV} \\ p\ (uud) \quad & m_p \approx 938.28 \text{ MeV} \end{aligned}$$

These are the two lightest spin $\frac{1}{2}$ baryons. Recall that the down quark has charge $-1/3$ and the up quark charge $+2/3$, so the proton has charge $+1$ while the neutron has no electric charge.

Already, there is something of a surprise here. The up and down quarks each have mass of a few MeV. Yet the proton and neutron each have mass of around 1000 MeV. How is this possible given that the proton and neutron are supposed to contain three quarks each?

The answer to this is simple: when we say that the proton and neutron each contain three quarks, we are hiding a more painful truth. The proton and neutron, and indeed all other hadrons, are in reality enormously complicated objects. The most accurate description is in terms of complicated, strongly interacting fields. In a particle language, we could describe them as containing many hundreds of quarks, anti-quarks and gluons, all interacting in a complicated fashion. The statement that the proton and neutron are composed of three quarks is really shorthand for the fact that they contain three more quarks than anti-quarks. These three additional quarks are sometimes called *valence quarks*, to distinguish them from the surrounding sea of quark-anti-quark pairs.

To put this in perspective, we could ask the following hypothetical question: suppose that the quarks were actually massless. What would the mass of the proton be? The surprising answer is that the mass of the proton would be more or less unchanged, still weighing in at a little less than 940 MeV! The same is true of the neutron. The masses of the proton and neutron care almost nothing about the mass of the three valence quarks: instead they are entirely dominated by the strong coupling scale $\Lambda_{\rm QCD}$ and their mass is few times $\Lambda_{\rm QCD}$.

You may have heard it said that the Higgs is responsible for all the mass in the universe. This is a fairly blatant lie. Later, in Section 4, we will learn that all elementary particles, including the quarks, do indeed get their mass from the Higgs boson. But the overwhelming majority mass in atoms is contained in the protons and neutrons that make up the nucleus, and this mass has nothing to do with the Higgs boson. It is entirely due to the urgent thrashing of strongly interacting quantum fields.

The irrelevance of the quark masses also has a more subtle implication. The masses of the proton and neutron are almost equal, despite the fact that the down quark is twice as heavy as the up. This reflects the fact that, at least as far as the strong interaction is concerned, the two particles behave almost identically. If we do an experiment with protons that is mediated by the strong force, then the same experiment performed with neutrons will yield exactly the same answer. This almost-symmetry of nature is referred to as *isospin*. Note, however, that it only holds for the strong force. The proton and neutron do not behave the same under the electromagnetic force, since the proton is charged while the neutron is not. Neither, it turns out, do they behave the same under the weak force.

There is, however, one place where the masses of the quarks are important. The fact that the down quark is heavier than the up quark does contribute a tiny amount to the mass and is the reason that the neutron is very slightly heavier than the proton. This

is important because it means that beta decay proceeds by a neutron decaying into a proton, rather than the other way around. We'll learn more about this in Section 4 where we discuss the weak force.

3.3.2 Delta Baryons

We could ask: why can't we have a baryon with, say, three up quarks? The answer to this lies in the Pauli exclusion principle. The full explanation is a little subtle, but the upshot is that we can have three up quarks in a baryon, but only if all their spins point in the *same* direction. (At first glance this might seem the wrong way around since, in chemistry, the Pauli exclusion principle dictates that electrons in the same orbital state have opposite spins. But quarks have that additional colour degree of freedom, and there is an anti-symmetry there which, in turn, requires a symmetric alignment of spins.)

When all the spins point in the same direction, the baryon itself has spin $\frac{1}{2}+\frac{1}{2}+\frac{1}{2}=\frac{3}{2}$ baryon. Now there are four choices, all of which are known as Delta (Δ) baryons. These particles have more or less equal mass, but different charges

$$\left.\begin{array}{ll} \Delta^{++} & (uuu) \\ \Delta^{+} & (uud) \\ \Delta^{0} & (udd) \\ \Delta^{-} & (ddd) \end{array}\right\} \quad m \approx 1232 \text{ MeV}$$

Here the superscripts, ++, +, 0 and - specify the electric charge of the particle.

We don't see Δ baryons floating around in the world. They have a lifetime of around 10^{-24} seconds, after which they decay, typically into a proton or neutron together with a meson called a pion. For example,

$$\Delta^{++} \to p + \pi^{+} \quad \text{and} \quad \Delta^{-} \to n + \pi^{-}$$

The lifetime of 10^{-24} seconds is much shorter than we can measure and particles with such short lifetimes are called, quite reasonably, *unstable*. Even when moving close to the speed of light Δ baryons don't travel far enough to register directly in particle detectors. Instead, they reveal themselves in more indirect means as so-called *resonances* in certain experiments. We'll describe this further in Interlude C.1.

The lifetime of the Δ baryons is actually the characteristic timescale of the strong force: $T_{QCD} = R_{QCD}/c \approx 10^{-24}$ seconds. If anything happens due to the strong force, it usually happens on roughly this timescale.

3.3.3 Strangeness

Let's now consider baryons that we can construct from the first three quarks: up, down and strange.

We'll first consider baryons with spin $\frac{1}{2}$. In addition to the proton and neutron, we now have four further baryons that contain a single strange quark, called sigma (Σ) baryons

$$\begin{aligned} \Sigma^-(dds) \quad & m \approx 1197 \text{ MeV} \\ \Sigma^0(dus) \quad & m \approx 1193 \text{ MeV} \\ \Sigma^+(uus) \quad & m \approx 1189 \text{ MeV} \end{aligned}$$

and the lambda (Λ) baryon

$$\Lambda^0(dus) \quad m \approx 1116 \text{ MeV}$$

Again, the superscript labels the electric charge of the baryon. You may have noticed that the quark content of the Σ^0 and Λ^0 are the same. The difference lies in the details of the wavefunctions for the up and down quarks. (Technically, some different minus signs mean that all Σ baryons have isospin 1, while the Λ has isospin 0.)

There are also two types of baryons that contain two strange quarks, called cascade, or xi (Ξ) baryons

$$\begin{aligned} \Xi^-(dss) \quad & m \approx 1322 \text{ MeV} \\ \Xi^0(uss) \quad & m \approx 1315 \text{ MeV} \end{aligned}$$

None of these baryons are familiar from our everyday experience. This is because they again decay, typically to protons and pions. However, here there is a surprise: although these new baryons have a mass in the same ballpark as the Δ's, they live for significantly longer. In particular, the $\Sigma^\pm$, the Λ^0 and the $\Xi^{0,-}$ all live for a whopping 10^{-10} seconds.

Now, 10^{-10} seconds may not sound like much. Indeed, it's difficult to imagine having a rich and fulfilling life in this time. But it's an aeon compared to the 10^{-24} seconds that the Δ baryons live. This is a puzzle: why do these new baryons have a such a comparatively long life, even though their masses are comparable to the Δ?

A partial answer to this is to invoke a new conservation law. We know that electric charge is conserved in all interactions. It turns out that there is another quantity – *strangeness* – which is conserved. Or, at the very least, almost conserved. Strangeness is simply a count of the number of strange quarks.

Here "almost conserved" means conserved by the strong interaction. The strong interactions cannot change the number of strange quarks and so particles like $\Sigma^{\pm}$, Λ^0 and $\Xi^{0,-}$ do not decay straight away. The decay only proceeds through the weak interaction, and this takes significantly longer. We'll describe how these decays occur in Section 4. In contrast, particles like the Δ baryons decay directly through the strong interaction, and this happens much faster.

(As an aside: there's always one complication. It turns out that, among the collection of strange baryons, there is one which is unstable: the $\Sigma^0 \to \Lambda^0 + \gamma$ with a lifetime of around 10^{-20} seconds. But this is allowed by the strong force because the number of strange quarks is unchanged. The Λ^0, as we've seen, then waits another 10^{-10} seconds before it too decays.)

As a general rule of thumb, hadrons can decay through one of the three forces: strong (like the Δ's), electromagnetic (like Σ^0) or weak (like $\Sigma^{\pm}$, Λ^0 and Ξ.). The lifetimes of these particles reflect the decay process:

- Strong decay: $\sim 10^{-22}$ to 10^{-24} seconds.
- Electromagnetic decay: $\sim 10^{-16}$ to 10^{-21} seconds.
- Weak decay: $\sim 10^{-7}$ to 10^{-13} seconds.

Where you sit within each range depends on other factors, such as the relative masses of the parent and daughter particles. Particles that live for up to 10^{-10} seconds are referred to (I think, somewhat tongue in cheek) as *stable*. In contrast, any particle that lasts 10^{-20} seconds or shorter is, like the Δ baryon, referred to as *unstable* or a *resonance*.

It should be clear from the discussion, however, that there's nothing very qualitatively different between a stable particle like the Λ and a resonance like the Δ. Both will decay in less than the blink of an eye. But a lifetime of 10^{-10} seconds mean that, with good technology, you can take a photograph of the particle's track in a cloud chamber or bubble chamber. You can see many such photographs in Interludes B and C. When a particle leaves such a vivid trace, it's hard to deny its existence. In contrast, we're never going to take a photograph of something that lasts 10^{-20} seconds. But that doesn't mean that it's any less real! It just leaves its signature in more subtle ways.

3.3.4 The Eightfold Way

There is a clear pattern to the masses of the spin $\frac{1}{2}$ baryons. To highlight this, we can place the 8 baryons in the following shape:

$$n\ (ddu) \qquad\qquad p\ (uud)$$

$$\Sigma^-\ (dds) \qquad\qquad \Sigma^0, \Lambda^0\ (dus) \qquad\qquad \Sigma^+\ (uus)$$

$$\Xi^-\ (dss) \qquad\qquad \Xi^0\ (uss)$$

The particles are arranged in rows of increasing strangeness and, correspondingly, in increasing mass. The particles in the top row have $m \approx 940$ MeV; those in the second row $m \approx 1190$ Mev; and those in the final row $m \approx 1320$ MeV. Meanwhile, particles in the same \ diagonal have equal charge.

We see that adding a strange quark increases the mass of a baryon by roughly 140 ± 10 MeV. This can be largely, but not entirely, accounted for by the mass of the strange quark, $m_{\text{strange}} \approx 95$ MeV.

Recall that there is an approximate symmetry, relating up and down quarks, called isospin. If we squint, we could enhance this to a larger, and even more approximate, symmetry relating up, down and strange quarks. This subsumes the previous isospin symmetry, but isn't quite as a good because the strange quark is significantly heavier than the other two. Indeed, it only holds if we're willing to pretend that 1320 MeV $\approx$ 940 MeV. Nonetheless, it is true that, at least as far as the strong force is concerned, all 8 baryons in the table have more or less the same properties. For example, a given meson will scatter off all 8 in pretty much the same way. The symmetry, first proposed by Gell Mann and, independently, Ne'eman in 1961 is called $SU(3)$ flavor symmetry or, more poetically, the *eightfold way*.

The $SU(3)$ flavor symmetry relating the three lightest quarks is not to be confused with the $SU(3)$ colour symmetry that underlies the strong force. It's just that the number 3 appears a bunch of times in the Standard Model. (Don't read anything mystic into this. It's just a small number.) With decades of hindsight, the eightfold way appears more accidental than fundamental. Nonetheless, it was important historically as a useful organising principle in cataloging the many hundreds of hadrons that were discovered.

The same almost-symmetry is sitting within the baryons of spin $\frac{3}{2}$. These can be placed in the following pattern, again organised by increasing strangeness

$$\Delta^- \ (ddd) \qquad \Delta^0 \ (ddu) \qquad \Delta^+ \ (duu) \qquad \Delta^{++} \ (uuu)$$

$$\Sigma^{\star -} \ (dds) \qquad \Sigma^{\star 0} \ (dus) \qquad \Sigma^{\star +} \ (uus)$$

$$\Xi^{\star -} \ (dss) \qquad \Xi^{\star 0} \ (uss)$$

$$\Omega^- \ (sss)$$

Once again, the masses in each row are roughly constant: $m \approx 1232$ MeV in the first row; $m \approx 1385$ MeV in the second; $m \approx 1533$ MeV in the third; and with the Ω^- weighing in at $m \approx 1672$ MeV. In each case, the increase is again roughly 140 to 150 Mev and is due to the extra strange quark.

Notice that the middle 7 baryons in this table have the same quark content as the spin $\frac{1}{2}$, strangeness 1, baryons that we met previously. These should be thought of excitations of these previous baryons, with the spin of one of the constituent quarks flipped, changing the overall spin from $\frac{1}{2}$ to $\frac{3}{2}$.

The real novelties in the table above are the three outliers, in which all quarks are the same. As we mentioned above, the Pauli exclusion principle prohibits the existence of spin $\frac{1}{2}$ baryons with three identical quarks, so they appear here for the first time. Two are particularly important: the Δ^{++} was the first particle to be found without charge ± 1 (or 0) and helped enormously in piecing together the story of the underlying quarks. The Ω^- baryon, meanwhile, holds a special place in the history of science because Gell-Mann used the simple quark model described above to predict its mass and properties before it was discovered experimentally. He therefore followed Mendeleev and Dirac in predicting the existence of a "fundamental" particle of nature (where, as should by now be clear, the meaning of the word "fundamental" is time-dependent).

There are eight baryons with spin $\frac{1}{2}$, and ten baryons with spin $\frac{3}{2}$, with the former lending its name to the "eightfold way" used to describe the whole enterprise. But, underlying this set-up are just 3 quarks — up, down and strange — and, as we mentioned above, an approximate $SU(3)$ symmetry that rotates them. Why do we start with the number 3, and end up with 8 and 10 baryons respectively? In fact, there is a good reason for this, although it's difficult to explain without going into the mathematics of group theory. It turns out that any group has a bunch of different ways in which it

can express itself, known as *representations*. And these representations come only with very specific numbers. Although it's not obvious, the numbers 8 and 10 are naturally associated to the group $SU(3)$.

Finally, I mention that, in principle, it should be possible to calculate the masses of the proton, neutron and all other baryons directly from our knowledge of QCD dynamics. While this is somewhat beyond what we can do with pen and paper, we can simulate QCD on a computer, and get pretty accurate predictions for the masses of baryons that we've seen above, certainly good to the 5% level. The same is true for the mesons that we will meet in the next section. There is now no doubt that the complexity seen in the hadron spectrum can be entirely explained by the dynamics of QCD.

More Baryons

The lists above do not exhaust the baryons that have been discovered. There are further baryons containing charm and bottom quarks. For example, in addition to the Σ^+, comprised of uus, there is also a Σ_c^+ comprised of uuc and Σ_b^+ comprised of uub, and similar stories for many of the other baryons. It is, however, difficult to argue for any approximate symmetry among these baryons, since the mass of the heavy quarks is much greater than those of the lighter quarks and greater than Λ_{QCD}.

There are also excited states of all these baryons, in which the quarks orbit each other, not dissimilar to the way in which the electrons orbit the proton in the excited states of the hydrogen atom.

There are not, however, baryons containing top quarks. The top quark is so heavy that such baryons are predicted to decay in around 10^{-25} seconds, even faster than the characteristic timescale $T_{QCD} \approx 10^{-23}$ seconds of the strong force. This means that such "top baryons" decay before they even form. Needless to say, none have been observed.

3.4 Mesons

We now turn to mesons, bound states of a quark and an anti-quark. Many hundreds have been discovered. Here we describe some of the most important.

3.4.1 Pions

We will again start by assuming that we've only got up and down quarks to play with. Once again, we can put the quark spins in the same direction, or in opposite directions.

We start by putting the spin of the quark and anti-quark in opposite directions. This results in mesons with spin 0. There are three such mesons that we can build from the up and down quarks, known as *pions*. Their masses and quark content are given by

$$\begin{array}{cc} \pi^+ \ (\bar{d}u) & m \approx 139 \text{ MeV} \\ \pi^0 \ \frac{1}{\sqrt{2}}(\bar{u}u - \bar{d}d) & m \approx 135 \text{ MeV} \\ \pi^- \ (\bar{u}d) & m \approx 139 \text{ MeV} \end{array}$$

The π^- is the anti-particle of π^+ and the two have exactly the same mass. The neutral pion, π^0 is a combination of up and down quarks as shown. It has no electric charge, but a very similar mass to the $\pi^\pm$. The similar masses reflect the isospin symmetry which says that, as far as the strong force is concerned, the up and down quarks have the same properties.

Despite their similar masses, the neutral and charged pions have rather different lifetimes. The neutral pion decays through the electromagnetic force to two photons

$$\pi^0 \to \gamma + \gamma$$

It has a lifetime of around 10^{-17} seconds. In contrast, the charged pions π^+ and π^- decay through the weak force. We'll see in Section 4 that they typically decay to a muon and a neutrino

$$\pi^+ \to \mu^+ + \nu_\mu \quad \text{and} \quad \pi^- \to \mu^- + \bar{\nu}_\mu$$

They live for 10^{-8} seconds, an eternity in the subatomic world and much longer than any of the baryons except the proton and neutron.

There is one, very important characteristic that distinguishes mesons from baryons. Mesons, made of a quark and anti-quark, have integer spin and are therefore bosons. Baryons, made of three quarks, have half integer spin and are therefore fermions.

Back at the beginning of Section 2, we explained that fermions are "matter particles" while bosons are "force particles". It should come as no surprise to learn that baryons, like protons and neutrons, are matter particles. After all, you're made up of them. But it may be less familiar to hear that mesons, like the pion, are force particles. What force do they mediate?

The answer to this is quite lovely: the pions give rise to an attractive force between the baryons. In particular, they give rise to an attractive force between any collection of protons and neutrons. It is this force that binds the protons and neutrons together inside the nucleus.

The existence of a scalar particle, mediating the interaction between protons and neutrons, was predicted by Yukawa in 1935, more than a decade before the pion was discovered. Yukawa observed that a massive scalar particle would give rise to a Coulomb-type force, but with an exponential suppression due to the mass. The potential energy between any two particles takes the form

$$V(r) \sim -\frac{e^{-mr}}{r} \tag{3.4}$$

This is called the *Yukawa potential*. When $m = 0$, it agrees with the more familiar Coulomb potential. For very small distances, $r \ll 1/m$, it is more or less the same as the Coulomb potential. But the force drops off very quickly at distances $r \gg 1/m$. Yukawa had the simple insight that the force that binds the nucleus together should exert itself over distances comparable to the size of the nucleus. From this, he predicted the mass of the pion to be around 200 times the mass of the electron, or about 100 MeV. As we see, he was not far off.

The force that binds the nucleus together is usually simply referred to as the strong nuclear force. But it would be better to give it a different name — say "mesonic force", or "Yukawa force" — to highlight the fact that it is really a residual, secondary effect. At the fundamental level the strong force is mediated by gluons and binds quarks together. But it binds them together in two ways: one to create baryonic matter particles, and another to create mesonic force particles. The upshot is that there are two layers to the strong force: we start with one force and a set of matter particles — gluons interacting with quarks — and end up with a very different force and a new set of matter particles — the mesonic force interacting with protons and neutrons. In this sense, both the particles in the nucleus, and the force that holds them together, are *emergent* phenomena, arising from something more fundamental underneath.

We might, then wonder: do similar transformations await us as we go to yet smaller scales? Could there be some other, very different degrees of freedom on the smallest scales, from which the Standard Model emerges. The answer, of course, is: we don't know.

3.4.2 The Eightfold Way Again

Let's now throw the strange quark into the mix. In addition to the three pions, there are five further spin 0 mesons we can build. (Actually six, but one of these, the η' has slightly different properties so we'll postpone its discussion for now.) These are a

collection of particles called *kaons*

$$\begin{aligned} K^+ \ (\bar{s}u) \quad & m \approx 494 \text{ MeV} \\ K^0 \ (\bar{s}d) \quad & m \approx 498 \text{ MeV} \\ \bar{K}^0 \ (\bar{d}s) \quad & m \approx 498 \text{ MeV} \\ K^- \ (\bar{u}s) \quad & m \approx 494 \text{ MeV} \end{aligned}$$

The charged kaons are, like the charged pions, relatively long lived: their lifetime is around 10^{-8} seconds. They too decay via the weak force.

The lifetime of the neutral kaons is a somewhat more complicated story: rather curiously they appear to have two different lifetimes, either 10^{-7} seconds or 10^{-10} seconds, depending on how you count! That's kind of weird. Moreover, rather unexpectedly, it turns out to be an indirect hint of one of the deepest properties of the Standard Model: the fundamental laws of physics are not the same if you run them forwards and backwards in time! We will postpone discussion of this topic to Section 4.3.4.

Finally, there is one meson that is a combination of up, down and strange quarks called, uninspiringly, the eta (η) meson. Its mass and quark content are

$$\eta \ : \quad \tfrac{1}{\sqrt{6}}(\bar{u}u + \bar{d}d - 2\bar{s}s) \qquad m \approx 548 \text{ MeV}$$

This is similar in spirit to the π^0 meson: for each type of quark there is also an anti-quark sitting within the meson. This allows it to decay quickly, in 10^{-19} seconds, to two photons.

Together with the pions, these 8 mesons sit in a pretty pattern governed by the eightfold way symmetry relating the three fundamental quarks. We again construct rows of increasing strangeness, where a strange anti-quark $\bar{s}$ counts as negative strangeness:

$$K^0 \ (\bar{s}d) \qquad\qquad K^+ \ (\bar{s}u)$$

$$\pi^- \ (\bar{u}d) \qquad\qquad \pi^0, \eta \qquad\qquad \pi^+ \ (\bar{d}u)$$

$$K^- \ (\bar{u}s) \qquad\qquad \bar{K}^0 \ (\bar{d}s)$$

We haven't written the quark content of the π^0 and η only in an attempt to keep table looking vaguely aesthetic.

Viewed purely in terms of masses, the eightfold way looks less convincing for the mesons than for the baryons. An additional strange quark or anti-quark now costs roughly 350 MeV, and the kaons and η are two to three times heavier than the pions. Despite this, it turns out that the masses of all these mesons are well understood theoretically, with the eightfold way an important part of the derivation.

Although I won't recount the full story here, there is part of it that is worth highlighting. Recall that the masses of the proton and neutron are set almost entirely by Λ_{QCD}, rather than the masses of the up and down quark. Indeed, as we mentioned previously, if the up and down quarks were massless then the mass of the proton and neutron would remain pretty much unchanged. At first glance, it looks like the same might be true of the mesons above. After all, the mass of the pions is much closer to $\Lambda_{QCD} \approx 200$ MeV than to $m_{\rm up}$ and $m_{\rm down}$, which are a few MeV. However, it turns out that this guess is completely wrong! If the mass of the up and down quarks vanish, then the mass of the pions would also vanish! Similarly, if the mass of the up, down and strange quarks all vanished, then all 8 mesons above would have vanishing mass. This may seem like a theoretical curiosity since, in the real world, the masses of the quarks are distinctly not zero. Nonetheless, it turns our that this simple observation is enough to govern many of the properties of these mesons. You can read more about this beautiful story in the chapter on *chiral symmetry breaking* in the lectures on Gauge Theory.

There is one final spin 0 meson that is a bit of a loner. It is even given a rubbish name. The eta-prime (η') meson has quark content and mass given by

$$\eta' \quad \tfrac{1}{\sqrt{3}}(\bar{u}u + \bar{d}d + \bar{s}s) \qquad m \approx 958 \text{ MeV}$$

The eta-prime is significantly heavier than the other 8 mesons, despite having a very similar quark content to the eta meson. There is again a beautiful story behind this associated to the so-called *axial anomaly*, one of the more subtle and deep aspects of quantum field theory. This too is described in the lectures on Gauge Theory.

More Mesons

So far we have only discussed spin 0 mesons, in which the spins of the constituent quark and anti-quark point in opposite directions. We could also arrange for these spins to be aligned. In this case, we end up with mesons of spin 1.

For example, there is a collection of three spin 1 mesons containing only the up and down quarks. These are called rho mesons, and can be viewed as excitations of the three pions. They have masses ~ 770 MeV and decay quickly to pairs of pions.

There are many many further mesons, including excitations of those already mentioned, and mesons that involve the charm and bottom quark. Once again, the top quark decays too quickly and does not form mesons.

Two sets of these mesons deserve a special mention. The first is *charmonium*, a bound state of charm and anti-charm quark. It also goes by the dual name J-psi (J/ψ),

$$J/\psi \ (\bar{c}c) \qquad m \approx 3.1 \text{ GeV}$$

Its lifetime is around 10^{-21} seconds. The discovery of this particle in 1974 was the first glimpse of the charm quark and will be described in Section C.3.

There are a collection of lighter mesons that contain just a single charm quark. These are called (somewhat peculiarly) *D-mesons*. The lightest are:

$$\begin{aligned} D^0 \ (c\bar{u}) \qquad & m \approx 1865 \text{ MeV} \\ D^+ \ (c\bar{d}) \qquad & m \approx 1869 \text{ MeV} \end{aligned}$$

These are remarkably long lived particles, with the D^+ living 10^{-12} seconds, and the D^0 about half this time. The long lifetime is because these particles decay only through a somewhat subtle property of the weak force. We will learn more about this in Section 4.3.

Similarly, the bottom quark was first discovered in *bottomonium*, also known as the upsilon (Υ)

$$\Upsilon \ (\bar{b}b) \qquad m \approx 9.5 \text{ GeV}$$

This has a lifetime of 10^{-20} seconds. Once again, it is neither the lightest nor the longest lived meson containing a b-quark. The lightest *B-mesons* are

$$B^+ \ (u\bar{b}) \text{ and } B^0 \ (d\bar{b}) \qquad m \approx 5280 \text{ MeV}$$

Despite being significantly heavier, they actually live (very) slightly longer than the D-mesons, with a lifetime of around 1.5×10^{-12} seconds. Again, this is down to intricacies of the weak force.

C Interlude: The Rise of the Machine

From the 1920s onwards, it became clear that the alpha particles emitted in radioactivity, with their 5 MeV of energy, would not suffice to understand the nucleus. As we saw in Interlude B, in the short term most discoveries would come courtesy of cosmic rays. But, long term, new accelerator technology was needed.

The cathode ray tube is, in many ways, the first particle accelerator, albeit one built before the constituents were even known to be particles. The key idea is a simple one: if you drop a large voltage over some distance then charged particles will pick up speed. You can then use the resulting beam to smash into other things, as Röntgen did in his discovery of X-rays.

A variant of this idea was taken forward by John Cockcroft and Ernest Walton. Working in Cambridge in 1932, under the ever watchful eye of Rutherford, they built a voltage multiplier capable of accelerating protons to 700 keV. They then used this to great effect, inducing the first artificial transmutation of a nucleus,

$$p + {}^7\text{Li} \longrightarrow 2\alpha$$

In more popular terminology, they succeeded in splitting the atom. These days, the "high-tension laboratory", where Cockcroft and Walton performed their later experiments, has been converted into the "Cockcroft lecture theatre", where we teach Newtonian mechanics and special relativity to first year undergraduates.

Both the cathode ray tube and the Cockcroft-Walton accelerator are linear accelerators: the charged particles move in a straight line. The next great breakthrough – due to Ernest Lawrence – was a simple one: make the particles bend.

C.1 The Cyclotron

Lawrence was inspired by a simple fact in classical mechanics. Take a charged particle of mass m and charge q, restricted to move in the (x, y)-plane, with a magnetic field B in the z-direction. If you give the particle an initial kick of speed v, then it moves in a circle and comes back to its starting position in time

$$T = \frac{2\pi m}{qB} \tag{C.1}$$

The lovely fact is that this time doesn't depend on the speed v! If you set the particle off with a bigger velocity, then it will travel in a bigger circle, but always come back in the same time.

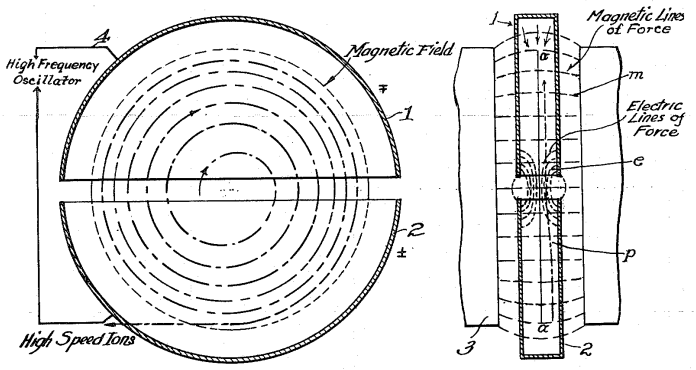


Figure 22. The principle behind the cyclotron, from Ernest Lawrence's 1934 patent application.

This is the key principle behind the *cyclotron*. The charged particles move perpendicular to a fixed magnetic field. They are trapped in two boxes, called *dees*, so named because they are shaped like the letter D. The two dees are placed back to back, with a small gap between them like so: ↀD. A voltage is placed across the gap, ensuring that the particles are accelerated every time they cross from one dee to the other. The rub, of course, is that first they cross the gap in one direction, then in the other. This means that if you want them to be accelerated each time, rather than decelerated, then you have to flip the polarity of the voltage in the time it takes them to travel a semi-circle. The good news, as we've seen in (C.1), is that this time doesn't depend on how fast the particles are going. This means that you can tune the AC voltage to a frequency, resonant with the particles and the particles will speed up, travelling in ever wider circles until they spit out the end where they can be used for whatever purpose is needed.

A prototype cyclotron, known as the 9-inch, was first constructed in Berkeley in 1930. By 1932, Lawrence and his student, Stanley Livingston, had succeeded in building an 11-inch machine that could reach energies of 1 MeV and were able to quickly reproduce the Cockcroft-Walton results splitting the atom. For the rest of the 1930s, Lawrence exhibited a single-minded focus on reaching higher and higher energies. By 1936 he reached 8 MeV, by 1939 20 MeV. Indeed, such was his desire to reach higher energies, it seemed to barely occur to Lawrence that he should, perhaps, occasionally pause to do some science with his machines.

Furthermore, there is a limit to how far the key equation (C.1) can be pushed. As protons reach energies of around 25 MeV, the effects of special relativity have to be

taken into account and their mass starts to increase with velocity. This knocks the timing, since the faster electrons now take slightly longer to make their orbit. To compensate, one has to tune the magnetic field to keep the timing in sync. This kind of machine, still with particles spiralling outwards but with a varying magnetic field, goes by the catchy name of a *synchrocylcotron*, or SC for short.

Soon after the war, flush with money from the military, Lawrence completed his first SC, a 184-inch monster machine with a magnet that weighed 10,000 tons and accelerated alpha particles to 380 MeV, later increased to 720 MeV. The question was: what to do with it?

The Neutral Pion

The neutral pion π^0 holds two claims to fame. It was the first particle to be predicted on grounds of symmetry. And it was the first particle to be discovered in a collider on Earth, rather than in a cosmic ray shower.

First the experimental situation. Recall from Interlude B.3 that the charged pion was discovered in cosmic ray showers by Powell and his team in 1947. In fact, rather embarrassingly, by that point Lawrence's SC had been producing pions for over a year. But, such was the focus on reaching higher energies, no one had put any thought into building detectors. It was only when Cécil Lattes, one of Powell's collaborators, moved to Berkeley in 1948 that they placed photographic emulsion plates in the accelerator and found pions in great numbers.

The neutral pion is a different matter altogether. Since it carries no electric charge, it leaves no tracks and neither cloud chambers nor photographic emulsions can be used to find it. Instead, it can only be seen by more indirect means through its decay to two photons

$$\pi^0 \longrightarrow \gamma + \gamma$$

The smoking gun is the detection of two simultaneous photons, whose energy, momentum and angular momentum correlations can be traced to the decay of a single particle. The neutral pion was discovered in this manner in 1950 by Steinberger, Panofski and Steller, using the Berkeley SC.

Next, the theory. In Section 3 we explained the *eightfold way*, the idea that there is an approximate symmetry between the up, down and strange quarks which leads to patterns in the masses of baryons and mesons. We also briefly mentioned a precursor to this idea, first pointed out in the early 1930's by Heisenberg. This is the idea that, at

least as far as the strong force is concerned, the neutron and proton behave in almost identical fashion. They have similar mass. Moreover, the binding force between nn, pp and np is more or less the same. Even with the very limited experimental data of the early 1930s, Heisenberg intuited that this was important. At the time the symmetry between the proton and neutron was called *isotopic spin*[11], these days shortened to just *isospin*.

However, the idea of isospin runs into trouble when you appreciate the obvious: the proton and neutron have different electric charges. If the strong force is mediated by charged pions $\pi^{\pm}$ alone, as was thought by the late 1930s (recall, people still thought that the muon was responsible at this point!) then charge conservation meant that the interactions experienced by the proton and neutron would necessarily be different. The way out was suggested by the theorist Nicholas Kemmer in 1938: you get to keep isospin symmetry, but only if there exists a third, neutral pion π^0.

Resonances

Neutral particles are not the only ones that are hard to see. Any hadron or meson that decays through the strong force, rather than the weak force, will have a lifetime of around 10^{-24} seconds. That's not a huge amount of time, even by the standards of particle physics. Even allowing for relativistic time dilation, these particles are not going to travel far enough to snap a photograph of them. Instead, we need more indirect methods to detect them.

The method of choice is to observe the effect of these new particles on the old. To explain this I first need to introduce a new concept: that of *cross-section*.

When two elementary particles come close to each other, there is some probability that they will interact. This interaction may result in them scattering off each other, or transforming – even if only briefly – into some other particle. Roughly speaking, the *cross-section* is the probability that they interact in some way, rather than pass through each other.

We can also speak less roughly. The cross-section, as the name suggests, is actually an area. As such it can't quite be a probability (which must be dimensionless) but it's closely related. The idea is that the cross-section is the area – or size – that the particles present to each other as they approach. As an analogy, if you're throwing balls in an attempt to hit some target, it's the cross-sectional area of the target that

[11]This is, of course, a daft name. A much better name would be *isobaric spin*, since isobars are elements with the same atomic mass but different combinations of protons and neutrons.

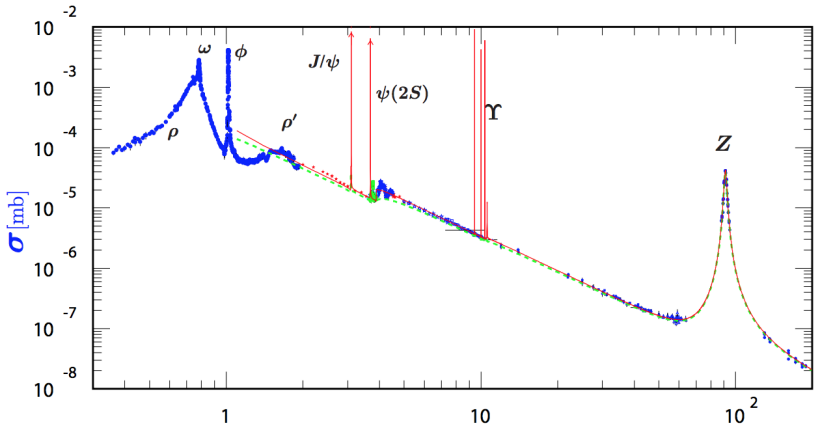


Figure 23. A collection of data from many experiments studying $e^+e^- \longrightarrow$ hadrons. The horizontal axis is the centre of mass energy, measured in GeV. A number of different mesons can be clearly seen at low energy, with the Z-boson at high energy. This plot was taken from the particle data group plots of cross-sections and related quantities.

will determine your success. A barn door has a bigger cross-section than a beer can. This, in turn, means that games where you try to hit a barn door are somewhat less entertaining than those where you try to hit a beer can.

Importantly, the cross-section for some interaction depends on the energy of the colliding particles. Typically, in particle physics one finds that the cross-section drops as the energy of the incoming particles increases. But, every now and then, one sees a pronounced bump in the cross-section. This bump is called a *resonance* and is the telltale sign that there's a new, third particle, appearing in the story.

The bumps appear when the energy of the incoming particles is tuned to the mass of some new particle. This allows for a new interaction, in which the two incoming particles briefly morph into the new one which will then, typically, subsequently decay. These decay products may be the original particles, in which case it will just look like they've scattered off each other, or they may be something new entirely.

The early data on resonances is not particularly clean. (We'll give an example below.) However, the idea is clearly illustrated in Figure 23 which collects together the results of many decades worth of experiments of e^+e^- collisions, where the end products are hadrons. The horizontal axis depicts the (log of the) energy, measured in GeV. The vertical axis depicts the (log of the) cross-section. You can see the overall downwards trend of the cross-section as the energy increases, but most striking are the various

peaks. At lower energies these peaks correspond to meson states that we met in the last section. Way up, at close to 100 GeV, we see the Z-boson that mediates the weak force. We'll discuss this more in the following section.

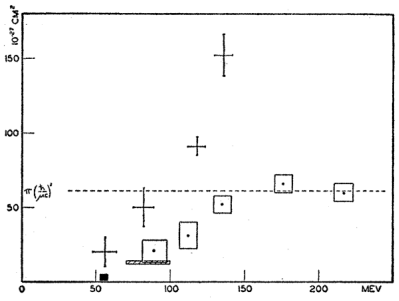
The shape of the resonance contains a lot of information about the underlying particle. The energy at which the resonance occurs tells us the mass of the particle. Meanwhile, the width of the resonance tells us its lifetime: the bigger the width, the quicker the particle decays. Returning to Figure 23, you can see that the width is barely discernible on the red spikes in the middle. This is because, as we saw in the last section, these are the lightest states containing charm and bottom quarks respectively. The presence of these new quarks limits their decay options, resulting in a much longer lifetime than might naively be expected.

Many of the particles that we'll meet as these lectures progress were detected through their resonance effect on scattering. You can read more about the basics of resonances in quantum mechanical scattering in the lectures on Topics in Quantum Mechanics.

Delta Baryons

The first novel particle detected as a resonance was the collection of four Δ-baryons. Viewed through today's lens of the quark model, they are Δ^{++} (uuu), Δ^{+} (uud), Δ^{0} (udd) and Δ^{-} (ddd). Each has mass 1232 MeV and a lifetime of around 5×10^{-24} seconds.

The year was 1952, the place Chicago. Enrico Fermi and his team had built a synchrocyclotron, based on the principles introduced by Lawrence. They extracted a beam of pions from the machine and directed it at hydrogen gas to watch how the pions scattered off protons. They didn't see anything as distinctive as a bump. They did, however, see a rise in the cross-section. Most strikingly, there was a clear difference in the cross-sections for π^+ (shown as crosses in the data to the right) and π^- (shown as rectangles)[12] . This needed an explanation.



[12]This figure is taken from the paper Total Cross Sections of Positive Pions in Hydrogen" by Anderson, Fermi, Long, and Nagle.

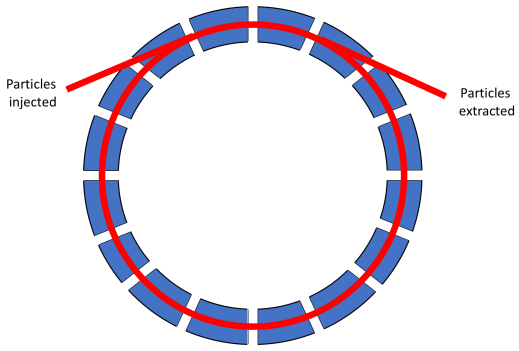
In fact, this effect had been anticipated earlier by a theorist called Keith Brueckner. He looked at the three scattering processes

$$\begin{aligned}\pi^+ + p &\longrightarrow \pi^+ + p \\ \pi^- + p &\longrightarrow \pi^- + p \\ \pi^- + p &\longrightarrow \pi^0 + n\end{aligned}$$

By invoking a set of intermediate states, which we now call the Δ-baryons, he argued that the processes could proceed through the creation of Δ^{++} or Δ^0 and should occur with relative probabilities in the ratio 9:2:1. This was what the Chicago team observed. Further analysis of the angular distribution of the scattering showed that the intermediate states must have spin 3/2. Historically, the discovery of the Δ^{++}, with its striking +2 charge, played an important role in elucidating the underlying structure of quarks.

C.2 The Synchrotron

Cylcotrons, with the particles spiraliing outwards, can reach energies no higher than 1 GeV or so. The next phase in accelerator development was the *synchrotron*. Here, particles travel in a circle of fixed radius. Their path is again bent by magnets, and accelerated in the gaps between the magnetics by electric fields. To keep the particle travelling in a circle, the magnetic field must be synchronised with the particle's velocity, so that the magnetic field becomes stronger as the particles pick up speed. A cartoon picture of the synchrotron is shown on the right.



The development of the synchrotron brought many new technical challenges, including the difficult issue of stabilising the beam. Key to the whole endeavour is the concept of *phase stability*, in which an alternating voltage is placed over the gaps acts to ensure that all particles converge on the same speed, with the faster ones slowing slightly and the tardy ones picking up speed. This results in particles sitting in bunches, rather than the continuous beam of the earlier cyclotron.

Another difference from the cyclotron is that you don't get to start the particles from rest. They must be injected into the synchrotron at some other accelerator. This, then, is the fate of accelerators upon retirement: they become injection machines for the next generation.

The early proton synchrotrons had the cool names, before the push for dull acronyms became too strong to ignore. In Brookhaven, the Cosmotron reached energies of 3 GeV; in Berkeley, the Bevatron 6 GeV. These began a long succession of machines, culminating more than half a century later, with the Tevatron at Fermilab reaching 1 TeV and the LHC at CERN reaching 13 TeV.

More Anti-Matter

The 6 GeV reached by the bevatron was not chosen arbitrarily. It was built with a particular science case in mind: the creation of anti-protons. The 6 GeV threshold allows their production through

$$p + p \longrightarrow p + p + \bar{p} + \bar{p}$$

The first proton on the left-hand side was in the beam; the second proton on the left-hand side was sitting in a fixed copper target. The resulting debris mostly consists of pions, with the occasional anti-proton lying within. The challenge was to find them.

You might have thought that the best way to detect anti-protons would be to watch them annihilate with protons. In fact, it turned out to be significantly simpler to identify them through their mass and charge. To this end, the experimenters first set up a series of magnets, designed to deflect unwanted positively charged particles and focus only negatively charged particles with very specific momentum into the detectors. The first set of detectors consisted of *scintillation counters*, an instrument for measuring the photons emitted by the ionised tracks left by a charged particle. Two scintillation counters were placed 12 m apart. Both anti-protons and pions in the beam had the same momentum, which meant that the heavier protons were slower and so took longer to travel the 12 m between the detectors. A whole 10^{-8} seconds longer. That was the first clue.

Next, the beam entered a pair of *Cerenkov detectors*. These detect an effect known as *Cerenkov radiation* which is emitted when particles travel through a material faster than light can travel through that material. The first Cerenkov detector fired whenever the particle was travelling too fast to be an anti-proton. The second Cerenkov detector had a special design which meant that it fired *only* for a small window of velocities, tuned to that of the anti-proton, and so failed to fire for the faster mesons.

With this elaborate set-up, over the course of two weeks in October 1955, Emilio Segrè and his team of Owen Chamberlain, Clyde Wiegand and Tom Ypsilantis found 60 anti-protons, nestled among 3.5 million pions. They announced the discovery at a press conference on October 19$^{\text{th}}$ 1955.

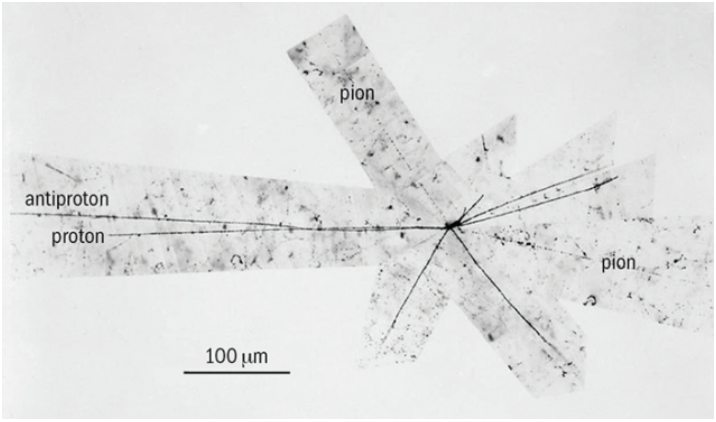


Figure 24. A photographic emulsion showing $p\bar{p}$ annihilation. The original (in vertical orientation) can be found in the 1956 Nuovo Cimento paper "On the Observation of an Antiproton Star in Emulsion Exposed at the Bevatron" by Chamberlain et. al.

This story also has some human twists. Indeed, as machines got bigger, so too did the opportunity for machinations. The clever magnetic lens design used in the experiment was due to an Italian physicist, Oreste Piccioni who, at the time, was a visitor in Berkeley. Although the group lead by Segrè adopted this design, they were reluctant to allow Piccioni to join their team when he returned to Berkeley full time in the summer of 1955. Instead, Piccioni joined a rival team, also searching for the anti-proton, lead by Edward Lofgren.

The two teams took it in turns to use the bevatron, two weeks on, two weeks off. The machine broke in the middle of Segré's second run and they could take no data. When the machine was finally repaired, nice-guy Lofgren yielded his time to allow Segrè to complete his run. Two weeks later they announced the discovery anti-proton. Four years after that, Segrè and Chamberlain collected their Nobel prize. Nice guys, it turns out, rarely fared well in the increasingly ruthless world of experimental particle physics.

However, there was some joy for both Lofgren and Piccioni. One year later, they were among the team who discovered the anti-neutron through the annihilation processes

$$p + \bar{p} \longrightarrow n + \bar{n}$$

It appears, however, that this brought little comfort to Piccioni. Some years later, he sued Segré and Chamberlain for $125,000 of their Nobel prize money. The case made it to the US supreme court, before being dismissed.

Figure 25. This bubble chamber photograph, taken at the Lawrence Berkeley Laboratory, shows a high energy γ ray colliding with an electron to produce an e^+e^- pair, which spiral in opposite directions. The original electron receives a huge kick and flies off to the right. At some point it emits a gamma ray which then turns into a second e^+e^- pair, this time with higher energy, visible as the V signature to the right of the picture.

Mesons and Baryons in Bubbles

The next leap forward was in detector technology. In the 1950s, Donald Glaser proposed the idea of a *bubble chamber*, a successor to the cloud chamber. The detector is filled with a liquid – ideally liquid hydrogen – that is kept under pressure at constant temperature, slightly below its boiling point. Just before the particles pass through, the pressure is reduced by a small amount, lowering the boiling point so that the liquid is *super-heated*, meaning that it remains in the liquid phase even though the temperature is above boiling. As a charged particle passes through, it gives the liquid the nudge it needs to start boiling, leaving behind a trail of bubbles. Examples of the particles' graceful arcs, as they spiral in an applied magnetic field, can be seen in Figures 25 and 26.

In addition, this was a time of increased automation. The number of events that could be recorded in a bubble chamber was far greater than in previous detectors. Teams of highly skilled, poorly paid women, scouring photographic emulsions for interesting forks and kinks just wasn't going to cut it anymore. Instead, both analysis and data storage required the use of computers. Experiments could be done in one institution, and analysed elsewhere.

Results came quickly. First, a collection of vector (i.e. spin 1) mesons were discovered, starting with the ρ and ω. The η meson was discovered at Johns Hopkins university, using data borrowed from Berkeley. (η translates to the letter H, which is short for "Hopkins" apparently.)

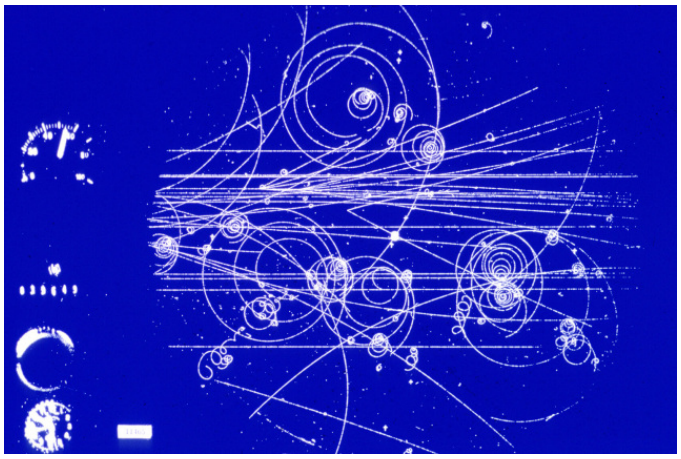


Figure 26. This photo was taken in 1960 at CERN. A stream of π^- mesons enter from the left. One of these pions hits a proton in the liquid hydrogen spraying new particles. Among these is a neutral Λ baryon, which doesn't leave a track but reveals its existence a little to right as it decays into a proton and pion, producing the characteristic V shape seen in the middle of the picture. The results of this decay have high energy and travel in straight line. Other, lower energy charged particles spiral in the magnetic field.

Baryons were also seen in quantity, both long-lived strange baryons and, indirectly, shorter lived resonances. An important breakthrough happened in 1964, when the Ω^- baryon was discovered. The Ω^- contains three strange quarks and its mass, lifetime and decay modes had been predicted earlier using the quark model.

The Ω^- discovery is shown in Figure 27. An incoming kaon collides with a proton in the liquid, yielding

$$K^-\left(\bar{u}s\right) + p(uud) \longrightarrow \Omega^-\left(sss\right) + K^+\left(u\bar{s}\right) + K^0(d\bar{s})$$

There is then a succession of further baryon decays, with

$$\begin{aligned}
\Omega^-\left(sss\right) &\longrightarrow \Xi^0\left(uss\right) + \pi^-\left(\bar{u}d\right) \\
\Xi^0\left(uss\right) &\longrightarrow \Lambda^0\left(uds\right) + 2\gamma \\
\Lambda^0\left(uds\right) &\longrightarrow p\left(uud\right) + \pi^-\left(\bar{u}d\right)
\end{aligned}$$

Each of these decays changes strangeness by -1, and so happens only through the weak force. Correspondingly, the lifetime is around 10^{-10} seconds for each, long enough for them to leave a trail a bubbles.

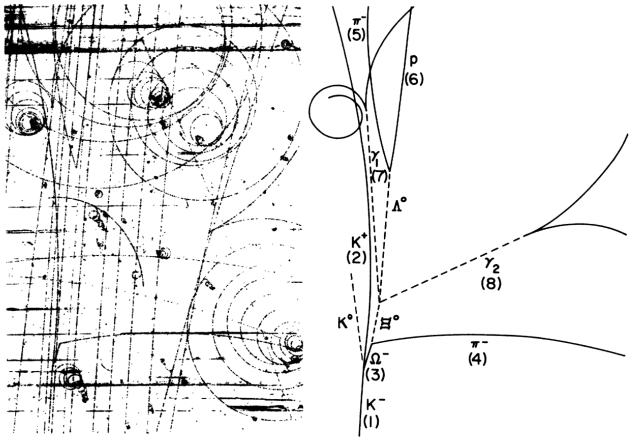


Figure 27. The original bubble chamber picture, and an accompanying line tracing, showing the discovery of the Ω^- baryon. This is taken from the 1964 paper "Observation of a hyperon with strangeness minus three" by Barnes et. al. Here "et al" refers to an additional 30 authors, heralding the large collaborations that were to come.

C.3 Quarks

The cornucopia of hadrons discovered throughout the 50s and 60s gave enough clues to lead to the quark model. The observed masses and lifetimes could be roughly accounted for by postulating three different types of constituent spin $\frac{1}{2}$ particles – up, down and strange – each of which carries three different internal degrees of freedom called colour. But that still left open the question of whether these quarks were real entities, or just mathematical accounting tricks.

Most physicists at the time assumed that quarks were merely a useful fiction. Their concern was the obvious one: if quarks are real, then why don't we observe them in isolation, where their fractional electric charge would stand out like a sore thumb. Gell-Mann ends the 1964 paper in which he first proposed quarks with the sentence

> "A search for stable quarks of charge $-\frac{1}{3}$ or $+\frac{2}{3}$ and/or stable di-quarks of charge $-\frac{2}{3}$ or $+\frac{1}{3}$ at the highest energy accelerators would help to reassure us of the non-existence of real quarks.

Zweig, who independently had the same idea in the same month – January 1964 – referred to quarks as "aces" and was marginally more optimistic. The final sentence of his paper reads

> "There is also the outside chance that the model is a closer approximation to nature than we may think, and that fractionally charged aces abound within us."

The situation was resolved by a series of experiments that, although initially designed to study resonances of the proton, turned out to be perfectly placed to instead explore its inner structure. These experiments, which ran from 1967 to 1973, took place at the Stanford Linear Accelerator, better known as SLAC, a 3 km long machine that accelerated electrons to 20 GeV, effectively by getting them to surf the crest of an electromagnetic wave.

The SLAC experiments didn't see anything as striking as fractional electric charge. Indeed, their data was, at first, murky and complicated. However, over the course of several years it became increasingly apparent that their results could only be explained if the electrons were scattering off some point-like constituent inside the proton and neutron. As we now explain, these experiments involve a process known as ...

Deep Inelastic Scattering

By the mid 1960s, it was apparent that the proton is not a point-like object but has a size of around 10^{-15} m. This was seen, for example, in the *elastic* scattering of electrons off protons. Here, the word "elastic" means that the electron has the same energy after the collision as before. For example, the Geiger-Marsden experiment that uncovered the structure of the atom involved the elastic scattering of alpha particles off the nucleus.

In contrast, in *inelastic* scattering, the electron collides with more destructive force, knocking the proton into a higher excited state, such as a spin $\frac{3}{2}$ baryon, or breaking it apart completely. This process can roughly be characterised as

$$e + p \longrightarrow e + \text{other stuff}$$

where we don't too much care about the other stuff. We care only about what happens to the electron. If the electron comes in at very high energies then it buries deep into the proton where it can probe whatever lies inside.

If, as was originally thought, the proton was structureless, then the cross-section for electron-proton scattering should rapidly decrease with energy. And this is indeed seen in elastic scattering. However, when the electron energy gets high enough to enter the inelastic regime, something different occurs. The first clue that something was afoot was simply that more electrons were scattered at low angles than expected. However, when the data was plotted in a particular way, something more surprising stood out.

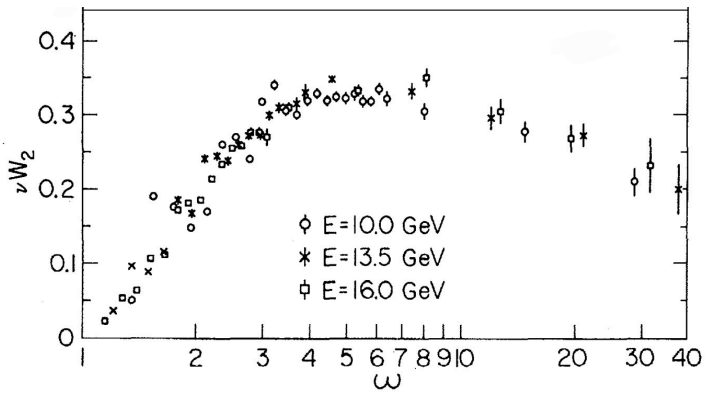


Figure 28. It's neither pretty, nor obvious, but this data provided the first hint of the existence of quarks. The proton structure function is plotted on the vertical axis, while $\omega = 1/x$ is on the horizontal axis. This graph is taken from the 1969 paper "Observed Behavior of Highly Inelastic Electron-Proton Scattering".

To explain this, I need to first describe a little bit of scattering theory. The cross-section for electron-proton scattering depends on an object called the *proton structure function*. And this, in turn, depends on the energy E of the incoming electron, the energy E' of the outgoing electron and the angle θ by which the electron is deflected. In general, any cross-section depends on these variables in two combinations

$$\nu = E - E' \quad \text{and} \quad q^2 = 4EE'(1 - \cos\theta)$$

The surprise was that, at high energies, the proton structure function doesn't depend on both ν and q^2: it depends only on the dimensionless ratio

$$x = \frac{q^2}{2M\nu}$$

where M is the mass of the proton. The initial data, shown in Figure 28, plots the proton structure function for scattering by a small angle ($\theta = 6^\circ$ in this case.) The fact that the data for three different energies all lie on the same curve is showing that the structure function depends only on the combination x. This is known as *Bjorken scaling*, after the theorist Bjorken who first suggested this behaviour (and first suggested that the experimenters plot their data in this unlikely manner.)

The question, of course, is what does scaling mean? The intuitive idea, largely due to Feynman, is that scaling is telling us directly that the electron is scattering off a point-like object inside the proton. Very roughly speaking, this is because the

proton structure function doesn't depend on any scale, because x is a dimensionless variable. But the only object with no scale is a point. (Feynman's original argument is characteristically creative and involves going to a frame of reference in which the point-like object in the proton has infinite momentum.)

Feynman referred to the point-like objects inside the proton as *partons*, noticeably avoiding the term "quark" in his original paper. I do not know if this was for the purposes of scientific agnosticism or scientific antagonism. (Gell-Man and Feynman were colleagues, collaborators and, perhaps above all, rivals!) Either way, the omission was appropriate. Simple models in which the proton and neutron are each composed of three quarks do not provide a good fit to the data. This is because, as we explained in the previous section, the proton and neutron are themselves enormously complicated objects. The cartoon picture in which they each contain three quarks is a long way from reality and ignores the morass of gluons and quark-anti-quark pairs that also sit inside. Deep inelastic scattering provides the tool to probe this complexity. Good fits to the data could only be achieved by including partons that are gluons and quark-anti-quark pairs, in addition to the three valence quarks.

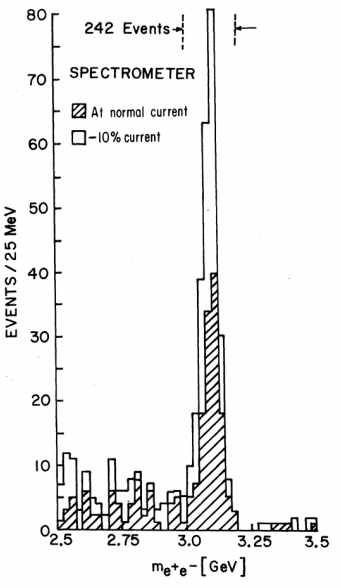
Deep inelastic scattering also provides a method to indirectly test the electric charges carried by the partons. It turns out that the cross-section is proportional to the sum of the squares of the charges of the partons. For a proton, with uud quarks, this sum gives $\left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 = 1$, while for a neutron with udd, it is $\left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 = \frac{2}{3}$. The simplest quark model then predicts that the cross-section for neutrons should be $\frac{2}{3}$ that of protons. After taking into the account the many subtleties described above, this is confirmed by experiment.

Although the parton model provided a good explanation for the experimental results, there was one mystery that remained. The partons that lead to exact scaling behaviour are free particles. Subsequent deviations from scaling suggested that there were some interactions between the partons, but these were necessarily small. Yet how could partons be confined within the proton if the interactions were so small? This issue was resolved by the discovery of asymptotic freedom in Yang-Mills like theories by Gross, Wilczek and, independently, Politzer in 1973.

Charmed

In November 1974, a new quark was found. This was the charm. It came as a surprise to many physicists and went a long way towards cementing belief in the quark model.

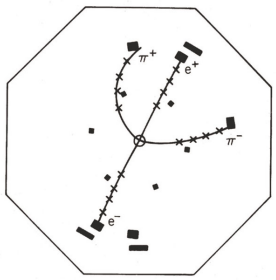
The discovery was made simultaneously on the West coast in SLAC and on the East coast in Brookhaven. Since both teams got to name the particle, it now goes by the double-barrelled J/ψ.



The J/ψ weighs in at 3.1 GeV, with a lifetime of almost 10^{-20} seconds. This is much longer than expected for a particle that is so heavy, a fact that reveals itself in the very narrow resonance shown on the right.

It didn't take long to understand that that this resonance is so narrow because it contains a new quark-anti-quark pair, bound as $\bar{c}c$ meson. It is now also known as *charmonium*. In fact, in many ways the charmed mesons are easier to understand than their lighter cousins. Because the charm quark is so heavy (about 1.3 GeV), it has a very small Compton wavelength. This means that two charm quarks in a meson sit so closely together that they are in the asymptotically free regime of QCD, where the coupling strength is fairly small. This makes it easier to understand their properties.

Within a few weeks of the original J/ψ discovery, a collection of further resonances had been found, all agreeing well with theoretical expectations. These new resonances were found only at SLAC so their name took preference, and they are called ψ'. In a cute twist, the reconstructed event decay of a ψ' takes the eponymous form shown on the right.



While most physicists had not expected the charm quark, there were some who had previously argued for its existence on theoretical grounds. The most compelling was due to Glashow, Iliopoulos and Maiani (or GIM) who, in 1970, analysed a subtle property of the weak force known as *flavour changing neutral currents*. (We'll describe some basic properties of the weak force in Section 4.) An example of this phenomenon is the decay $K^0 \longrightarrow \mu^+ + \mu^-$. This happens very rarely, but it was difficult to understand why: theoretical expectations suggest that a process involving the exchange of an up quark and W-boson should give a much larger decay rate. The GIM mechanism proposed the existence of a charm quark, with charge $+2/3$, and showed that this gives a contribution to the decay almost exactly cancels the contribution from the up quark.

After the discovery of the charm quark, many things fell into place. The key ideas underlying the Standard Model, from asymptotic freedom of the strong force, to symmetry breaking of the weak force (described in Section 4) had all been developed previously. But the discovery of the charm quark prompted a synthesis of these ideas, with all phenomena described by four quarks, two charged leptons, and two neutrinos, coupled through forces associated to $SU(3) \times SU(2) \times U(1)$. The inappropriately modest name *Standard Model* dates from this time. For this reason, physicists who lived through the discovery of charm sometimes refer to this period as the *November Revolution*.

Colliders

The discovery of charm also involved an important experimental leap: colliders.

All the accelerators that we've discussed so far take a beam of particles and smash it into a fixed target. This has the obvious advantage that it's easy to hit a fixed target. But it also has a disadvantage because conservation of momentum means that any particles you create fly off at close to the speed of light. It's not so much the fact they're moving that's the problem, but that much of the beam energy is wasted because it goes into the kinetic energy of the final product. If you succeed in accelerating particles of mass m to an energy E which then hits a fixed target, then the energy available to create new particles is $\sqrt{2mE}$. Clearly it would be better we could make use of more of that precious beam energy.

These problems go away in a *collider*. This consists of two beams with equal and opposite momentum, which are then brought together at one or more intersection points where the collision takes place.

Needless to say, the technological hurdles in getting this to work are formidable. Not least is the problem of *luminosity*, meaning the number of collision events taking place. The density of particles in each beam is significantly less than that of a fixed target and, correspondingly, the number of collisions is greatly reduced. To compensate for this, each beam should consist of many pulses of particles, with collisions happening repeatedly and often. This, in turn, requires that the beams have long lifetimes. Typical numbers involve beam lifetimes of several hours, allowing them to circulate 10^{10} times or so. These accelerators also go by the name of *storage rings*.

The discovery of charm at Brookhaven used a standard fixed target experiment. A beam of electrons ploughed into a beryllium target, and they found their J particle as the huge peak in the centre of mass energy of e^+e^- pairs.

In contrast, the SLAC team used an e^+e^- collider, called SPEAR, short for the Stanford Positron Electron Asymmetric Rings. These storage rings took electrons and positrons from the linear collider and brought them into head-on collisions. They discovered their ψ particle as a resonance in the cross-section for $e^+ + e^- \longrightarrow$ hadrons.

As an aside, the acronym "asymmetric" in the acronym SPEAR dates from an earlier proposal in which e^+ and e^- were accelerated in different rings. After a budget cut, the design changed to a single ring, but the acronym stayed.

Round Three

Back in 1932, there were 100 days where physicists could revel in a simplistic world containing only electrons, protons and neutrons. Then the discovery of positrons burst their bubble.

In the 1970s, physicists had a little over a year in which they could believe in a nice symmetrical world with two generations of fermions, each containing two quarks, an electron-type particle and a neutrino. Late in 1975, a group working at the SPEAR experiment found something odd and entirely unexpected. Their original paper doesn't beat around the bush, opening with the blunt statement:

> "We have found 64 events of the form
>
> $$e^+ + e^- \longrightarrow e^\pm + \mu^\mp + \geq 2 \text{ undetected particles}$$
>
> for which we have no conventional explanation."

The paper suggests that these events could be due to a new charged, heavy lepton or to a new charged heavy boson. It took some years to realise that the former is the case: the original e^+e^- pair collide to form what we now call the *tau leptons*,

$$e^+ + e^- \longrightarrow \tau^+ + \tau^-$$

The taus subsequently decay as, for example,

$$\tau^- \longrightarrow \nu_\tau + e^- + \bar{\nu}_e \quad \text{and} \quad \tau^+ \longrightarrow \bar{\nu}_\tau + \mu^+ + \nu_\mu$$

giving rise to the observed signature.

The discovery of the τ lepton upset the balance. It would be another 20 years until it was restored, with the bottom quark, top quark and tau neutrino filling out the set. The bottom quark was discovered not long after. In 1977, a fixed target experiment at Fermilab found a strong, narrow resonance at 9.5 GeV, the upsilon Υ with quark content $\bar{b}b$.

Finding the top quark was another matter. A rough guesstimate for its mass can be made by a quick glimpse at the first five quarks:

$$\begin{array}{lll} m_d = 4.7 \text{ MeV} & m_s = 96 \text{ MeV} & m_b = 4.2 \text{ GeV} \\ m_u = 2.2 \text{ MeV} & m_c = 1.3 \text{ GeV} & m_t = ? \end{array}$$

Given this pattern, what would you guess for m_t? Perhaps 40 GeV? In the 1970's, a couple of e^+e^- colliders failed to find the top quark at 30 GeV. In the 1980's, a p^+p^- collider failed to find it at 80 GeV. By this time it was clear that the top quark was so heavy that it would not form a detectable $\bar{t}t$ meson state like J/ψ or Υ. This is because (see Section 4) it could decay directly to a W-boson through

$$t \longrightarrow W^+ + b \quad \text{and} \quad \bar{t} \longrightarrow W^- + \bar{b}$$

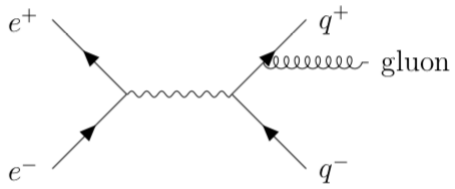
This decay happens in a shorter time scale than that associated to the strong force. This means that, if we want to find the top quark, we need to find clear evidence for the decay products that it leaves behind.

Because the top quarks are created as $\bar{t}t$ pairs, they will decay to $\bar{b}b$ pairs, together with a W^+W^- pair. The W-bosons themselves then decay, sometimes to quarks $W^\pm \to \bar{q}q$ and sometimes to leptons $W^\pm \to l^\pm\nu$ where $l = e, \mu$ or τ. The question is: how to see these decay products?

We know what charged leptons look like, and we know that neutrinos are just going to escape unnoticed. So the real question is: how do we see the quarks? Because these are confined, we can't see quarks directly. However, when quark-anti-quark pairs are formed from collisions, something rather dramatic happens. Each one of the pair flies off in a different direction but, because they hate to be alone, they pull further quark-anti-quark pairs from the vacuum as they go. The end result is that each quark morphs into a collection of hadrons, all moving in roughly the same direction. This is called a *jet* and the phenomenon of quarks turning into a multitude of mesons and baryons is called *hadronisation*.

Jets were first seen in SPEAR e^+e^- collisions in 1975 and proved yet more evidence for the reality of quarks. With a lot of work, it is sometimes possible to go backwards and, from the jet, reconstruct the kind of quark that started it.

As an aside to our main story, in the late 1970s, the first indirect evidence for the existence of the gluon came through a process associated to the following Feynman diagram:



One of the emitted quarks radiates a gluon, but the gluon can no more live on its own than the quarks. The result is 3 jets; two from the quarks, one from the gluon.

Back to the story of the top, the cleanest signature of top production occurs when both W bosons decay into leptons, giving two leptons and two jets

$$t\bar{t} \longrightarrow l^+ + l^- + \nu + \bar{\nu} + \bar{b} + b$$

This has a low background from other processes, but happens only rarely in top decay as well. Another option is that one of the W-bosons decays into leptons and the other into quarks, giving

$$\begin{aligned} t\bar{t} \longrightarrow \quad & l^+ + \nu + \bar{q} + q + \bar{b} + b \\ \text{or} \quad & l^- + \bar{\nu} + \bar{q} + q + \bar{b} + b \end{aligned}$$

which has a single lepton, and four jets. This process, which is depicted in Figure 29, happens more often, but also has a higher background. The final decay, with no lepton and six jets gets swamped by the background.

After a number of hints in the early 1990s, the discovery of the top quark was finally announced in 1995. It was found at the Tevatron in Fermilab, a p^+p^- collider that reached energies of 1 TeV. Collisions took place at two similar but complimentary detectors, situated in different places around the ring, and were analysed by two independent rival collaborations known as D$\varnothing$ (pronounced dee-zero) and CDF. A top-quark event from the CDF collaboration is shown on the next page.

The top took so long to track down because it was much heavier than anyone had anticipated. It finally weighed in at

$$m_t = 170 \text{ GeV}$$

This remains the heaviest fundamental particle that we know. Why is it so much heavy? We have, I think it's fair to say, no idea. Surely it is telling us something important.

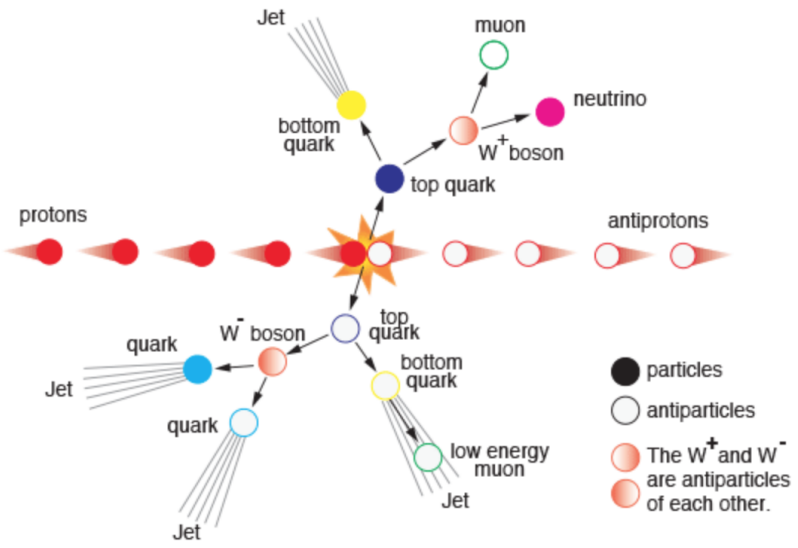
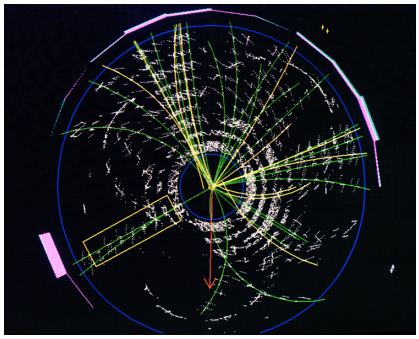


Figure 29. A schematic picture of a top quark event, taken from the Fermilab website.



With the tau, bottom and top quark in place, all that was left was the tau neutrino. This was discovered by the DONUT experiment at Fermilab in 2000. They used the Tevatron to create a beam of neutrinos which included ν_τ. These were then directed at nuclear emulsion targets where they collided with iron nuclei. The tell-tale sign of a ν_τ neutrino is the creation of a τ lepton, which leaves a small 1 mm track in the emulsion. Four such events were seen.

4 The Weak Force

It is now time to turn to the weak force. For many years, there was just one manifestation of the weak force, namely beta decay

$$n \;\to\; p + e^- + \bar{\nu}_e$$

We now know that this can be understood in terms of the down quark decaying into an up quark, electron and anti-electron neutrino

$$d \;\to\; u + e^- + \bar{\nu}_e$$

There is nothing in the strong force or electromagnetism that would allow one type of quark to morph into another. We need to invoke something new.

That "something new" turns out to be something old and familiar. Nature has a tendency to re-use her good ideas over and over again, and the weak force is no exception. Like the strong force, it too is described by Yang-Mills theory. The difference is that the matrices are now 2×2 instead of 3×3. However, as we'll see in this section, it's not just the strengths of the forces that differ and the weak and strong forces manifest themselves in a very different manner.

The three forces of Nature together provide the foundation of the Standard Model. In mathematical language these forces are characterised by a group

$$G = SU(3) \times SU(2) \times U(1)$$

where the 3×3 matrix fields of $SU(3)$ describe the strong force, and the 2×2 matrix fields of $SU(2)$ describe the weak force. However, rather surprisingly the fields of $U(1)$ do *not* describe the force of electromagnetism! Instead, they describe an "electromagnetism-like" force that is called *hypercharge*. The combination of $SU(2)$ and $U(1)$ is sometimes referred to as *electroweak theory*. We will learn in Section 4.2 how electromagnetism itself lies within.

The weak force has few obvious manifestations in our everyday life and, in many ways, is the most intricate and subtle of all the forces. It is intimately tied to the Higgs boson and, through that, the way in which elementary particles get mass. Moreover, both the most beautiful parts of the Standard Model, and those aspects that we understand least, are to be found in the weak force.

Figure 30. Parity violation of Chien-Shiung Wu.

4.1 The Structure of the Standard Model

When describing the strong force, we saw that it affects some particles (we call these quarks) while leaving other untouched (we call these leptons). Our first task now should be to describe which particles are affected by the weak force.

You might think that we could simply list those particles that feel the weak force. But, as we will see, things are not quite so straightforward. It turns out that the weak force acts on exactly half of the particles in the universe. But it does so by acting on exactly half of each and every particle!

4.1.1 Parity Violation

There is one defining characteristic of the weak force and hypercharge that differentiates them from the strong force (and from electromagnetism). They do *not* respect the symmetry of *parity*.

This fact was discovered by Chien-Shiung Wu, on a cold winters day, in New York City, in December 1956. Wu's experiment was technically challenging, but conceptually very simple. She placed a bunch of Cobalt atoms in a magnetic field and watched them die. Cobalt undergoes beta decay

$$^{60}\text{Co} \;\rightarrow\; ^{60}\text{Ni} + e^- + \bar{\nu}_e + 2\gamma$$

with a half-life of around 5.3 years. The two photons arise because cobalt first decays to an excited state of the nickel nucleus, which subsequently decays down to its ground state emitting two gamma rays. The whole point of the magnetic field was to make sure that the nucleon spins of the atoms were aligned. Wu discovered that the electrons were preferentially emitted in the opposite direction to the nucleon spin

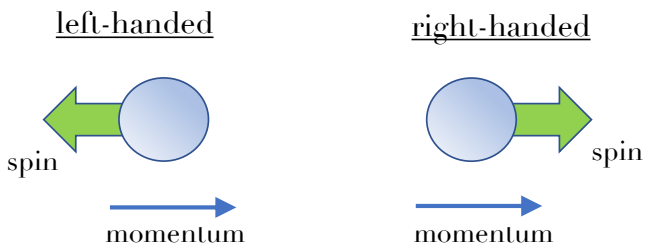


Figure 31. The handedness of a massless particle is determined by the relative direction of its spin and momentum.

This may sound innocuous, but we now realise that it was one of the most significant discoveries in all of particle physics. The key point is that when we say the nuclei spin in a given direction, we mean in a *right-handed* sense. Take your right-hand and curl your fingers round in the way the nuclei are spinning: then your thumb points in the "direction of spin" which, as Wu observed, is opposite to the motion of the electrons.

Now suppose that, for some bizarre reasons, Wu looked at her experiment reflected in a mirror. The directions of the electrons would remain unchanged, but the spin of the nuclei would be reversed, because right-hands are reflected into left-hands. This means that, viewed in a mirror, Wu would have come to the opposite conclusion: the electrons would be preferentially emitted in the *same* direction as the nuclei spin.

To put this in context, stare at the two photographs of Wu and her experiment shown in Figure 30. If you look closely you can tell which photograph is the original and which is flipped about an axis. (For example, one way to do this is to note that the writing is only legible on the left-hand picture.) Wu discovered that the same is true of the laws of physics at a fundamental level: you can tell if you're looking directly at sub-atomic particles, or viewing them reflected in a mirror. There are things that can happen in a mirror that cannot happen in our world! This property is known as *parity violation*.

How can we write down theories which violate parity, meaning that they describe a world which looks different when reflected mirror? The key is something that we learned back in Section 2.1.4: any massless spin $\frac{1}{2}$ particle decomposes into two pieces, called *left-handed* and *right-handed*. Recall that a right-handed particle is one whose spin is aligned with its momentum, while a left-handed particle has spin and momentum anti-aligned. This distinction only makes sense for massless particles since they travel at the speed of light and so all observers, regardless of they're own motion, agree on the direction of spin and momentum.

To write down a theory which violates parity is then straightforward: we simply need to ensure that the left-handed particles experience a different force from the right-handed particles.

The weak force accomplishes this in the most extreme way possible: only left-handed particles experience the weak force. Right-handed particles do not feel it at all. For reasons that we now explain, this is the key property of the weak force and one of the key properties of the Standard Model.

There are quite a few things that we will need to unpick regarding the weak force. Not least is the fact that, as stressed above, the distinction between left-handed and right-handed particles is only valid when the particles are massless. A remarkable and shocking consequence of parity violation is that, at the fundamental level, all elementary spin $\frac{1}{2}$ particles are indeed massless! The statement that elementary particles – like electrons, quarks and neutrinos – are fundamentally massless seems to be in sharp contradiction with what we know about these particles! We learn in school that electrons and quarks have mass. Indeed, in the introduction to these lecture notes we included a table with the masses of all elementary particles. How can this possibly be reconciled with the statement that they are, at heart, massless? Clearly we have a little work ahead of us to explain this! We'll do so in Section 4.2 where we introduce the Higgs boson.

4.1.2 A Weak Left-Hander

We're now in a position to explain how the three forces of the Standard Model act on the matter particles. The short-hand mathematical notation for the forces is

$$G = SU(3) \times SU(2) \times U(1)$$

Let's first recall some facts from the previous chapter. The strong force is associated to the "$SU(3)$" term in the equation above. As we explained in Chapter 3, the analog of the electric and magnetic fields for the strong force are called gluons, and are described by 3×3 matrices. (This is what the "3" in $SU(3)$ means.) Correspondingly, each quark carries an additional label, that we call *colour* that comes in one of three variants which we take to be red, green or blue.

While quarks come in three, colour-coded varieties, the leptons – i.e. the electron and neutrino – do not experience the strong force and hence they come in just a single, colourless variety. In the introduction, we said that each generation contains four particles: two quarks and two leptons. However, a better counting, including colour,

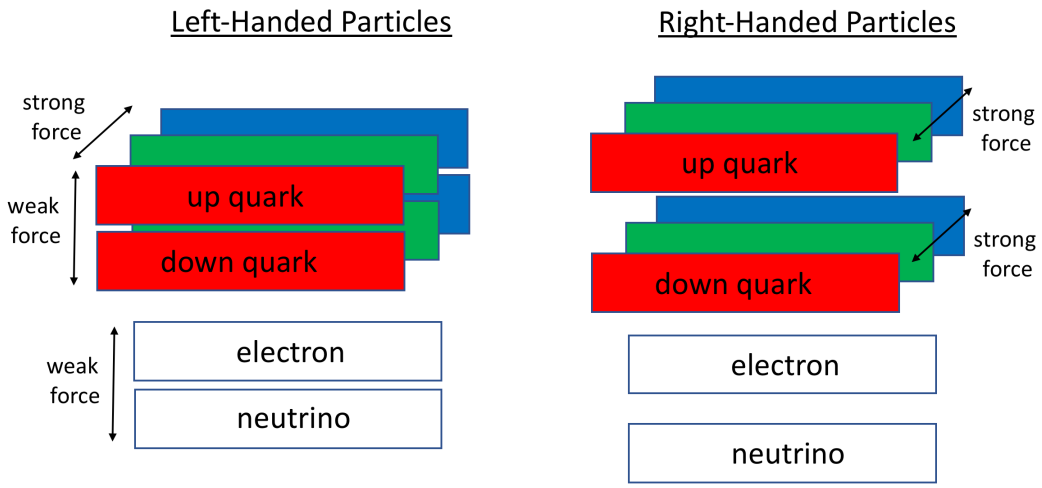


Figure 32. The jigsaw of the Standard Model. Only quarks experience the strong force. Only left-handed particles experience the weak.

shows that each generation really contains 8 particles (strictly 8 Dirac spinors). There are $3+3$ from the two quarks, and a further $1+1$ from the two leptons.

The next step in deconstructing the Standard Model is to note that each spin $\frac{1}{2}$ particle should really be decomposed into its left-handed and right-handed pieces. Only the left-handed pieces then experience the weak force $SU(2)$. If we were to follow the path of the strong force, you might think that we should introduce some new degree of freedom, analogous to colour, on which the weak force would act. A sort of weak colour. Like pastel. In fact, that's not necessary. The "weak colour" is already there in the particles we have.

This is illustrated in Figure 32. The right-hand particles are the collection of coloured quarks and colourless leptons. The left-handed particles are the same, except now the weak force acts between the up and down quark, and between the electron and neutrino. In other words, the names of distinct particles — up/down for quarks and electron/neutrino for leptons — are precisely the "weak colour" label we were looking for! We've denoted this in the figure by placing the weak doublets in closer proximity.

This should strike you as odd. For the strong force, the red, blue and green quarks all act in the same way. We say that there is a symmetry between them. However, it's very hard to make the same argument for the "weak colour" label. The electron and neutrino are very different beasts. If we're really introducing "weak colour" in analogy with actual colour, surely there should a symmetry between them. What's going on?

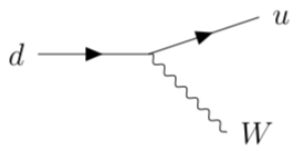
This brings us to our second striking fact: at the fundamental level, there is no distinction between a left-handed electron and left-handed neutrino! Nor is there a distinction between a left-handed up quark and right-handed up-quark. These particles share all their properties. However, as the story of the Standard Model plays out, the Higgs field intervenes. In addition to giving the elementary particles mass, the Higgs fields also leaves the electron/neutrino and up/down pairs with the distinctive characteristics that we observe. This aspect of the Standard Model is called *symmetry breaking* and will be described in Section 4.3.2.

Gauge Bosons

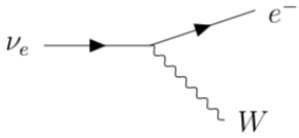
As with the other forces that we've met, there are spin 1 particles associated to the weak force. These are the analogs of the photon for electromagnetism and the gluon for the strong force. The spin 1 particles associated to the weak force are called *W-bosons* and *Z-bosons*. We'll see the difference between the W and Z later when we discuss the Higgs.

The type of spin 1 particles – like the photon, gluon, W and Z – that mediate forces are rather special and collectively go by the name of *gauge bosons*. (Here "gauge" is pronounced to rhyme with "wage".) The kind of Yang-Mills type theories that underly these are called *gauge theories*.

We can draw Feynman diagrams associated to the weak force. The W-boson interacts with the quarks, changing a down into an up like this

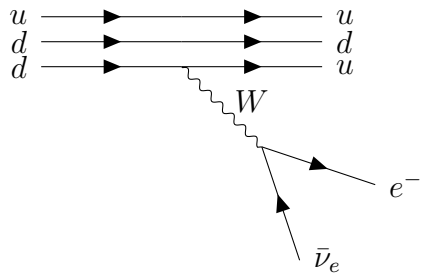


Note that the colour of the quark remains unchanged. For example, a red down-quark will turn into a red up-quark. Similarly, the W-boson interacts with the leptons like this.



In each case, the conservation of electric charge tells us that the W-boson must carry charge -1. We'll describe this in more detail later.

Already from these diagrams, we can start to understand how beta decay works. Recall that the quark content of the neutron is udd, while that of the proton is uud. One of the down quarks in a neutron decays into an up quark by emitting a W-boson. This is subsequently followed by the decay of the W-boson into an electron and anti-neutrino.



Note that we get an anti-neutrino rather than a neutrino because the arrow is now running backwards in time. We see that beta decay doesn't proceed through a direct interaction between quarks and leptons. It's mediated by the W-boson, in the same way that the electromagnetic force is mediated by the photon.

The fact that the down quark decays into an up quark, and not the other way round, can be traced to the fact that the down quark is (marginally) the heavier of the two. Moreover, the W-boson is heavier than the combination of the electron and neutrino (now the difference is not so marginal) so this decay too is allowed. These facts, like so many other things, come from the Higgs boson. By now, you should probably be getting the feeling that the Higgs plays an important part in all of this!

Hypercharge

Shortly we'll see how electric charge arises in this story. But first, we need to understand its counterpart at a more fundamental level. This is something called *hypercharge* and is associated to the $U(1)$ factor in the Standard Model. It is a force that is closely related to electromagnetism. However, the various particles don't have the same charges under hypercharge as they do electric charge. Importantly, hypercharge does not respect parity, and left-handed and right-handed particles carry different charges. These hypercharges are listed in Table 2, together with a summary of how the strong and weak forces act.

We've normalised the hypercharges in the conventional way so that they come in units of 1/6. However, this is just convention and there's nothing deep in this choice. We could just as well multiply everything by a factor of six, so all hypercharges are integer. Notice that one advantage of the current normalisation is that the hypercharge

Particles		Strong	Weak	Hypercharge
Left-handed	quarks	yes	yes	+1/6
	leptons	no	yes	-1/2
Right-handed	up quark	yes	no	+2/3
	down quark	yes	no	-1/3
	electron	no	no	-1
	neutrino	no	no	0

Table 2. The Standard Model forces acting on each of the fermions in a single generation.

for the right-handed particles coincide with their electromagnetic charge. This, as we shall see, is no coincidence. However, the hypercharges for the left-handed particles are rather different.

Before we move, I should mention that there is one caveat. (Isn't there always!) We don't yet have direct evidence for the existence of the right-handed neutrino and there is a possibility that it doesn't exist! Indeed, many people would say that the right-handed neutrino should not be included in the list of particles in the Standard Model. From the table you can see that the right-handed neutrino is neutral under all three of the forces in the Standard Model and this makes it very challenging to detect. We'll see the indirect evidence for its existence in Section 4.4 where we describe more about neutrinos in general.

4.1.3 A Perfect Jigsaw

The particles and forces listed in Table 2 summarise 150 years of work (dated from Röntgen's discovery of X-rays), dedicated to understanding the structure of matter at the most fundamental level. The first thing that comes to mind when you see it is: what a mess! The individual elements comprise some of the most gorgeous objects in theoretical physics – the Dirac, Maxwell and Yang-Mills equations. And yet any semblance of elegance would seem to have been jettisoned at the last, with the different components thrown together in this strange higgledy-piggledy fashion. Why this collection of forces and particles? In particular, why this strange collection of hypercharges?

Happily, there is an a wonderful and astonishing answer to these questions. The beautiful truth is simply: it could barely have been any other way.

The reason for this has its roots in Wu's observation that the world does not respect the symmetry of parity. As we explained above, we can account for this in the Standard Model by ensuring that left-handed fermions experience different forces from right-handed fermions. However, it turns out that this is not quite as straightforward to achieve as I've made out.

To explain why, I need to tell you a few further facts about the mathematics underlying quantum field theory. Quantum field theories in which left- and right-handed fermions experience different forces are called *chiral theories*. It turns out that chiral theories are particularly fragile objects, always teetering on the brink of mathematical inconsistency. Put bluntly, most chiral theories that you write down don't make any sense. If you write down a random collection of particles, with left- and right-handed components experiencing different forces, it is overwhelming likely that the equations will spit back stupid, and obviously wrong answers like "1=0".

If you want to write down a sensible chiral quantum field theory then there are a bunch of hoops that you have to jump through. These hoops are mathematical consistency conditions that the different forces must obey if the theory is to be sensible. One simple way of obeying these consistency condition is to ensure that left- and right-handed particles feel the same force but such theories don't exhibit parity violation. If you want to write down a theory with parity violation, you're obliged to work harder and find a delicate balance between the forces experienced by the left-handed particles and those experienced by the right-handed particles.

I won't describe all the consistency conditions, but here's a taster. Let's call the hypercharge of each fermion $\tilde{Q}_f$ where f labels the fermion. Then one consistency condition reads

$$\sum_{\rm left-handed} \tilde{Q}_f^3 = \sum_{\rm right-handed} \tilde{Q}_f^3 \tag{4.1}$$

To check that this works for one generation of the Standard Model particles listed in Table 2, you have to remember that each quark comes with three colours, while the left-handed fermions are really a pair under the weak force. We then have

$$3 \times 2 \times \left(\frac{1}{6}\right)^3 + 2 \times \left(-\frac{1}{2}\right)^3 = 3 \times \left(\frac{2}{3}\right)^3 + 3 \times \left(-\frac{1}{3}\right)^3 + (-1)^3 + 0^3$$

A successful solution to the mathematical consistency conditions, like the one above, is known technically as *quantum anomaly cancellation*. It's not a particularly enlightening name. For now, I can only tell you that underlying these conditions are some of the

deepest and most powerful ideas from the mathematics of topology. Indeed, the subject of "quantum anomalies" is currently where the fields of mathematics and physics have their richest intersection. (You can read more about quantum anomalies, in their full technical glory, in chapter 3 of the lecture notes on Gauge Theory.)

Viewed through the lens of quantum anomalies, the Standard Model morphs from a seeming mess into a perfect jigsaw. The forces act on the various particles so that the mathematical consistency conditions are satisfied. If you try to change any small piece, the whole thing falls apart and ceases to be a sensible physical theory.

I should stress that the Standard Model is not the only chiral gauge theory. If you allow for entirely different forces, or different collections of particles, then you can write down other such theories. But the Standard Model is, arguably, the simplest chiral gauge theory. Although, at first glance, the Standard Model looks like a jumble of random forces and particles, it is instead a beautiful and surprising theoretical construct. You just have to look at it the right way.

Hypercharge and Fermat's Last Theorem

It's difficult to explain the quantum anomaly consistency conditions at a deeper level without getting into the full mathematics. But here's a quick calculation that I'm particularly fond of that will hopefully give a sense of what's going on. Take the set of particles listed in Table 2, and assign them the properties under the strong and weak force that are listed. But allow them to have almost arbitrary values $\tilde{Q}_f$ of hypercharge. The "almost" is there because I'll impose two restrictions. First, we'll take the right-handed neutrinos to be neutral. (This can be motivated on the grounds that we're not really sure that they exist!) Second, we'll take the hypercharges $\tilde{Q}_f$ to be rational numbers, of the form p/q where p and q are integers. There are good theoretical reasons to think that charges should be quantised in this way. Then we ask the question: what values of hypercharges satisfy the mathematical consistency conditions?

It turns out that some of the hypercharges can be immediately related to others through the consistency conditions. And some remain arbitrary. If you follow through the calculation, and do some fairly complicated change of variables, the equation (4.1) ends up turning into the equation

$$X^3 + Y^3 = Z^3$$

where X, Y and Z must all be integers. This is a very famous equation! Fermat's last theorem tells us that the equation has no non-trivial solutions. There are, however, trivial solutions like $1^3 + 0^3 = 1^3$. If you take this trivial solution and plug it back into

the complicated change of variables, you'll discover the set of hypercharges listed in Table 2. These hypercharges may look random, but they're related to some very deep and beautiful mathematics. In particular, they're related to Fermat's last theorem!

Further Generations

As we explained in the introduction, the pattern of particles listed in Table 2 is repeated twice more. The second generation consists of the strange and charm quarks, together with the muon and muon-neutrino. The third generation consists of the bottom and top quarks, together with the tau and tau-neutrino.

We don't understand why there are three generations. However, the quantum consistency conditions tell us that each generation must come in a complete set. For example, there was a 20 year gap between the discovery of the tau lepton in 1975 and the discovery of the top quark in 1995. (The bottom quark was discovered in 1977). Yet no one doubted that the top quark was there because mathematical consistency required it. The whole theory doesn't make any sense without it the top quark.

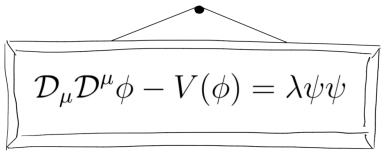
Similarly, if we one day discover a fourth electron-like object that is just a heavier version of the electron, muon and tau, then we know that there has to be a fourth neutrino and a pair of quarks waiting to be found. The four particles come in an irreducibly interwoven set.

4.2 The Higgs Field

Finally, it's time to introduce the famous Higgs boson. This is, it turns out, the simplest particle in the Standard Model. But the way in which it interacts with other fields is, by far, the most intricate. And, as we shall see, it is ultimately the Higgs boson that ties everything else together.

The Higgs boson is the simplest particle because it has spin 0. Indeed, it is the only fundamental particle that we know of without spin. Fields without spin are also referred to as *scalar fields*.

Recall that spin $\frac{1}{2}$ fields are described by the Dirac equation and spin 1 fields by the Maxwell or Yang-Mills equations. Both were pretty enough to be put in picture frames in earlier chapters. Any spin 0 field, like the Higgs boson, is described by the *Klein-Gordon* equation. It's nowhere near as beautiful as earlier equations, largely because it is too simple: it lacks the subtleties and surprises that make the Dirac and Yang-Mills equations so special. Nonetheless, it would be slightly churlish to deny it a place on the wall so, for what it's worth, here is the Klein-Gordon equation



In this equation, $V(\phi)$ is the Higgs potential while the terms on the right-hand-side describe the couplings to the fermions in the theory. Both of these will be described in more detail below.

As we've alluded to earlier in these lectures, the Higgs field plays a number of roles and we'll elaborate on these as we go along. For now we mention only that the Higgs field does not experience the strong force, but it does feel both the weak force and hypercharge. We should therefore augment Table 2 listing the forces experienced by each particle with one further entry:

Particle	Strong	Weak	Hypercharge
Higgs	no	yes	+1/2

Table 3. The forces experienced by the Higgs boson

Note that, because the Higgs field has no spin, it doesn't decompose further into left- and right-handed pieces. It just is.

Like all other fields, ripples of the Higgs field give rise to particles. This is the Higgs boson, the last of the Standard Model particles to be discovered. In weighs in at a mass

$$m_H \approx 125 \text{ GeV}$$

making it the second heaviest particle in the Standard Model, after the top quark.

However, the real importance of the Higgs lies not in the particle (although that's certainly interesting!) but, as we'll now explain, more in a property of the field itself

4.2.1 The Higgs Potential

Given that the Higgs field is simpler than all the others, why does it play such an important role? Well, there's something that a spin 0 field can do that higher spin fields cannot: they can "turn on" in the vacuum.

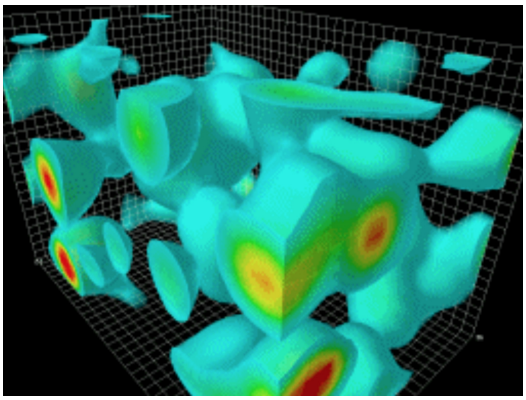


Figure 33. The fluctuations of a quantum field, taken from the simulations of Derek Leinweber.

To understand what this means, recall that in the introduction we gave some intuition for a quantum field as an object that's constantly fluctuating. More precisely, the vacuum of space should be viewed as a quantum superposition of many different field configurations. An example of a typical configuration of the gluon field in the vacuum is reproduced in Figure 33. Given this, it would seem that all fields "turn on" in the vacuum. However, importantly, for any field with spin the *average* of all these field fluctuations always vanishes in the vacuum.

The reason for this is simple: if a field with spin has a non-zero average, then it has to point in some direction. For example, if the average of the electric field $\mathbf{E}$ is non-vanishing then it picks out a direction in space. But the vacuum of space must look the same in every direction, so the average must be zero. (This statement is a little quick, but there are mathematical theorems that make it more precise.)

However, a scalar field is spinless, so doesn't point in any particular direction even when turned on. If we denote the field as $\phi(\mathbf{x}, t)$ then it's possible that, even in the vacuum, the average of all the fluctuations doesn't vanish, so

$$\langle \phi(\mathbf{x}, t) \rangle = \text{constant} \neq 0$$

where the angular brackets denote the average. In technical language, we say that the field *condenses*, a term that has its origin in the study of phase transitions and the condensation of water from vapour.

This new possibility brings up two immediate questions. What determines whether the field condenses? And what are the consequences if it does? Here we'll answer the first of these questions, postponing the second to Section 4.2.2.

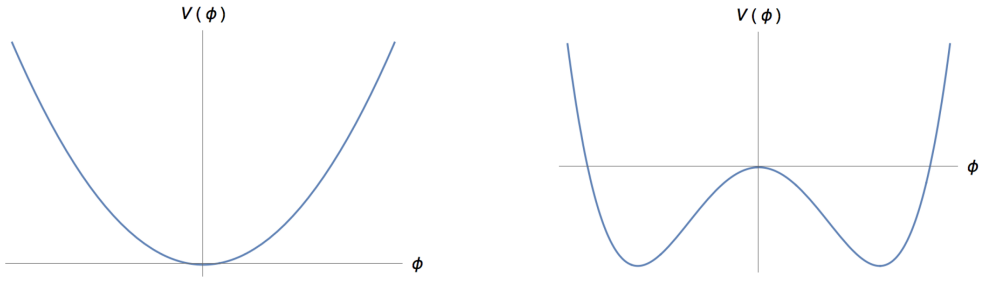


Figure 34. Two possible shapes for the potential for a scalar field. The Higgs field has a potential like that shown on the right, so that the field condenses, with $\langle\phi\rangle \neq 0$ in the vacuum.

The fate of a scalar field is not something that we get to chose. It is determined dynamically by the theory. Any scalar field experiences a potential energy that we call $V(\phi)$. This is a function that tells us how much energy it costs for the field to take certain values. Roughly speaking, there are two different shapes that these potentials take in theories of particle physics. These are shown in Figure 34.

In the vacuum, the scalar field sits at the minimum of the potential. If the potential has the shape shown on the left of Figure 34, then $\langle\phi\rangle = 0$ in the vacuum. However, if the potential has the shape shown on the right of Figure 34, then $\langle\phi\rangle \neq 0$ in the vacuum, and more interesting things happen. It turns out that the potential for the Higgs field in the Standard Model has the shape shown on the right, and this is what endows the Higgs with its power. (This statement is roughly true. A more accurate depiction of the Higgs potential will be given in Section 4.3.2.)

This, of course, brings up another interesting question: why does the Higgs potential in our world have the shape on the right, and not the shape on the left? We don't know the answer to that. At present, it is an input into the Standard Model and, hopefully, will be explained by some more complete theory in the future.

Here we are focussing on the question of whether the minimum of $V(\phi)$ sits at $\phi = 0$, or $\phi \neq 0$. But another question that we could ask is the value of $V(\phi)$ itself at the minimum. In the context of particle physics, this plays no role: it is just like any other potential energy, where only potential differences really matter and you can always add a constant to $V(\phi)$ without changing the physics. However, once we include gravity into the mix, the value of the potential energy becomes very important and contributes towards the cosmological constant. We'll say more about this in Section 5.3.1.

The Higgs Expectation Value

The upshot of the discussion above is that, even in the vacuum, the Higgs field averages to something non-vanishing. That something non-vanishing turns out to have the dimensions of energy. It is

$$\langle \phi \rangle \approx 246 \text{ GeV} \tag{4.2}$$

This is known as the Higgs *vacuum expectation value*. It is one of the key fundamental scales in the universe.

So what are the consequences of this? Well, they are pretty dramatic: the vacuum expectation value (4.2) turns out to give a mass to everything that it touches. This is known as the *Higgs mechanism* or the *Higgs effect*. The particles that get masses include both the W-bosons and Z-bosons that mediate the weak and hypercharge, together with all the spin 1/2 matter particles of the Standard Model. We'll postpone a discussion of fermion masses to Section 4.3, but the punchline is simply that the quarks, electrons and neutrinos get the masses that we advertised back in Section 1. Here, we will begin by focussing on the masses of the W-bosons and Z-bosons and some of their consequences. But first we will attempt, and largely fail, to give some intuition for why the Higgs field gives mass to particles at all.

Analogies for the Higgs Mechanism

The result of the Higgs expectation value $\langle \phi \rangle \neq 0$ is utterly startling. The Higgs field is like the ancient king Midas, but instead of turning everything to gold it makes everything massive. (Both make things heavier.) Why?

This is not an easy question to answer at the level of these lectures. What's perhaps unusual about this is that it's not at all difficult to understand the Higgs effect at the level of equations. Indeed, it's one of the simpler calculations in quantum field theory, but that doesn't change the fact that you do first need to learn quantum field theory. Largely, the difficulty in translating from equations to everyday language lies in the fact that the Higgs effect is really a phenomenon that is to do with fields rather than particles and we simply don't have much intuition for how these objects behave.

If we want to avoid the mathematics, we're obliged to rely on analogy. And, sadly, good analogies that relate the Higgs boson to more familiar, everyday phenomena are hard to come by. Here, for example, is a bad analogy. We could say that the Higgs field is like some kind of treacle. If you drag a spoon through treacle, you experience more resistance than if you drag it through water. Something similar happens with the

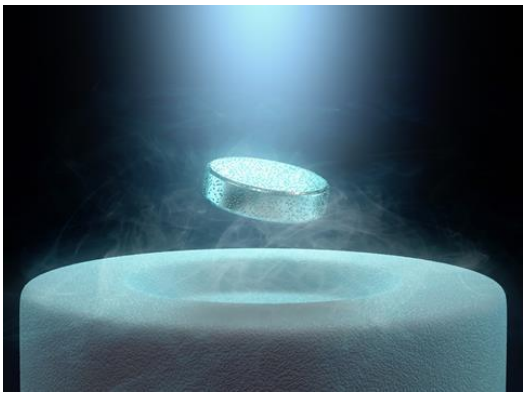


Figure 35. A superconductor and the Higgs boson are described by identical mathematics. (Image taken from the Royal Society of Chemistry).

elementary particles that interact with the Higgs. As we've seen, at the fundamental level all matter particles (as well as the W-boson and Z-boson) are actually massless and so should travel at the speed of light. But as they move in the vacuum, they have to plough through the Higgs field and this slows them down, effectively giving them a mass.

In fact, this isn't a totally terrible analogy. But if you try to push it any further it will quickly break down. For example, the reason that the spoon slows down is because of friction forces, which means that it continually loses energy to the treacle. In contrast, there's no friction force going on with the Higgs, nor any constant transfer of energy from the particle to the Higgs field. The analogy just stops working at this point.

There is, however, one completely excellent analogy for the Higgs effect, although not one that's particularly everyday. This can be found in phenomenon of superconductivity. In certain certain metals (for example, aluminum or lead) the electrical resistivity plummets to zero as they are cooled to low temperatures. At the same time, the metal expels any magnetic field, giving rise to dramatic technological applications like levitating trains.

The theory behind superconductivity is well understood. First, the electrons bind together into pairs. This, already, is somewhat surprising since the electrostatic Coulomb force repels two electrons and so it seems unlikely that they would want to bind together. However, in a metal the sound waves (also known as phonons) give a second, attractive force between electrons and in favourable conditions this can win over, causing the electrons to bind. While the individual electrons have spin, the bound state

can be spinless (rather like the mesons, formed of two quarks, that we met in Section 3.) It turns out that the potential for this bound state looks like the right-hand graph in Figure 34, and it condenses. The result is that, inside a superconductor, the photon gets a mass and this underlies behind all the subsequent phenomena, including the resistance free conduction and the expulsion of magnetic fields.

On a mathematical level, the analogy between the Higgs effect and superconductivity is exact: the key aspects of the equations describing the two are identical. It is a manifestation of the remarkable unity of physics, where the same ideas crop up in diverse situations.

4.2.2 W and Z Bosons

There are gauge bosons associated to all three forces in the Standard Model. Because the Higgs doesn't experience the strong force, the associated gauge bosons – which we named gluons in Section 3 – are unaffected by the Higgs boson. This is why we could discuss them earlier without dealing with all these subtleties.

In contrast, both the $SU(2)$ weak force and the $U(1)$ hypercharge interact with the Higgs interacts. Which means that they get a mass. The result is three, massive spin-1 bosons:

$$\begin{aligned} W^\pm \text{ bosons:} \quad & M_W \approx 80 \text{ GeV} \\ Z \text{ boson:} \quad & M_Z \approx 91 \text{ GeV} \end{aligned}$$

Here the W^+ and W^- particles are distinguished by their $\pm$ electric charge. They are the anti-particles of each other. In contrast, the Z-boson is neutral. It is its own anti-particle. Note that both masses are a factor of 3 or so below the Higgs expectation value (4.2). This is not a coincidence: the expectation value sets the scale of the masses, with the reduction due to the strength with which the Higgs field interacts with these spin 1 bosons.

The fact that the force-carrying particles become massive greatly changes the properties of the force. Instead of the familiar $V \sim 1/r$ Coulomb potential of electromagnetism (or the less familiar $V \sim r$ confining force of QCD), the massive W- and Z-bosons give a potential energy between particles that takes the form

$$V(r) \sim \frac{e^{-Mr}}{r} \tag{4.3}$$

where r is the distance between two particles, while M is the mass of the W or Z boson. We've met a force of this kind before: it takes the same form as the Yukawa

force (3.4) that is mediated by pions, and binds the protons and neutrons together in a nucleus. These forces have the characteristic feature that they become negligibly small for distances $r \gg 1/M$.

When, in Section 3, we discussed the force mediated by mesons, their mass was $m \approx 140$ MeV and the force extended over a distance $\approx 2 \times 10^{-15}$ m. Which, of course, is the size of the nucleus.

For the weak force, the masses of the force-carrying particles are almost 1000 times heavier than the meson, so the distance over which the weak force acts is almost a 1000 times smaller than the size of the nucleus. That's a pretty small distance scale! Indeed, it's so small that it means our old Newtonian way of thinking about forces as acting between particles really isn't particularly useful anymore. Unlike electromagnetism and the strong force, there are no examples where the weak force can be viewed of as an attractive force that make things stick together. There are no atoms or mesons of the weak force. Instead, as we've already seen, the primary role of the weak force is one of decay. Its job is to rent asunder that which the strong put together. This affects the neutron through beta decay and many other particles whose lives are cut tragically short by the weak force. We will describe this in more detail shortly.

The One That Got Away

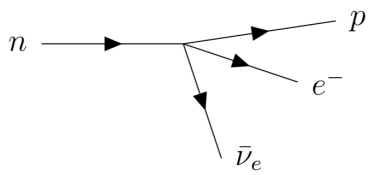
There's one final, very important twist to the story above. As we've seen, the gauge bosons associated to both the $SU(2)$ weak force and the $U(1)$ hypercharge interact with the Higgs. However, there's one special combination of these gauge bosons where the weak interaction is precisely cancelled by the hypercharge interaction. In other words, one combination of spin 1 bosons manages escape the attention of the Higgs and therefore remains massless. This is the photon.

This gives us a new perspective on the familiar electromagnetic force. Electromagnetism is not one of the three fundamental forces in the Standard Model, but instead is a combination of the weak force and hypercharge. It is special, because it is the only combination uncontaminated by the Higgs boson.

For this reason, the $SU(2) \times U(1)$ weak and hypercharge forces are referred to collectively as the *electroweak* theory. It is sometimes said that they are a unification of the weak and electromagnetic forces. While it's certainly true that the weak and electromagnetic forces are intricately interwoven, it's not quite correct to say that they're "unified". Indeed, it might be better to say that they're divided by the Higgs: the weak force became mired in the Higgs condensate, while the electromagnetic force is the one that got away.

4.2.3 Weak Decays

Enrico Fermi was the first person to understand beta decay. In 1934, just 18 months after the discovery of the neutron, he proposed a simple quantum field theory in which a neutron can decay into a proton, an electron and an anti-neutrino. The associated Feynman diagram is



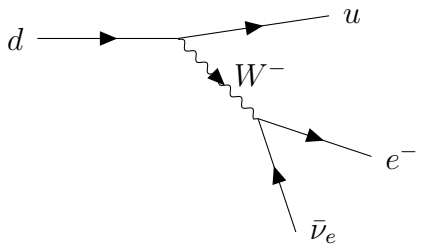
Thus, right from the beginning, the story of the weak force was one of decay.

Fermi's theory was one of the great breakthrough's of particle physics. Not only did it give a correct explanation for beta decay, but it was the first time that a quantum field theory was written down in which a particle of one type can transmute into something else. Any idea that the neutron was composed of a proton plus electron (or, as was also suggested, the proton was composed of a neutron + positron) were consigned to the waste bin.

It took several more decades to understand what things look like if we zoom in a little further. The structure of the electroweak theory in its essentially complete form was first understood by Steven Weinberg, but many others got close including Sheldon Glashow, Abdus Salam and John Ward.

W-bosons

We know that if we zoom into the neutron and proton we find quarks. Beta decay occurs when a down quark changes into an up quark. However, this process doesn't happen through a direct interaction of four fermions: it is mediated by the W-boson, and looks like this.



Let's unpack this a little. There are two vertices that involve the W-bosons. The relevant sub-diagrams look like this:

We've met both of these before, except this time we've specified that it's the W^- boson involved in the process. This is needed to ensure the conservation of electric charge. As usual, diagrams with W^+ can be formed by flipping the arrows or, equivalently, replacing particles with anti-particles.

There are a few things to say about beta decay when mediated by a W-boson. First, the reason that down quarks decay into up quarks, rather than the other way around, is because the mass of the down quarks is heavier than the masses of the decay products

$$m_d > m_u + m_e + m_{\nu_e}$$

We'll explain more about how these masses arise from the Higgs field in Section 4.3, but at this stage I'll simply point out that we don't currently have any understanding of why the masses are ordered this way rather than some other way. Nonetheless, this is crucial for beta decay to take place while conserving energy. The remaining energy $\Delta E = m_d c^2 - m_u c^2 - m_e c^2 - m_\nu c^2$ goes into the kinetic energy of the decay products.

If the masses had been ordered in some other way, then we would have a world in which, say, the electron could decay into a down and anti-up quark, together with a neutrino. Indeed, you can easily write down a Feynman diagram for such a process, but it's forbidden in our world because of energy conservation.

However, this begs a new question. The first sub-diagram in beta decay looks like a down quark decaying to an up quark and a W^- boson. But the W^- boson is way heavier than the up and down quark. Hence, energy conservation would suggest that such a decay is impossible. However, internal particles in diagrams – sometimes called *virtual particles* – are not subject to the same strict rules of energy conservation as external legs. (We described a similar idea in Section 2 when we first introduced Feynman diagrams.) Although beta decay proceeds through the creation of a W-boson, the existence of this W-boson is fleeting. Heuristically, we sometimes say that the W-boson briefly borrows some energy from the vacuum, and this is allowed by the Heisenberg uncertainty principle as long as it is paid back in a suitably short space of time. A better explanation is to admit that all is fields, and the kind of random fluctuations needed to create the W-boson are part and parcel of quantum field theory.

Whatever words we choose to drape around this, the fact that the intermediate W-boson is much heavier than any of the incoming or outgoing particles has consequence: it reduces the probability for the decay to take place and this in turn means that the lifetime of particles that decay through the weak force can be much larger than other timescales in particle physics. In fact, the lifetime of particles depends both on the mass of the W-boson and the mass differences between the initial and final particles.

So, for example, the half-life of a neutron is around 10 minutes, a very human number, while the half-life of uranium is around 4.5 billion years. And, of course, most nuclei do not decay at all, with the strong force providing a safe haven that stabilises the neutron.

That Time, Before We Were Born, When We Nearly All Died

As something of an aside, there's a wonderful race-against-the-clock story from the early universe that someone should turn into a Hollywood movie.

A long long time ago – a few fractions of a second after the big bang to be precise – protons and neutrons lived together in happy equilibrium. Beta decay happened, but so too did inverse beta decay and the two reactions were in a delicate balance. However, this state of harmony could not last forever. At around 2 seconds after the big bang, an imbalance kicked in and the inverse beta decay no longer occurred.

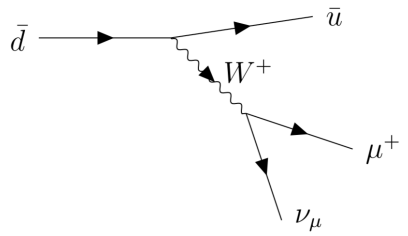
From then on, the neutrons were on their own. If any were to die, there was nothing to replace them. With a half-life of just 10 minutes, they needed to quickly find a sanctuary before they were all killed off by beta decay, a fate which would leave the later universe a very dull place to live.

The refuge was obvious: they should bind together with a proton to form a deuterium nucleus, or with a couple of protons to form a helium nucleus. Sadly the early universe back then was a hot and violent place, and every time the neutrons tried to bind together with a proton, the resulting nucleus was quickly smashed apart. For some time, the fate of neutrons – and all future life – hung in the balance. The neutrons needed to wait long enough for the universe to cool so that they could form nuclei, but time was not on their side. Eventually, around 6 minutes after the big bang, the temperature dropped sufficiently and the first stable deuterium nuclei formed, followed quickly by helium. The universe was saved, leaving a future that could be filled with atoms and stories.

The scary part about the above tale is that the 6 minutes needed for nuclei to form seems to have nothing to do with the 10 minutes needed for neutrons to decay. They come from entirely different pieces of physics. We should all feel very lucky to have survived this perilous time in history. You can learn more about the calculations underlying this in the lectures on Cosmology.

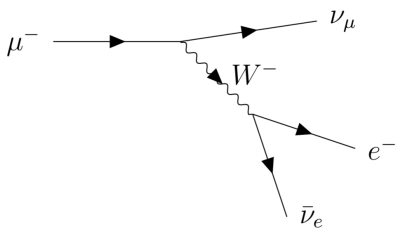
Pion Decay

Neutrons are not the only victim of the weak force. A world without the weak force would be awash with pions which, as we saw in Chapter 3, are the lightest of the particles formed from quarks. The vast majority of the time (something like 99.99%) charged pions $\pi^+ = u\bar{d}$ decay through the weak force to muons. This occurs through the Feynman diagram:



The resulting $\bar{u}$ anti-up quark then combines with the other up quark in the pion, and the two rapidly decay into photons. The lifetime of the charged pion is about 10^{-8} second.

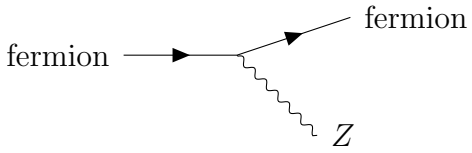
The resulting muons don't live too long either although, as we already saw in Interlude C, they hang around longer than the pions. Their demise is also due to the weak force and they decay to electrons and neutrinos through the process



The lifetime of the muon is around 2×10^{-6} seconds. All other particles involving quarks and leptons from the second and third generation have the same fate, decaying through the weak force to the more familiar particles from the first generation.

Z-Bosons

The Feynman diagrams involving Z-bosons are similar to those involving photons that we met in Section 2 in the sense that they don't change the type of fermion with which they interact:

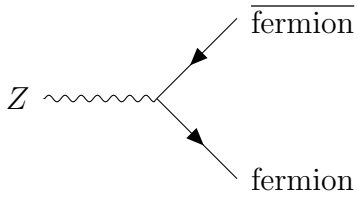


The fermion on the two legs is the same, but now can be either u, d, ν_e or e^-. Similar Feynman diagrams involving higher generations of fermions also exist for both W and Z interactions.

The effects of the Z−boson are less immediately dramatic than those of the W. There are a number of processes that have been key for us in unravelling the structure of the weak force. In particular, Z-interactions allow neutrinos to scatter off any other lepton or quark, and this is one of the key ways in which these particles are detected.

Second, there is an important story involving the lifetime of the Z-boson. This gives our best current understanding to the question: how many generations of fermions are there? Of course, we've discovered three. But are there more to be found?

We're not fully confident of the answer to this, but there is one very nice experiment which suggests that that we've found them all. This comes from looking at how the Z-boson decays. As we've seen, this can occur through a diagram of the form



where the fermion-anti-fermion pair can be either quarks or leptons. All that's needed is that the mass of the final two fermions is less than the 91 GeV mass of the Z.

About 20% of the time, the Z-boson decays to a pair of neutrinos. As we'll explain in more detail, the neutrinos are extremely light with a mass less than $\lesssim 1$ eV. If there were more generations of fermions they would, as we've seen in Section 4.1.3, necessarily include further neutrinos. But, assuming that these additional neutrinos are not widly heavier than the first three, they would affect the lifetime of the Z-boson and so could be detected indirectly.

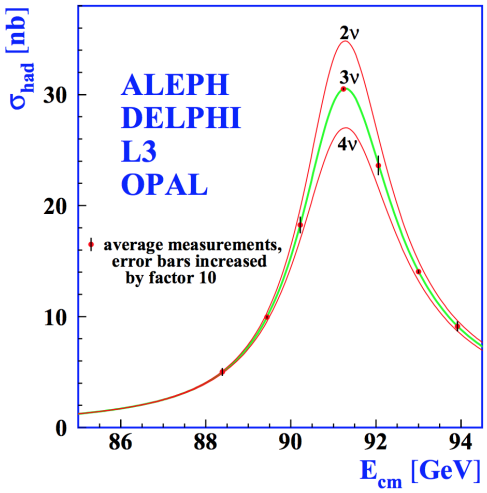


Figure 36. The Z-boson seen as a resonance in e^+e^- scattering. The three graphs show the shape of the graph that would be seen if there were 2, 3 or 4 species of neutrinos. The data points fall beautifully on the middle option. This plot was taken from a joint paper, combining many different experimental results.

A careful experimental study of the lifetime of the Z-boson then gives a striking result. The number N_ν of neutrinos that the Z decays into is

$$N_\nu = 2.994 \pm 0.008$$

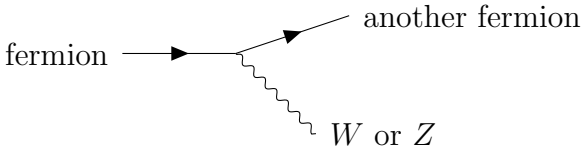
Which, if you're looking for an integer, is the same thing as 3. If there are further generations to be found, then the neutrinos must be so heavy that they don't affect the lifetime of the Z-boson. In other words, any further neutrino must be something like 10^9 times heavier than the three that we know and love.

Similar, but less striking, arguments about higher generations can also be made from a study of Higgs boson decay.

The Fermi Constant

For both QED and QCD we made a big deal out of the coupling constants and the way they change as you look at different energy scales. But, so far, we've made no comment about the strength of the coupling for the weak force.

There is such dimensionless coupling which we write as α_W. It's analogous to the fine structure constant α in QED. Whenever you see a Feynman diagram like the one below, it contributes to a process with probability proportional to α_W:



Rather surprisingly, the value of α_W is not particularly small. Like all coupling constants, it changes with scale. At the scale of 100 GeV, it takes the value

$$\alpha_W \approx \frac{1}{30}$$

This is somewhat larger than $\alpha \approx 1/137$ of electromagnetism! So why is the weak force so weak?

The answer, like everything to do with the weak force, lies in the Higgs mechanism. The weak coupling constant runs under renormalisation group and, like the strong force, actually increases as we look at lower energies. However, the Higgs ruins all of this. At the scale of the Higgs mechanism, the coupling freezes to roughly the value $\alpha_W \approx 1/30$. More importantly, the W and Z bosons get a mass at this scale and this greatly limits the range over which the weak force can operate to roughly distances $r \sim 1/M_W$, We saw this already in the effective Yukawa-type force (4.3) that arises from the exchange of massive W and Z bosons. Ultimately the reason that the weak force is so weak is because the distance over which it operates is so small, rather than the intrinsic weakness of the force itself.

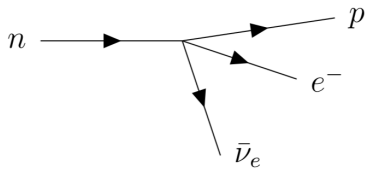
We can make this more quantitative. It's often more useful to characterise the strength of the weak force using a dimensionful coupling constant

$$G_F = \frac{\pi}{\sqrt{2}} \frac{\alpha_W}{M_W^2} \tag{4.4}$$

where the strange value of $\pi/\sqrt{2}$ is there for historical purposes. G_F is called the *Fermi constant* and takes the value

$$G_F \approx 1.166 \times 10^{-5} \text{ GeV}^{-2}$$

To understand the importance of this coupling, let's go back to our favourite beta decay. If we squint and ignore the W-boson, we return to Fermi's original theory where the process looks like a direct interaction between four fermions



The probability of such a decay is characterised by G_F which, as we see from the definition (4.4), incorporates both the original weak coupling constant α_W and the finite range of the W-boson due to its mass. More precisely, to get a probability we should construct a dimensionless object out of G_F so the probability of any low-energy event with $E^2 \ll G_F^{-1}$ is really $\sim G_F E^2$. Because G_F^{-1} is so big, any weak process that happens at low energy is suppressed. This is why the weak force is weak.

W and Z Self-Interactions

To complete our collection of Feynman diagrams, I'll finish by pointing out that there are interactions between the $W^\pm$ and Z bosons themselves. This follows because, just like for gluons, the underlying equations describing the weak force are the Yang-Mills equations, whose defining property is the presence of interaction terms between spin 1 particles. These Feynman diagrams take the form

where each leg can be any $W^\pm$ or Z boson, as long as charge conservation is obeyed at the vertex. These diagrams are mostly important in higher order corrections to the kinds of processes that I've described above.

4.3 Flavours of Fermions

In this section, we will describe how the quarks and leptons interact with the Higgs field. The six quarks are often referred to as *flavours*. As we will see, the structure of flavour is, in many ways, the most fiddly and poorly understood part of the Standard Model. It is certainly where the vast majority of the parameters in the Standard Model reside and, most likely, the best place to look for clues about what lies beyond.

4.3.1 Yukawa Interactions

The Higgs field talks to quarks and leptons through *Yukawa interactions*.

There is room for confusion in this name. The original interactions postulated by Yukawa were designed to explain how neutrons and protons bind together inside a nucleus, with a scalar field providing the mediating force. We now known that the scalar particle Yukawa had in mind is the pion and, as explained in Section 3, is composed of two quarks.

However, Yukawa's basic idea – that a scalar field can interact with two fermions – is reprised in the Standard Model with the Higgs field arising as the fundamental scalar, and the name "Yukawa interactions" has been co-opted in this new context. More precisely, the Yukawa interactions in the Standard Model couple the Higgs field to one left-handed fermion and one right-handed fermion.

Like all forces, there are dimensionless numbers that tell us the strength of the interaction between the Higgs field and the fermions. These dimensionless numbers are called *Yukawa couplings* and we will denote them by λ. The Higgs field then gives a mass to each fermion that is directly proportional to the strength of the Yukawa coupling,

$$\text{mass} = \frac{\lambda}{\sqrt{2}} \times 246 \text{ GeV} \tag{4.5}$$

where 246 GeV is the value of the Higgs expectation value (4.2) that we met earlier. The factor of $1/\sqrt{2}$ in this formula is just convention.

Each fermion thas a different Yukawa coupling, and hence a different mass. The Yukawa couplings for the various quarks are:

$$\begin{aligned}
\text{top}: &\quad \lambda \approx 1 &\Rightarrow\quad & m_t \approx 173 \text{ GeV} \\
\text{bottom}: &\quad \lambda \approx 2.5 \times 10^{-2} &\Rightarrow\quad & m_b \approx 4.2 \text{ GeV} \\
\text{charm}: &\quad \lambda \approx 7.5 \times 10^{-3} &\Rightarrow\quad & m_c \approx 1.3 \text{ GeV} \\
\text{strange}: &\quad \lambda \approx 5.5 \times 10^{-4} &\Rightarrow\quad & m_s \approx 96 \text{ MeV} \\
\text{up}: &\quad \lambda \approx 1.3 \times 10^{-5} &\Rightarrow\quad & m_u \approx 2.2 \text{ MeV} \\
\text{down}: &\quad \lambda \approx 2.7 \times 10^{-5} &\Rightarrow\quad & m_d \approx 4.7 \text{ MeV}
\end{aligned}$$

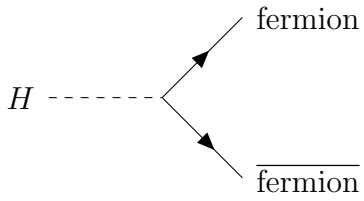
Meanwhile, the Yukawa couplings for the electron and its cousins are

$$\begin{aligned}
\text{tau}: &\quad \lambda \approx 1 \times 10^{-2} &\Rightarrow\quad & m_\tau \approx 1.8 \text{ GeV} \\
\text{muon}: &\quad \lambda \approx 6.1 \times 10^{-4} &\Rightarrow\quad & m_\mu \approx 106 \text{ MeV} \\
\text{electron}: &\quad \lambda \approx 2.9 \times 10^{-6} &\Rightarrow\quad & m_e \approx 0.5 \text{ MeV}
\end{aligned}$$

Although we've reduced the masses of the various quarks and leptons to dimensionless coupling constants λ, we currently have no understanding of why the Yukawa couplings take these values. The Yukawa couplings span 5 orders of magnitude in order to explain the quark masses, and a further order of magnitude to explain the electron mass. We don't know why. Is it coincidence that the top Yukawa coupling is almost exactly one? Again, we simply don't know. All of these are, from the perspective of the Standard Model, fundamental constants and we have no deeper understanding of them.

You may have noticed that we haven't yet discussed neutrinos. These have masses at least six orders of magnitude smaller than the electron, or 10^{-12} times smaller than the top quark. Thankfully there is a plausible reason for this vast discrepancy in mass, but we postpone the discussion until Section 4.4.

The fact that the masses of all fermions derive from the Higgs field has a consequence that could only be tested after the Higgs boson was discovered. The Higgs boson couples to all fermions with the Feynman diagram interaction



The strength of this interaction is dictated by the same Yukawa coupling λ that determines the mass of the fermion. This means that once the Higgs boson is produced, we have an entirely different way of measuring the Yukawa couplings by determining the relative probability that the Higgs boson decays to various fermion-anti-fermion pairs. So far, decays to the top and bottom quarks and the tau have been measured, all in agreement with theoretical expectations.

4.3.2 Symmetry Breaking

Symmetry is one of the key concepts in particle physics. For example, the $SU(3)$, $SU(2)$ and $U(1)$ labels that we've been using to describe the three forces are really mathematical expressions of symmetry. Until now, we've somewhat underplayed the idea of symmetry in these lectures, largely because it's a fairly formal mathematical idea and analogies tend to get bogged down in useless diversions. However, in describing how fermions get a mass we have a chance to elaborate on this. At the same time, we'll also get a better understanding of how electromagnetism emerges from the electroweak force.

For an example of symmetry, we can first look to the strong force. As we've seen, each quark comes with a colour: red, green or blue. But there is a symmetry underling the choice of colours. To see this, you take a collection of particles and swap all the colours around,

$$\text{red} \longrightarrow \text{green} \longrightarrow \text{blue} \longrightarrow \text{red}$$

The end result will then look identical to the set-up we started with. This happens for the proton because it has a collection of quarks with one colour of each and that

doesn't change under a permutation of colours. Meanwhile, as we described in Section 3.2, a meson has a quantum superposition of colours $\bar{r}r + \bar{g}g + \bar{b}b$, so that too remains unchanged.

What's the analogous statement for the weak force? We've seen that "weak colour" is the up/down and electron/neutrino label for left-handed particles. This means that, at the fundamental level there is a symmetry that swaps

$$\text{up} \longleftrightarrow \text{down}$$

and

$$\text{electron} \longleftrightarrow \text{neutrino}$$

for left-handed particles. Under such an exchange, all physics should remain the same.

Now, although there is such a symmetry at the fundamental level, it certainly doesn't manifest itself in our world. If you were to exchange all the electrons in your body for neutrinos, bad things would happen. So what's going on? What happened to this fundamental symmetry of nature?

The answer is that the symmetry is broken by the Higgs boson. To understand this, we first note that the Higgs experiences the $SU(2)$ weak force. This means that the Higgs too must have a "weak colour" label, something that we've neglected to mention so far. The Higgs field is actually a pair of complex-valued Higgs fields

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$$

where the two fields ϕ_1 and ϕ_2 are mixed by the weak force.

Because there are actually two Higgs fields, the potential $V(\phi)$ should really be plotted in 3d. We have drawn this in Figure 37. Note that the shape of the potential doesn't change if you rotate it around the vertical axis, and this reflects the symmetry of the weak force. There's no sense in which the potential prefers the ϕ_1 direction to the ϕ_2 direction: both are on the same footing. If you slice this potential along any direction, then you get the 1d graph that we previously drew on the right-hand side of Figure 34.

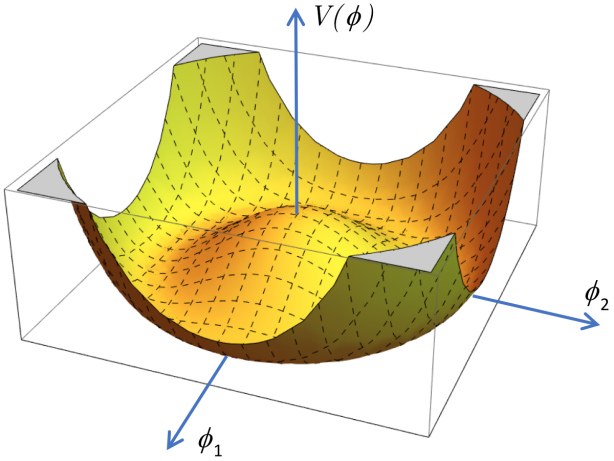


Figure 37. A better depiction of the Higgs potential. This is sometimes called the "mexican hat" potential for its sombrero-like quality.

This new 3d potential doesn't have isolated minima. Instead, there is a ring of minima, all lying at some fixed distance $\sqrt{\phi_1^2 + \phi_2^2}$ from the origin. In the vacuum, the Higgs field must sit somewhere in this valley. But where? Nothing tells the Higgs field where to sit and, moreover, because of the symmetry, any choice is as good as any other. Nonetheless, like a ball on a roulette wheel, the Higgs must choose somewhere to sit. And choose it does. Because it doesn't matter where the Higgs field sits, we may as well decide that it lies along the ϕ_1 axis, although any other position would be just as good. This is shown in the figure to the right.

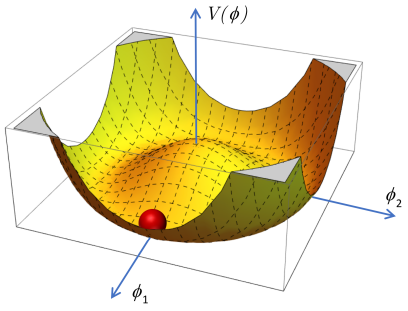


Figure 38. The choice of Higgs expectation value breaks the symmetry

Once the Higgs field has nestled in place, it's no longer true that the ϕ_1 direction is the same as the ϕ_2 direction, because the Higgs field sits in one of these directions and not the other. We say that the symmetry has been broken. More precisely, we say that the symmetry has been *spontaneously broken*.

Ultimately, this is the reason why electrons and neutrinos (and up and down quarks) behave so very differently in our world. The laws of physics endow the left-handed versions of the particles with identical properties. But then the Higgs field comes along and spoils it, choosing to sit in one place and rather than another. The ϕ_1 direction

in which the Higgs field sits coincides with the "electron direction" for leptons, and the "down quark" direction for quarks. The ϕ_2 direction is where the neutrino and up quark live.

Finally, we can complete the story of how electromagnetism emerges from the electroweak force. The photon is the combination of the weak and hypercharge forces that does not fall into the clutches of the Higgs boson. The right-handed particles know nothing about the weak force, so for them the coupling to the photon (which we call electric charge) is identical to the coupling to hypercharge. Indeed, if you look at Table 2 you'll see that the hypercharge for the right-handed particles is the same as their electric charge.

However, the left-handed particles feel both the weak and hypercharge and the resulting electric charge is a combination of both. How that combination plays out depends on whether the particle is aligned with the Higgs field, or orthogonal to the Higgs field. For those left-handed particles that are aligned with the Higgs field, the resulting electric charge is

$$\text{electric charge} = \text{hypercharge} - \frac{1}{2}$$

while for those that lie orthogonal to the Higgs field, the electric charge is

$$\text{electric charge} = \text{hypercharge} + \frac{1}{2}$$

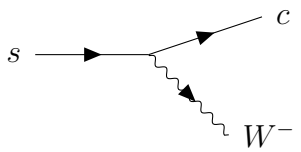
We can now check what this means for the left-handed quarks and leptons by looking back at Table 2. After symmetry breaking, the left-handed quark splits into two particles with electric charges $\frac{1}{6} \mp \frac{1}{2} = -\frac{1}{3}$ and $+\frac{2}{3}$. These, of course, are the charges of the down and up quark respectively. Meanwhile, the left-handed leptons also split into two, now with charges $-\frac{1}{2} \mp \frac{1}{2} = -1$ and 0. These are the electric charges of the electron and neutrino.

4.3.3 Quark Mixing

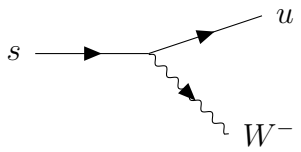
We've now described all the forces of the Standard Model. However, within the structures that we've outlined above, there are certain particle interactions that we cannot explain. We can't, for example, explain the decay of hadrons that contain quarks from the higher generations.

To see the problem, consider the kaon K^-, whose quark content is $\bar{u}s$. We stated in Section 3.4 that kaons decay with a lifetime of around 10^{-8} seconds. But where do they go? Within the strong and electromagnetic forces, there's no way for quarks to

change their type, so it must be a weak interaction that does the job. But the weak interaction that we've described above doesn't allow quarks to change generation. For a strange quark, we have the Feynman diagram



But then we're left with a charmed quark and there's nowhere for that to go. Instead, for a kaon to decay we would need an interaction that mixes the generations, like this:



If such a decay was possible then the resulting up quark could annihilate with the $\bar{u}$ in the kaon, while the W^- can decay into, say, an electron and anti-neutrino in the usual way.

So what are we missing? Is it possible for the weak force to mediate decays that change quarks from one generation into another? The answer to the second question is: yes. The answer to the first question is that we're missing something rather subtle about the meaning of the word "particle"!

So far, we have been using two, somewhat different meanings of the word "particle" and tacitly assuming that they coincide. These are:

- A particle is an excitation of the field that has a fixed energy. Or, because $E = mc^2$, an equivalent way of saying this is that we can assign a specific mass to the particle. In the language of quantum mechanics, we say that it is an *energy eigenstate*.

- A particle is the object that interacts with a particular force. This is really pertinent only for the weak force which, as we've seen, turns one particle into a different particle: say the down quark into an up quark.

The subtlety comes about because, for the weak force, these two ideas of what it means to be a particle don't quite agree. The excitations of the field with a fixed energy aren't the same thing as the excitations of the field that have a specific interaction with the weak force. Another way of saying this is to recall that the mass of the particle comes

from the Higgs field. So what's really going on here is a mismatch between the way the Higgs interacts with fermions and the way the W-bosons interact with fermions. The two interactions are not quite aligned.

Let's keep the our original names for quarks – u, d, s, etc – for the particles that have a definite mass. We'll then denote the particles that experience the weak force with primes – u', d', s' and so on. It turns out that we can always simply define the up-sector quarks to be aligned,

$$u = u' \quad \text{and} \quad c = c' \quad \text{and} \quad t = t' \tag{4.6}$$

But the down-sector quarks are then misaligned. It's simplest to explain what's going on if we first ignore the bottom quark. The misalignment is then given by

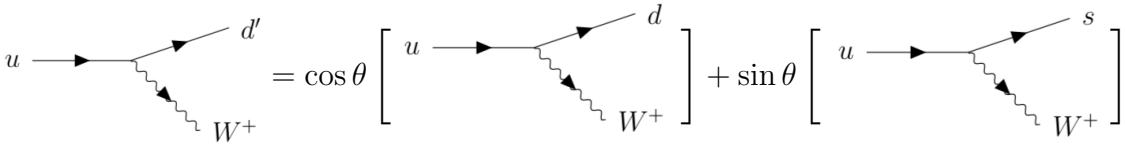
$$\begin{aligned} d' &= d\cos\theta + s\sin\theta \\ s' &= s\cos\theta - d\sin\theta \end{aligned} \tag{4.7}$$

Here θ is known as the *Cabbibo angle*. It is a fundamental parameter of Nature, an example of what's called a mixing angle. We'll see many more of these shortly. The Cabbibo angle is measured experimentally to be

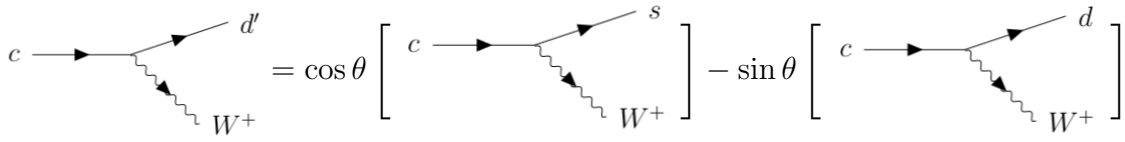
$$\sin\theta \approx 0.22 \quad \Rightarrow \quad \theta_c \approx \frac{\pi}{14} \approx 13^\circ$$

Why this number and no other? We don't know! We don't currently have any deeper explanation for this.

The formulae (4.7) might remind you of the equations for a rotations, and that's exactly the right way to think about it. The (d', s') quarks that feel the weak force are rotated relative to the (d, s) quarks that interact with the Higgs. This phenomenon is called *quark mixing*. It means that the Feynman diagrams that we previously wrote down for the weak force should be amended. The correct Feynman diagram involving the up quark is

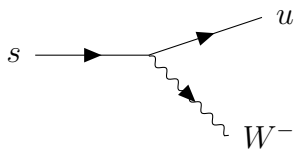


and similarly for the charm quark



How should we think about this? It seems to say that if, for example, a charm quark decays by emitting a W-boson, then the end product is both a strange quark and a down quark, in some combination. But, of course, we're in a quantum world here. And just as a particle can be in two places at the same time, or a cat both dead and alive, the decay product of a charm is indeed a quantum superposition of a down quark and a strange. As usual, this manifests itself in our experiments as probability. We get probabilities by taking the square of a Feynman diagram: so the probability of $c \to s + W^+$ is proportional to $\cos^2\theta$ while the probability of $c \to d + W^+$ is proportional to $\sin^2\theta$.

The phenomenon of quark mixing resolves our earlier puzzle: it's now quite possible for a meson like the kaon to decay, because there is an escape route for the strange quark, with Feynman diagrams like this now allowed:



The only price we pay is that the probability for such events to happen is reduced by $\sin^2\theta \approx 0.05$. This results in an increased lifetime for mesons containing strange quarks.

This story repeats with the addition of an extra generation. Now there is mixing between the down, strange and bottom quarks, and the simple rotation (4.7) is replaced by a more complicated matrix equation

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{4.8}$$

The 3×3 matrix is known as the Cabbibo-Kobayashi-Maskawa (or CKM) matrix. The upper-left 2×2 sub-matrix agrees with the Cabbibo mixing (4.7), but now we have the possibility of mixing between all three quarks.

You might reasonably ask: what made the up-sector special? Why is the up-sector aligned, as in (4.6), while the down-sector has the complicated CKM matrix? The answer is that this is just a choice. There's some freedom in the equations to guarantee a partial alignment between the weak and Higgs forces. The convention is to pick the up-sector to be aligned because then the misalignment looks simple for the relatively

light strange mesons, with no need to invoke the charm quark in the argument. But if you were feeling a little perverse, there's nothing to stop you redefining everything with the down-sector aligned and the up-sector askew, or even some combination of the two.

The components of the CKM matrix have been accurately measured experimentally. It turns out that some of the elements can be complex numbers and we'll explain the significance of this in Section 4.3.4. For now, we give just the absolute values of each element which are roughly

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \approx \begin{pmatrix} 0.97 & 0.22 & 0.004 \\ 0.22 & 0.97 & 0.04 \\ 0.009 & 0.04 & 0.999 \end{pmatrix}$$

You can see the Cabbibo angle sitting there in $V_{us} = \sin\theta \approx 0.22$. The full CKM matrix extends the Cabbibo angle to 10 parameters – the 9 above, together with a complex phase that we'll discuss in Section 4.3.4.

Just like we have no understanding of why the Cabbibo angle takes its particular value, nor do we have any good understanding of the CKM matrix. As you can see, it's not far from a diagonal matrix, with the Cabbibo terms being the only ones that aren't tiny. We don't know why.

It's worth pausing to take in a bigger perspective here. In the first part of this chapter, we described how the matter content of the Standard Model interacts with the different forces. There we found a beautiful consistent picture – a perfect jigsaw – in which the interactions were largely forced upon us by the consistency requirements of the theory. For a theoretical physicist, it is really the dream scenario. This, however, contrasts starkly with the story of flavour. Even focussing solely on the quarks, we find that there are 6 Yukawa couplings that determine their mass, plus a further 10 entries of the CKM matrix that determine their mixing. And none of these parameters are fixed, or understood at a deeper level.

Somewhat ironically, much of this complexity can be traced to the simplicity of the Higgs. The strong and electroweak forces are described by Yang-Mills type theories, and these come with mathematical subtleties that are ultimately responsible for the quantum consistency conditions that constrain their interactions. But there are no such subtle constraints for the Higgs boson. It is a simple, spin 0 particle, that can do as it pleases and the result is the plethora of extra parameters that we've seen.

Turning this on its head, there is a possibility that the flavour sector of the Standard Model may well offer a unique opportunity. The structure of quark masses, together with the CKM matrix, surely contains clues for what lies beyond the Standard Model. Why the hierarchy of masses? Why is these values of the CKM matrix? Moreover, there is a peculiar pattern: with $V_{us} = \sin\theta$, the other remaining values of the CKM matrix are not far from $V_{cb} \approx \sin^2\theta$ and $V_{ub} \approx \sin^3\theta$. Is this coincidence, or telling us something deep and important? We don't know. But hopefully, one day, we will find out.

4.3.4 CP Violation and Time Reversal

There is one last surprise waiting for us in the CKM matrix. We started this chapter by describing how the symmetry of parity is not respected in the microscopic world: the laws of physics look different when reflected in the mirror. As we saw, this was a pivotal moment in the history of particle physics, giving an important clue about the chiral structure of the Standard Model.

The violation of parity goes hand in hand with the violation of another symmetry, know as *charge conjugation*. We could ask: are the laws of physics the same for particles and anti-particles? And the answer is no. Anti-particles do not feel the same forces as particles. Moreover, this follows immediately from parity violation: the weak interaction involves left-handed neutrinos, or right-handed anti-neutrinos. So if you just change a particle to an anti-particle – say a left-handed neutrino to a left-handed anti-neutrino – then it doesn't experience the weak force anymore.

However, it remains a logical, and indeed compelling, possibility that the laws of physics are invariant under the simultaneous action of parity and charge conjugation. This symmetry is called CP, the C for "charge conjugation" and the P for "parity". A universe that exhibited CP symmetry would have the property that the laws of physics look the same if we swap all particles with anti-particles *and* look at them reflected in the mirror.

So are the laws of physics actually invariant under CP? You may not be surprised to hear that the answer is no.

CP violation was discovered in 1964. We describe the experimental evidence in Interlude D.2, focussing here on the theory. While violation of parity has an enormous effect on the theoretical underpinnings of the Standard Model, the violation of CP appears to be almost an afterthought: it turns out that the laws of physics violate CP if the CKM matrix contains a complex number! It does, and so CP is violated.

There is a way to quantify how much CP is violated in the Standard Model. From the CKM matrix, one can define a single number called the *Jarlskog invariant*, J. The maximum theoretical value it can take is $J \leq 1/2\sqrt{3} \approx 0.3$. Its experimentally measured value is $J \approx 3 \times 10^{-5}$. We see that the CP violation, at least as manifested among quarks, is really very small.

The experimental consequences of CP violation among quarks are, to put it mildly, rather subtle. It shows up in the decay of certain neutral mesons involving higher generations (like the kaon $K^0 = d\bar{s}$). We'll explain this in more detail in Section D.2.

Why should we care about CP violation? Well, there are two reasons. The first is that our universe is, rather fortunately, full of matter but with very little anti-matter. It's thought that this imbalance occurred naturally in the early universe, but for this to happen there have to be processes where matter and anti-matter behave differently. This, it turns out, requires CP violation. So although small, it may well have had extraordinarily large consequences. We'll discuss this further in Section 5.3.3.

There's also a twist to this tale. It turns out that a world with just two generations of fermions has no room for CP violation in the CKM matrix! In that case, the only quark mixing comes from the Cabbibo angle (4.7) and there's no complex number in sight. So, while it may seem like the bottom and top quark aren't much good for anything today, without them we may well find ourselves in a universe with as much matter as anti-matter.

Time Reversal

The second reason that CP symmetry is important is because it's closely related to yet another symmetry, that of *time reversal*.

A theory is said to obey the symmetry of time reversal if there's no way to distinguish between the laws of physics running forwards in time and running backwards. Of course, in our macroscopic everyday world, the arrow of time is obvious. Underlying this is the concept of *entropy* which characterises the tendency for things to become disordered and muddled for the simple reason that there's more ways to be messy than to be neat. Ultimately, the entropic arrow of time can be traced to initial conditions. For a reason that we don't fully understand, the universe started off neat and tidy, so that it could ultimately descend into disorder, creating interesting structures like galaxies, planets and life along the way.

This entropic arrow of time is a powerful idea, largely because it doesn't depend on what's happening on the microscopic scale. In particular, it cares little for whether the laws of physics at the fundamental level are time reversal invariant. Nonetheless, it's interesting to ask whether time reversal is a symmetry of the world. If we were able to take a movie of interactions of fundamental particles, then play it backwards, would we notice the difference?

It turns out that the answer to this question is very closely tied to CP violation. The three symmetries of parity, charge conjugation and time reversal are an intwined triumvirate. There is a theorem that states that all laws of physics must be invariant under the combination of all three symmetries, what is known as CPT, with T standing for "time reversal". This means that if you make a movie of some dance performed by fundamental particles, then it's guaranteed to look the same if you watch it played backwards, in a mirror, with all particles exchanged for anti-particles.

The CPT theorem tells us that if CP is violated, then so too is T, for only then can the combination CPT be saved. The upshot is that the experimental discovery of CP violation is, admittedly indirectly, telling us that the laws of physics must look different running forwards and backwards in time. In other words, even at the microscopic level, there is an arrow of time in the universe.

Now, that sounds like a very big deal. Nonetheless, it's striking how little impact this fact has, not just on our daily lives, but on our deeper understanding of physics. As we mentioned above, it is likely that CP violation or, equivalently, time reversal violation is responsible for the preponderance of matter over anti-matter in the universe and that's not something to sneeze at. But that happened a long time ago! Has the violation of time reversal really not had any significant consequences since?! Moreover, it's not even straightforward to give a clear description of the microscopic process that runs differently forwards and backwards in time. The experimental observation of CP violation are described in Section D.2, and involve some rather subtle aspects of meson decay. We'll do slightly better in Section 4.4 when we discuss possible CP violating effects in neutrinos where an interpretation in terms of time reversal is easier to come by.

However, on the theoretical level it's interesting to compare the effects of parity violation with the effects of time reversal violation. At first glance, they seem very similar: one is a flip of spatial coordinates, $\mathbf{x} \to -\mathbf{x}$, the other a flip of time $t \to -t$. Yet the discovery of parity violation had profound consequences, leading immediately to the need for chiral matter, the associated delicate consistency conditions between

left- and right-handed fermions that, ultimately, gives us much of the structure of the Standard Model. This stands in sharp contrast to the theoretical consequences of time reversal violation which imply only that a single parameter, buried within the CKM matrix, is a complex rather than a real. It makes you wonder if there's something that we're missing!

4.3.5 Conservation Laws

Some things never change. In physics, we call these conservation laws. They comprise some of the most useful and powerful laws of nature. Among the conservation laws that we learn early in our physics education are the conservation of energy, momentum, angular momentum, and electric charge.

The familiar conservation laws listed above are all exact. No known process violates these laws. Moreover, they are sewn in the mathematical fibre of quantum field theory at such a deep level that it seems likely that they are here to stay.

In addition to these exact conservation laws, we have a number of "almost conservation" laws, together with a couple of "probably almost conservation" laws. These are the subject of this section.

First, the "almost conservation" laws. These are things that are conserved if you ignore one kind of force or another, but ultimately do not hold when you consider the whole of the Standard Model. For example, the electromagnetic and strong interactions don't change the type of quark. This means that the number of up quarks, down quarks and strange quarks are each individually conserved by any electromagnetic or strong process. Of course, as soon as the weak interaction kicks in, the down quark can decay into an up quark as we have seen repeatedly in this section. But because the weak force is, as the name suggests, weak, these decays can take a long time. This means that there will be situations where we can ignore weak decays and view the number of up and down quarks – or, equivalently, the number of protons and neutrons – as an effectively conserved quantity. This almost conservation law is sometimes called *isospin* and was mentioned briefly in Section 3.

As we saw in the last section, a strange quark can decay only through the weak force and, even then, only through the process of quark mixing. This causes yet further suppression in decays of mesons containing strange quarks, and this is seen in the relatively long lifetimes of the kaons where now "relatively long" means around 10^{-8} seconds. Again, we can think of this as an "almost conservation" of a quantity called *strangeness*. Combining isospin with strangeness gives the eightfold way that we described in Section 3.

There is a similar story in the lepton sector. If we ignore the weak force, the number of electrons, muons, taus and their associated neutrinos are all unchanging. When we include the weak force, the muon and tau can both decay. Moreover, as we will see in the next section, the neutrinos also undergo a mixing process, analogous to the one we've seen for quarks, so the individual neutrino species are not, ultimately, conserved quantities either.

Nonetheless, when the dust settles there do seem to be two exact conservation laws that follow from the Standard Model as described so far.

- **Conservation of quark number**

  If you start with one kind of quark, it can decay into a different kind, usually ending up as the up quark since this is the lightest. But, if we count quarks as $+1$ and anti-quarks as -1, then the total number of quarks can't change.

  Confinement means that we don't see individual quarks, but rather the protons, neutrons and mesons that they bind into. The mesons contain a quark and anti-quark, and so there's no such thing as meson conservation. But the conservation of quark number means that the number of baryons (again, counting anti-baryons as -1) can't change. For this reason, "conservation of quark number" is usually referred to as *baryon conservation*.

  We see a good example of this in beta decay. The neutron decays, but leaves behind a proton. The total number of baryons before and after is the same.

- **Conservation of lepton number**

  Just as the total number of quarks can't change, neither can the total number of leptons. Again, we see this in beta decay. The neutron decays into a proton, electron and anti-neutrino. It creates an electron, which changes the lepton number by $+1$, but this is accompanied by an anti-neutrino which, because of the "anti-" contributes -1 to lepton number. The overall lepton number remains unchanged.

So the laws of physics seem to give two conserved quantities, baryon conservation and lepton conservation. And it's a true statement that we've never observed any process in which either these conservation laws are violated. Nonetheless, there are good reasons to think that they are not true conservation laws of nature. For this reason, I'll call them "probably almost conservation laws".

There are two reasons to think that baryon number and lepton number are not exactly conserved. While theoretically sound, neither of these is going to be high on

an experimenters list of things to do. However, as I explain below, there are at least two experiments that might, with some level of optimism, tell us more in the future.

Reason 1: Electroweak Instantons

The first reason that baryon and lepton number are not likely to be precisely conserved quantities is, in many ways, the most subtle since it hinges on some of the deeper ideas relating geometry and quantum field theory. When we first met Feynman diagrams in Section 2.2, we explained that they were a way to give approximate answers to questions in quantum field theory. But there are some effects in quantum field theory that Feynman diagrams miss completely! These effects are, in some sense, smaller than any given Feynman diagram. (They are also closely related to the ideas of quantum anomalies that we discussed in Section 4.1.3.) And, in the Standard Model, there is such an effect that can turn a baryon into a lepton.

This process is due to an object known as an *electroweak instanton*. It turns out that it can't turn a proton into a positron: the proton is absolutely stable in the Standard Model. However, it can turn a collection of three baryons into three leptons. (The factor of three is actually related to the existence of three generations!) This means, for example, that the ^{3}He nucleus is unstable to decay into three leptons, say a couple of positrons and an anti-neutrino.

Now, we haven't ever observed such a decay. And for good reason. If you compute how long it would take for a helium nucleus to decay by this process, you get a silly number: something like 10^{173} years. Our universe has lasted around 10^{10} years. Clearly, it's unrealistic to think that we would ever observe this process. Nonetheless, strictly speaking in the Standard Model the baryon and lepton numbers are not individually conserved. Instead, there is only a single conserved quantity

$$\text{Conserved quantity} = \text{(number of baryons) - (number of leptons)}$$

This quantity is usually called simply $B - L$.

Reason 2: Black Holes

Black holes aren't black. Hawking taught us long ago that they slowly emit radiation due to quantum effects. While there is much that we don't understand about quantum gravity, the existence of Hawking radiation stands out as one of the few robust and trustworthy calculations that we can do. The prediction of this radiation follows from the known laws of physics, and doesn't rely on any speculative ideas about what lies beyond.

If we wait long enough (and, again, we're talking ridiculously long times here), any black hole will eventually evaporate and disappear. So we can ask: what became of the stuff that we threw in?

First, the black hole can't destroy electric charge. If you throw, say, an electron into a black hole then the black hole itself now carries the electric charge. Moreover, this is visible outside of the event horizon because the black hole emits an electric field. That electric field can't just disappear. So, as the black hole evaporates, it must eventually spit out a charged particle – maybe an electron, maybe an anti-proton – which carries the electric charge. The process of black hole evaporation must respect conservation of electric charge. Similarly, black hole evaporation respects the conservation laws of energy, momentum and angular momentum.

In contrast, there is nothing to prevent black holes from destroying baryons and leptons. When a black hole forms from the collapse of a star, it will typically contain around 10^{57} protons, and roughly the same number of electrons. But there's nothing like an electric field outside the black hole that tells you how many baryons and leptons are sitting inside. Furthermore, as the black hole evaporates there's no reason that it should spit out these particles in tact. In fact, the vast majority of the mass of a black hole will be emitted in gravitational and electromagnetic radiation rather than baryons or leptons. In this way, we expect black hole evaporation to respect neither baryon number nor lepton number conservation.

Before we go on, I should stress an important point. There is an interesting and long standing problem about whether the *information* of what's thrown into a black hole is lost. We think that the answer is no. But, importantly, this isn't in contradiction with the violation of conservation laws!

To see why this is, consider the analogy of burning a book. In principle, the information written on the pages isn't lost: it's encoded in some impossibly complicated way in the correlations of the light and smoke that are emitted, and in the cinders that remain. Although the information is retained, the individual letters in the book are clearly lost.

In this analogy, the baryons in a black hole are like the letters. If you're clever enough and persistent enough, you may be able to detect that baryons were once present in the subtle correlations of the photons emitted by the black hole. But that doesn't change the fact that the baryons themselves have, almost certainly, gone for good.

Possible Experiments

No experimenter is lining up to test either of the theoretical arguments above. Nonetheless, the reasoning relies only on known, established laws of physics and tells us that, unlike electric charge, there is no fundamental reason for the conservation of lepton or baryon number. One might then wonder whether the violation of these conservation laws can be seen in some less extreme circumstance, one that could be tested here on Earth. There are two classes of ongoing experiments designed to test this.

Proton Decay

As mentioned above, within the Standard Model, the proton is absolutely stable. Nonetheless, it may well be that there is some physics beyond the Standard Model that causes protons to decay. Obviously, if we ever observed such a thing it would give us an important clue for what's happening on the next level.

So far, all we have are lower bounds. From our failure to detect proton decay in experiments we can infer that the lifetime of the proton is greater than around 10^{34} years. This is already quite an impressive statement, since its significantly longer than the age of the universe which is about 10^{10} years! But every litre of water contains about 10^{25} hydrogen atoms, so if you take 100 million litres of water, stare at it for a year, and fail to see a proton decay then you start to get close to the bound. There are a number of experiments around the world doing exactly this. The best current bounds come from the super-Kamiokande water Cerenkov detector in Japan. (We'll learn more about this detector in Section D.4 when we discuss experiments on neutrino oscillations.)

Neutrinoless Double Beta Decay

There is a rare, but well understood phenomena in which two neutrons in a nucleus simultaneously decay to two protons. This is called *double beta decay*.

For example, ^{48}Ca is a rare isotope of calcium, making up less than 0.2% of the naturally occurring element. It consists of 20 proton and 28 neutrons. A standard beta-decay process would take ^{48}Ca to ^{48}Sc, but this isotope of scandium has a smaller binding energy than calcium and so the beta decay process doesn't happen. Instead, ^{48}Ca decays through the much rarer double beta decay process

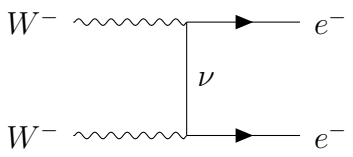
$$^{48}\text{Ca} \longrightarrow {}^{48}\text{Ti} + 2e^- + 2\bar{\nu}_e$$

The half-life is long: about 10^{20} years.

The double beta decay was first observed in the 1980s. As you can see, it violates neither baryon nor lepton number. However, it does raise the possibility for an even rarer occurrence: double beta decay without neutrinos. This would be a decay process of the form

$$^{48}\text{Ca} \longrightarrow {}^{48}\text{Ti} + 2e^-$$

With no neutrinos in the final state, lepton number is violated in this process. From the perspective of Feynman diagrams, the two neutrons would decay to protons by emitting two W^- bosons. These subsequently behave as follows:



Note that there's no arrow on the intermediate neutrino. Moreover, it's not possible to draw an arrow that lines up with both the external electron legs. Such a diagram is only valid if there's no way to distinguish between a neutrino and an anti-neutrino; indeed, this is what the observation of neutrinoless double beta decay would mean. We'll see how this might come about in the next section.

Despite many ongoing experiments around the world, neutrinoless double beta decay has yet to be observed. If it were seen, it would be a very big deal. Indeed, as we explain in the next section, it would give a key insight into the nature of neutrinos.

4.4 Neutrinos

No one would accuse a neutrino of being gregarious. They interact less than a first year undergraduate mathematics student forced to sit next to their theoretical physics professor at a matriculation dinner (to give a weirdly specific yet shudderingly memorable analogy).

For example, in the time it takes you to read this sentence, around 100 trillion neutrinos will have passed through your body. Most of them came from the Sun, but a significant minority have a cosmic origin, and have been streaming through the universe, uninterrupted since the first few seconds after the Big Bang. Moreover, in contrast to photons, the number of neutrinos hitting you doesn't change appreciably as day turns into night. The neutrinos from the Sun will happily pass right through the Earth and out the other side. This is vividly demonstrated in the picture of the Sun at night shown in Figure 39.

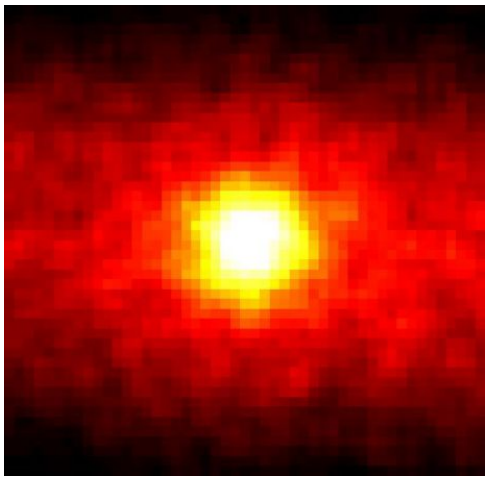


Figure 39. The Sun at night. This is a picture, taken by Super-Kamiokande, shows the neutrino flux coming from the Sun. The utterly remarkable fact is that the picture was taken at night, with the neutrinos passing through the Earth before hitting the detector.

There are two reasons why neutrinos are so intangible. The first is that they are the only particle to interact solely through the weak force. And, as we've seen, the weak force is weak. The second reason is that their mass is much much smaller than any other fermion which means that on the rare occasion that they do interact, they don't deliver much of a punch. The purpose of this section is to describe some properties of neutrinos in more detail.

4.4.1 Neutrino Masses

There is much that we don't know about neutrino masses. But we do know that the masses are not zero.

At the moment, we have no direct measurement of the mass of each neutrino. But we do have some precious information. First, we know that one neutrino must have a mass greater than

$$m_\nu \gtrsim 0.05 \text{ eV}$$

We'll explain *how* we know that neutrinos have a mass this big in Section 4.4.2.

Second, constraints from cosmology give us an upper bound on the sum of all neutrino masses. This comes from the imprint that neutrinos in the early universe leave on the cosmic microwave background radiation. (We'll say more about the intersection of

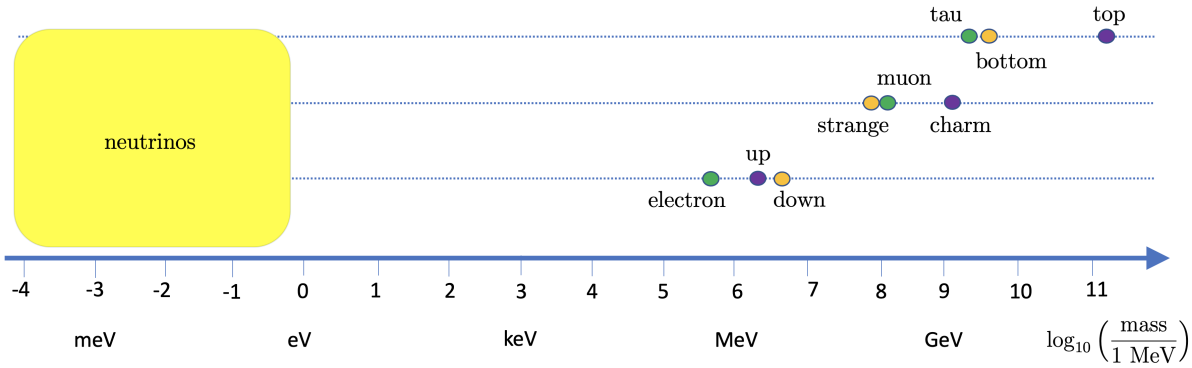


Figure 40. Fermion masses, arranged by generation. The charged leptons are green, the $-1/3$ quarks are orange, and the charge $+2/3$ quarks are purple. The neutrinos are way off to the left.

cosmology and particle physics in Section 5.3.). This bound is

$$\sum_\nu m_\nu \lesssim 0.25 \text{ eV}$$

There are still lots of possibilities consistent with these bounds. It may even be, for example, that one neutrino is massless while others have mass ~ 0.1 eV or so. Still, our ignorance notwithstanding, a rough summary of the masses of all fermions is shown in Figure 40. Even with our limited knowledge, it's clear that neutrinos aren't like the other particles. There is six orders of magnitude separating the mass of the top quark from the mass of the electron. Then there is a gap of another six order of magnitude before we get to the neutrinos. The first question we should ask is: why?

We don't have a definitive answer to this question. But we do have a plausible answer. As I will now explain, there are a number of different possibilities for the way that neutrinos get a mass, but some of these offer a clear explanation for why neutrino masses are so much smaller than those of the charged fermions. To understand this, we will need to delve once again into some intricacies of quantum field theory and the Dirac equation.

Let's first recall some facts that we learned previously. A massless fermion can come in one of two kinds: left-handed or right-handed. Furthermore, if a particle has one orientation then it's anti-particle necessarily has the opposite. Applied to neutrinos, this means that the particles that interact with through the weak force, which are necessarily the particles we detect in beta decay and other processes, are:

- A left-handed neutrino
- A right-handed anti-neutrino

When we refer to a "left-handed neutrino" below, this is shorthand for "left-handed neutrino and the right-handed anti-neutrino". They come together as a pair. You can't have one without the other.

In contrast, we've never observed the parity counterparts of these objects. This means that we've seen neither a right-handed neutrino nor a left-handed anti-neutrino in experiments. When we talk about a "right-handed neutrino" below, this will be shorthand for "right-handed neutrino and the left-handed anti-neutrino". Again, these come as an entwined pair.

The final important fact that we need is that massive particles arise by gluing together a left-handed fermion with a right-handed fermion. So the fact that neutrinos have a mass suggests that we should have both left- and right-handed neutrinos in the game, even though we can only directly observe the left-handed ones. However, nothing with neutrinos is as straightforward as it seems. One of the reasons they're special is because, as we saw in Section 4.1 (see, in particular, Table 2) the putative right-handed neutrino doesn't feel any force at all. It's this latter fact that is special to neutrinos and, as we now explain, opens up a new opportunity.

A Mass From the Higgs

First, it is quite feasible that neutrinos get a mass from the same mechanism as all other fermions, namely through interactions with the Higgs boson.

Recall that the interactions of all fermions with the Higgs field are characterised by a dimensionless number called a *Yukawa coupling*. The mass of the fermion is then given by (4.5)

$$\text{mass} = \frac{\lambda}{\sqrt{2}} \times 246 \text{ GeV} \tag{4.9}$$

where 246 GeV is the energy scale at which the Higgs settles down. The values of the Yukawa couplings range from $\lambda \approx 1$ for the top quark to $\lambda \approx 3 \times 10^{-6}$ for the electron.

This same story now repeats for the neutrinos. All you have to do is tune the value of the Yukawa coupling to $\lambda \approx 10^{-12}$, or whatever is needed to explain the three masses. In the context of neutrinos, this standard mechanism for generating mass is called a *Dirac mass*.

I stress that there's nothing logically wrong with this approach. The Yukawa couplings for charged fermions already range from 1 to 10^{-6} and so you may not feel unhappy by stretching the range down to 10^{-12}. But it does feel like the vastly different masses of neutrinos are crying out for a different explanation. Happily, one exists.

A Mass Without the Higgs and the Seesaw Mechanism

For all other fermions, the need for a Higgs field to induce a mass can be traced to the fact that the left-handed and right-handed particles experience different forces. The same is true for neutrinos, but with the key additional fact that the right-handed neutrino experiences no force at all.

But the lack of any force brings a level of freedom that other fermions do not enjoy. This is because the force that a particle experiences is used to distinguish between the particle and its anti-particle. For example, the difference between an electron and a positron lies in its electric charge. A particle that experiences no force — like the right-handed neutrino — may well be its own anti-particle.

We now tie this observation together with some facts about fermions. First, a fermion gets mass only by coupling together a left-handed and right-handed piece. Second, if we have a right-handed massless particle, then its anti-particle is left-handed. Third, if the particle experiences no force, then it can be its own anti-particle. When the dust settles, all of this means that the right-handed neutrino is unique among particles because it can get a mass *without* interacting with the Higgs field. It does so by interacting with itself! This kind of mass is called a *Majorana mass*, named after the Italian theorist Ettore Majorana who first realised this possibility in 1937.

We will discuss some consequences of the Majorana mass below but there is one important point that we flag up immediately. Forces are not the only way to distinguish particles from anti-particles: the conservation laws that we mentioned in Section 4.3.5 provide another. Recall that if we have a theory with conservation of lepton number, we count electrons and neutrinos as $+1$ and positrons and anti-neutrinos as -1. The converse of this statement is that if a neutrino is the same as an anti-neutrino, then there can be no conservation of lepton number!

This gives the smoking gun for a Majorana mass: in any theory where neutrinos have a Majorana mass, lepton number is not conserved. As explained in Section 4.3.5, we would then expect to see neutrinoless double beta decay. This is one reason why those ongoing experiments are so important: they will greatly help complete our understanding of the neutrino sector of the Standard Model.

So we learn that neutrinos can get masses in two ways:

- A Dirac mass m, from left-handed and right-handed neutrinos interacting with the Higgs field.
- A Majorana mass M, from the right-handed neutrino coupling to itself.

Of course, when we do experiments on a neutrino, we measure just one mass. Which one do we see? The answer is rather nice. It is the combination[13],

$$\text{mass} = \left| -\frac{M}{2} \pm \frac{1}{2}\sqrt{M^2 + 4m^2} \right| \tag{4.10}$$

The $\pm$ sign here is telling us that we should see two particles, each their own anti-particle, with different masses.

Now comes the key idea. The Dirac mass m comes from the Higgs mechanism, so is given by (4.9) for some unknown Yukawa coupling λ. But the Majorana mass M is unrelated to the Higgs mechanism and could be very large, $M \gg m$, perhaps coming from some unknown physics at a high energy scale that we have yet to understand. If this is true, the two masses in (4.10) are approximately

$$\text{mass} \approx M \quad \text{and} \quad \frac{m^2}{M} \tag{4.11}$$

The particle with mass M is essentially the right-handed neutrino and is very heavy. We have yet to detect this in any experiment. Meanwhile, the other particle, which can be identified with the left-handed neutrino that we observe experimentally, has mass m^2/M. The key point is that the Dirac mass need not be particularly small; it could take the same kind of value as the other quarks and leptons. But if the Majorana mass is bigger still, with $M \gg m$, the we would see a tiny neutrino mass m^2/M. If this is the way Nature works, then the tiny value of the observed neutrino mass comes about not because the Dirac mass m is very small, but because the Majorana mass M is very large. This in known as the *seesaw mechanism*.

As an example, here are some sample numbers. Suppose that the Dirac mass of the neutrino is as high as 100 GeV, comparable to the mass of the top quark. If the Majorana mass is around $M \approx 10^{15}$ GeV which, as we will discuss in Section 5.1.1, is the scale of grand unified theories, then we naturally get a neutrino mass of the right order of magnitude $\sim 10^{-2}$ eV.

[13]This formula comes from solving a quadratic equation. Specifically, it comes from finding the eigenvalues of the matrix $\begin{pmatrix} 0 & m \\ m & M \end{pmatrix}$ where the off-diagonal terms are the Dirac mass and the term in the lower-right is the Majorana mass.

A Mass Without a Right-Handed Neutrino at all!

The seesaw mechanism provides a very natural explanation for why neutrinos are so much lighter than everything else. However, we are notably relying on physics at a high energy scale, way beyond our current reach, and this mechanism clearly predicts the existence of a new particle of a big mass M that we have still to discover.

There's something a little unsettling about this. Because, at the end of the day, we introduced the right-handed neutrino only to find that it has a very high mass and can be ignored, leaving behind the light left-handed neutrino that we actually detect in experiments. But if the right-handed neutrino is so heavy that we can't see it, then surely there should be a way to describe the physics that it leaves behind without the need to invoke it!

Indeed, there is. Here is the way that it works. Although the left-handed neutrino experiences the weak force, it has the property, unique among all the fermions, that it can bind together with the Higgs field to produce something that is neutral under all forces. (You can see this if you compare the charges of fermions in Table 2 with the charges of the Higgs field in Table 4.2.) This means that although we can't introduce a Majorana mass for the left-handed neutrino alone, we can introduce a Majorana mass for the left-handed-neutrino-Higgs combo.

Clearly, this mass once again involves the Higgs field. However, it's different from the Yukawa terms that give the other fermions a mass. First, it involves only the left-handed neutrino, not the right-handed. Second, like any Majorana mass, it violates conservation of lepton number and so will give rise to neutrinoless double beta decay. And finally, and perhaps most importantly, it involves *two* interactions with the Higgs field and not just one[14]. This means that if $\langle\phi\rangle = 246$ GeV is the expectation value of the Higgs field, the Majorana mass of the left-handed neutrino will be proportional to ϕ^2. But this doesn't have the right dimensions to be a mass! This means that we must introduce another scale M, so that the resulting mass of the left-handed neutrino is roughly ϕ^2/M. This is the same form as we saw in the seesaw mechanism (4.11).

From this perspective, the mysterious new scale M is associated to some novel physics at a high energy that we don't yet understand. It could be the mass of the right-handed neutrino as in the seesaw mechanism, or it could be something else entirely. Either way, the irony of the seesaw mechanism remains: detecting a very small Majorana mass for the neutrino is clearly pointing to some new physics at a very high scale!

[14]In more precise language of effective field theory, it is a dimension 5 operator built out of Standard Model fields, in contrast to the dimension 4 Yukawa terms.

Some Other Way?

Above we've sketched the most plausible scenarios for neutrinos to get a mass. However, they're not the only possibilities. One could, for example, introduce further scalar fields that act like the Higgs boson, but carry different quantum numbers. If these too get an expectation value, it's possible to arrange for the neutrinos to get a mass. As you can see, our need to better understand the neutrino sector of the Standard Model has some urgency.

4.4.2 Neutrino Oscillations

So far we have described the different ways in which neutrinos can get a mass. But we haven't yet explained how we know that they have mass. After all, it's not like we can simply collect a bunch of neutrinos in a jar and weigh it. Instead, our information comes in a less direct manner.

We have met the key piece of physics already. In Section 4.3.3, we described how the assignment of mass to quarks is misaligned with the way the weak force acts on the quarks. This resulted in the phenomenon of *quark mixing*, described by the CKM matrix.

An entirely analogous phenomenon is at play in the lepton sector. It's simplest to explain what's going on by starting with two neutrinos, ignoring the third for now. To this end, we'll consider just ν_e and ν_μ. These are defined to be the neutrinos that couple to the electron and muon respectively, as in the following diagrams

But, just as with quarks, the ν_e and ν_μ particles that appear in these interactions are not the particles that have a well defined mass. Instead, there is a mixing and the neutrinos that have a specific mass are ν_1 and ν_2 defined by

$$\begin{aligned}\nu_1 &= \nu_e \cos\theta + \nu_\mu \sin\theta \\ \nu_2 &= \nu_\mu \cos\theta - \nu_e \sin\theta\end{aligned}$$

This is entirely analogous to the quark mixing that we saw in (4.7). It turns out that $\theta \approx 33^\circ$ for leptons, so that $\sin\theta \approx 0.55$. This is somewhat larger than $\theta_{\text{Cabbibo}} \approx 22^\circ$ that we saw in quark mixing. This, it turns out, is what storytellers call "foreshadowing".

At this stage of the argument, however, there's a slight change of perspective. In the context of quarks, when we hold a meson in our hand (metaphorically speaking) we know that it has a definite mass. The mixing then shows up because this meson interacts through the weak force with quarks of other generations

For neutrinos, this situation is reversed. If we've got a beam of neutrinos then it came from some phenomenon involving the weak force, usually associated in some way to beta decay. This means that we know our experiment emitted a neutrino like, say, ν_e with definite flavour but, at least if $\theta \neq 0$, this neutrino does not have a definite mass. What happens next is quite wonderful. The kind of particles that happily travel along without adventure are those with definite mass (known, in the language of quantum mechanics, as energy eigenstates.) But ν_e doesn't have this property. And this has a dramatic effect: as the beam travels some distance, the neutrinos oscillate from ν_e to ν_μ and then back again.

There is a fairly simple formula that describes how this happens. Suppose that the difference in the masses of ν_1 and ν_2 is

$$\Delta m^2 = m_2^2 - m_1^2$$

measured in eV. If the neutrinos have kinetic energy E (measured in GeV) and travel a distance L (measured in km) then the probability that ν_e transforms into ν_μ is given by

$$P(\nu_e \to \nu_\mu) = \sin^2(2\theta)\, \sin^2\left(1.27 \times \frac{L\,\Delta m^2}{E}\right) \tag{4.12}$$

The fact that this probability depends on sine functions is telling us that the change of flavour is an oscillation, in the sense that it goes back and forth. The formula contains two fundamental parameters: the mixing angle θ and the difference in masses Δm^2. To see oscillations, both need to be non-zero. The formula also contains two parameters that can vary from one experiment to another: the energy E of the beam and the length travelled L. In principle, by varying E and L, and seeing how one kind of neutrino morphs into another, we can determine the mixing angle θ and mass difference Δm^2.

We explain more about how these experiments are done in Section D.4. Here, instead, we focus on the results. For reasons that will become clear, I'll first describe what we know about the mixing angles and only then turn to the masses.

Neutrino Mixing Angles

With three generations, neutrino mixing is described by introducing a 3×3 matrix, entirely analogous to the CKM matrix that we met for quarks in (4.8). This is

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \tag{4.13}$$

On the left-hand side we have neutrinos ν_e, ν_μ and ν_τ that interact with their counterpart electrons through the weak force; On the right-hand side we have neutrinos ν_1, ν_2 and ν_3 that have definite mass. Relating them is a 3×3 matrix is known as the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix, or simply the neutrino mixing matrix.

The components of the PMNS matrix have now been measured to reasonable accuracy. The absolute values are roughly

$$\begin{pmatrix} |U_{e1}| & |U_{e2}| & |U_{e3}| \\ |U_{\mu 1}| & |U_{\mu 2}| & |U_{\mu 3}| \\ |U_{\tau 1}| & |U_{\tau 2}| & |U_{\tau 3}| \end{pmatrix} \approx \begin{pmatrix} 0.8 & 0.5 & 0.1 \\ 0.3 & 0.5 & 0.7 \\ 0.4 & 0.6 & 0.6 \end{pmatrix}$$

Some values are known fairly well; others less well. There are, for example, error bars of ± 0.1 on $U_{\tau 2}$.

The first thing to note is that the PMNS matrix is strikingly different from the CKM matrix describing the mixing of quarks[15]. In the quark sector, the CKM matrix was close to being the unit matrix, with just small off-diagonal elements. This meant that there was close alignment between the masses and the weak force.

But we see no such thing in the neutrino sector. The mixing is pretty much as big as it can be! Once again, we see that, in quantitative detail, the neutrinos really behave nothing like the charged fermions.

We do not have an explanation for the structure of the PMNS matrix. Indeed, its form came as a surprise to theorists. Surely it is telling us something important. It's just we don't yet know what!

[15] Recall that $\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}$ with $\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \approx \begin{pmatrix} 0.97 & 0.22 & 0.004 \\ 0.22 & 0.97 & 0.04 \\ 0.009 & 0.04 & 0.999 \end{pmatrix}$.

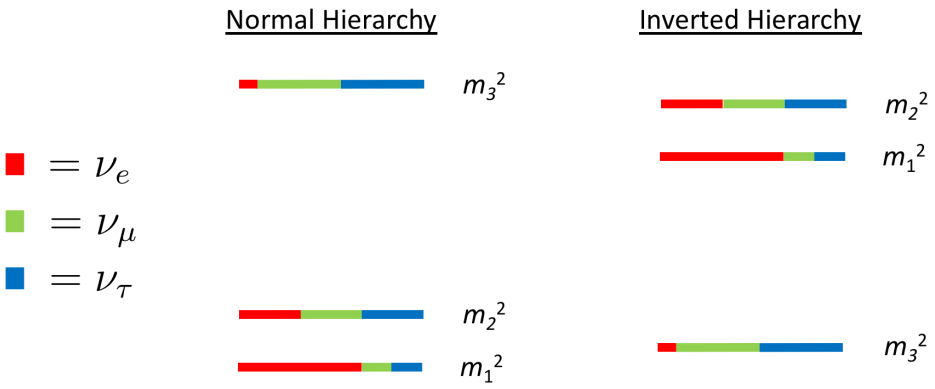


Figure 41. A colour coded description of the possible ordering of neutrino masses.

Neutrino Mass Differences

The mixing angles are a surprise. They tell us that each of the particles ν_1, ν_2 and ν_3 that have definite mass are not closely associated to the particles ν_e, ν_μ and ν_τ that have definite weak interaction.

Inverting the relation (4.13), we find can make the following statements about the neutrinos ν_1, ν_2 and ν_3 that have definite masses:

- ν_1 acts like an electron neutrino two thirds of the time, and as a muon or tau neutrino the other third.
- ν_2 acts like any one of the three neutrinos one third of the time.
- ν_3 acts like a tau neutrino 45% of the time and like a muon neutrino 45% of the time. The remaining 10%, it acts like an electron neutrino.

With this in place, we can now describe what we know about the mass differences. First, ν_1 is known to be lighter than ν_2 and the squares of their mass differ by

$$m_2^2 - m_1^2 \approx 7.4 \times 10^{-5} \text{ eV}^2$$

The difference in their masses is of order $\sim 10^{-2}$ eV, an order of magnitude smaller than the biggest mass. We also know the difference between the masses of ν_3 and ν_2 but, crucially, we don't yet know which one is heavier! We have

$$m_3^2 - m_2^2 = \pm 2.5 \times 10^{-3} \text{ eV}^2$$

Of course, if we could measure the mass difference between m_1 and m_3.then we would be able to resolve this $\pm$ ambiguity. As it stands, we just don't know the order of the masses.

The two possibilities are shown in Figure 41. Given the pattern seen in all other fermions, one might expect that the electron neutrino ν_e would be the lightest. Since the ν_e has the biggest overlap with ν_1, this would mean that ν_1 is lightest. This is referred to as the *normal hierarchy*. But, as we've seen, very little about the neutrinos follows our expectation. So another possibility is that ν_3, which contains very little of the electron neutrino, is the lightest. This is called the *inverted hierarchy*.

CP Violation Among Neutrinos

Neutrino mixing provides another forum in which the Standard Model can exhibit CP violation. And, as we described in Section 4.3.4, this also entails an opportunity for the violation of time reversal invariance.

In the quark sector, the violation of CP symmetry shows up as a complex number in the CKM matrix. The same is true in the neutrino sector. It turns out that if the neutrinos have only a Dirac mass, then there is a single complex number while if the neutrinos have a Majorana mass then there is an opportunity to introduce two more.

We do not currently have a good handle on these complex numbers. (Indeed, we have no handle whatsoever on the CP violation coming from Majorana masses.) However, preliminary results suggest that the complex numbers are non-vanishing and there is CP violation in the neutrino sector.

The good news is that, although difficult to measure, CP violation in the neutrino sector is conceptually rather more straightforward than in the quark sector. In particular, CP simply exchanges all left-handed neutrinos with right-handed anti-neutrinos. This means that if CP is preserved, the following probabilities are equal

$$CP \;\Rightarrow\quad P(\nu_\alpha \to \nu_\beta) = P(\bar{\nu}_\alpha \to \bar{\nu}_\beta)$$

where you can replace ν_α and ν_β with your favourite choice of flavour from ν_e, ν_μ and ν_τ. To detect CP violation, we (simply!) have to do an experiment that shows the probability for a neutrino to change flavour is not the same as the probability for its anti-particle to do the same thing.

Similarly, the violation of time reversal is also easier to state among neutrinos. In a world that is time-reversal symmetric, the probability of a neutrino morphing into a different one would coincide with the probability that this process is reversed,

$$T \;\Rightarrow\quad P(\nu_\alpha \to \nu_\beta) = P(\nu_\beta \to \nu_\alpha)$$

This is an experiment that could, in principle, be performed to discover a violation of time reversal invariance in our world.

Finally, as we discussed in Section 4.3.4, there is a mathematical theorem that says the laws of physics must be invariant under the combination CPT. In the language of neutrinos, this theorem tells us that

$$CPT \;\Rightarrow\quad P(\nu_\alpha \to \nu_\beta) = P(\bar{\nu}_\beta \to \bar{\nu}_\alpha) \tag{4.14}$$

A mathematical theorem is all well and good but it's nice to confront it with experiment. While the expectation is that CP and time reversal will both be found wanting in the neutrino sector, if it was found that the probabilities (4.14) don't coincide, that would surely rock our understanding of physics.

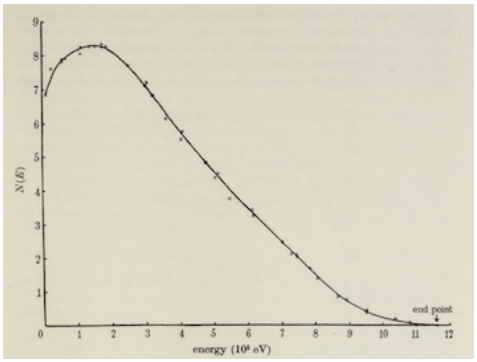
D Interlude: Big Science for Weak Things

For many decades, there was just one known manifestation of the weak force: beta decay. Much of our early understanding of the weak force came from careful study of this phenomenon.

We described some of the initial flurry of discovery in Interlude A. Radiation was found, largely by accident, in the darkness of Becquerel's desk drawer in 1896. By 1898, Rutherford had determined that the radiation from uranium consists of two different types – α rays and β rays – characterised by their penetrating power. And by 1900 it was understood that β rays were composed of the newly discovered particle called the electron. The next step was to further study the properties of these rays.

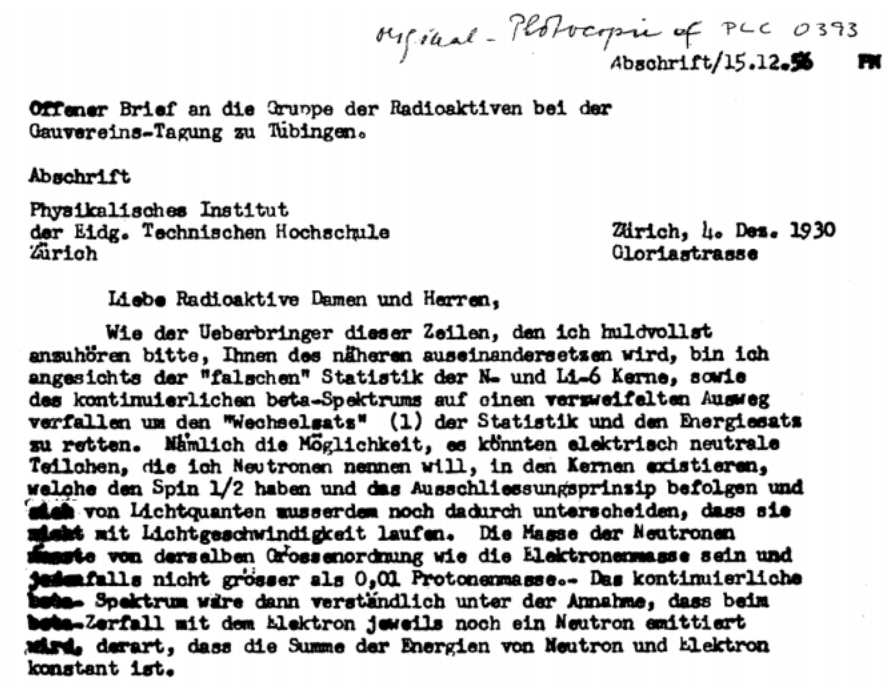
That, it turns out, was not easy. It had long been understood that α particles are emitted from the nucleus with a fixed velocity, and the general expectation was that the same would be true of β rays. However, the experiments were significantly harder. The leading experts in this game were Hahn, Meitner, and von Baeyer. Their first experiment confirmed that electrons were emitted with a uniform velocity; their second suggested a spread of velocities; their third a collection of distinct, but fixed, velocities. The situation was, to put it mildly, confusing.

The fog finally lifted in 1914. James Chadwick had completed his PhD under Rutherford and gone to work with Geiger in Berlin. (We met an older version of Chadwick in Section A.4 where we recounted his discovery of the neutron.) While the earlier experiments had used somewhat temperamental photographic plates to detect electrons, Chadwick used the counter recently invented by his postdoc mentor. His experiments made it clear that the electrons were emitted with a range of different velocities. A typical plot of energies is shown in the figure on the right[16].



1914 was not a great year to be an Englishman working in Berlin. Shortly after finishing his experiment, Chadwick was arrested and interned in the stables of a racecourse where he would spend the rest of the war, the monotony broken only by regular

[16]This was taken from a paper by G. J. Neary entitled "The β-Ray Spectrum of Radium E". (Radium E is now known as ^{210}Bi.)



Original - Photocopie of PCC 0393

Abschrift/15.12.30 PM

Offener Brief an die Gruppe der Radioaktiven bei der Gauvereins-Tagung zu Tübingen.

Abschrift

Physikalisches Institut
der Eidg. Technischen Hochschule
Zürich

Zürich, 4. Dez. 1930
Gloriastrasse

Liebe Radioaktive Damen und Herren,

Wie der Ueberbringer dieser Zeilen, den ich huldvollst ansuhören bitte, Ihnen des näheren auseinandersetzen wird, bin ich angesichts der "falschen" Statistik der N- und Li-6 Kerne, sowie des kontinuierlichen beta-Spektrums auf einen verzweifelten Ausweg verfallen um den "Wechselsatz" (1) der Statistik und den Energiesatz zu retten. Nämlich die Möglichkeit, es könnten elektrisch neutrale Teilchen, die ich Neutronen nennen will, in den Kernen existieren, welche den Spin 1/2 haben und das Ausschliessungsprinzip befolgen und sich von Lichtquanten ausserdem noch dadurch unterscheiden, dass sie nicht mit Lichtgeschwindigkeit laufen. Die Masse der Neutronen müsste von derselben Grössenordnung wie die Elektronenmasse sein und jedenfalls nicht grösser als 0,01 Protonenmasse.- Das kontinuierliche beta- Spektrum wäre dann verständlich unter der Annahme, dass beim beta-Zerfall mit dem Elektron jeweils noch ein Neutron emittiert wird, derart, dass die Summe der Energien von Neutron und Elektron konstant ist.

Figure 42. Liebe Radioacktive Damen und Herren.

deliveries of *Nature* and the occasional piece of scientific apparatus to help him pass the time.

Chadwick's work was not the end of the story, and a great deal of to-ing and fro-ing took place, largely between Lisa Meitner's group in Berlin and the pair of Chadwick and Charles Ellis, now back in Cambridge. But by the late 1920s, the situation was decisively settled: the spectrum of beta rays was continuous.

What to make of this? It was in sharp distinction to both alpha rays and gamma rays, where the radiation was emitted with a definite energy. Moreover, the emission of alpha and gamma rays was well understood as the transition between different states, just like the clean spectra of atoms. What could a continuous spectrum possibly mean?

D.1 The Neutrino

Two proposals were put forward in 1930. Bohr, ever the revolutionary, was keen to ditch energy conservation. Pauli, however, had a different idea: a new, hitherto undetected, particle. If two particles were emitted in beta decay, then the energy could be shared among them in different ways. There would then be no reason for the electron that we observe to carry a unique energy.

Pauli's first made this proposal public, just days after a messy divorce, in a famous "Dear Radioactive Ladies and Gentleman" letter sent to a conference in Tubingen. This is shown in Figure 42. The first paragraph translates as

> "I have come upon a desperate way out regarding the 'wrong' statistics of the N- and Li 6-nuclei, as well as to the continuous β-spectrum, in order to save the 'alternation law' of statistics and the energy law. To wit, the possibility that there could exist in the nucleus electrically neutral particles, which I shall call neutrons, which have spin 1/2 and satisfy the exclusion principle and which are further distinct from light-quanta in that they do not move with light velocity. The mass of the neutrons should be of the same order of magnitude as the electron mass and in any case not larger than 0.01 times the proton mass. The continuous β-spectrum would then become understandable from the assumption that in β-decay a neutron is emitted along with the electron, in such a way that the sum of the energies of the neutron and the electron is constant."

It is clear from the letter that Pauli is trying to solve two problems at once. We now know that these two problems in fact require two new particles. The first problem was described in Section A.4 and arose from the prevailing view that the nucleus is comprised of A protons with $A-Z$ electrons making up the charge difference. Pauli doesn't dissent from this viewpoint, but proposes that there are further neutral particles in the nucleus which ensures that the nucleon spin agrees with the sum of its constituents. As we saw, this problem was solved in 1932 by the discovery of the neutron.

The second problem addressed by Pauli is the one of interest here. The continuous spectrum of β-decay is resolved if a light, neutral particle is also emitted. Pauli calls this the neutron, but obviously that name was later taken. The Italian name *neutrino* was coined by Fermi in 1933.

Detection

Pauli was unduly nervous about his proposed neutrino. Indeed, his letter continues

> "I don't feel secure enough to publish anything about this idea, so I first turn confidently to you, dear Radioactives, with a question as to the situation concerning experimental proof..."

The neutrino hypothesis was accepted long before the particle was seen directly in experiment. It was an integral part of Fermi's theory of beta decay, which agreed too

well with the observed phenomenology. Moreover, the same theory made it clear how difficult a direct detection would be. As we've mentioned previously, a typical neutrino emitted in beta decay will interact only once as it travels through a light-year of lead.

The resolution to this problem is not to focus on a single neutrino. Instead, you need to go to a place where there are many. This was the goal of a team lead by Clyde Cowan and Fred Reines, working at the Los Alamos lab. Originally their idea was to detect the neutrinos emitted in a nuclear explosion, but they soon realised that a nuclear reactor offered a better (and presumably safer) alternative. Because of the ghostly nature of neutrinos, they named their original experiment "project poltergeist".

They set up their experiment in the Savannah reactor site, a nuclear reservation in South Carolina, USA. The nuclear reactor emits a flux of 10^{13} neutrinos per cm^2 per second, through beta decay

$$n \longrightarrow p + e^- + \bar{\nu}_e$$

The experimental challenge is to take the resulting neutrinos and observe the inverse process

$$\bar{\nu}_e + p \longrightarrow n + e^+$$

The bad news is that such processes are extremely rare. The good news is that both products n and e^+ have a distinctive signature. The experimental set-up consisted of two tanks, each holding 200 litres of water. These provide the targets for the neutrinos. The resulting positron quickly annihilates with an electron in the water, emitting two gamma rays. Slightly later, the neutron is captured. To aid this process, Cowan and Rienes dissolved cadmiun salts in the water. These are known to efficiently capture neutrons, emitting a third gamma ray in the process with a well understood energy spectrum. The signature of neutrinos is then two coincidental pulses of gamma rays, the first a pair; the second, delayed by a few milliseconds, a single photon, all of which were registered by liquid scintillators which surrounded the two tanks.

The experiment ran for almost 1400 hours and, when the reactor was on, detected roughly 3 coincidental pulses an hour.

Announcing a major discovery must be both exciting and nerve-wracking. Announcing the discovery to Wolfgang Pauli, not known to suffer fools gladly, doubly so. But on June 14$^{\text{th}}$ 1956, Cowan and Rienes sent Pauli a telegram that read:

> "We are happy to inform you that we definitely detected neutrinos from fission fragments by observing inverse beta decay of protons. Observed cross section agrees well with the expected six times ten to minus forty four square centimeters"

For some reason, Pauli's reply was never sent. It exists only in the form of a draft discovered later in his papers and reads: "Thanks for the message. Everything comes to him who knows how to wait."

The Muon Neutrino

It had long been known that when charged pions decay to a muon, they also emit a second spin 1/2 particle with all the properties (or lack thereof) of a neutrino.

$$\pi^+ \longrightarrow \mu^+ + \nu_\mu \quad \text{and} \quad \pi^- \longrightarrow \mu^- + \bar{\nu}_\mu$$

We now know, of course, that this is a different kind of neutrino from the one that appears in beta decay. The question is: how to tell ν_μ and ν_e apart?

One way is to run the inverse beta decay experiment again. What do you see if this neutrino collides with a proton? The two obvious possibilities are:

$$\nu_\mu + p \xrightarrow{?} n + \mu^+$$
$$\text{or} \quad \nu_\mu + p \xrightarrow{?} n + e^+$$

If the muon neutrino ν_μ is, in reality, the same object as the electron neutrino ν_e, then both of these processes should occur, presumably with comparable frequency. In contrast, if ν_μ is truly distinct from ν_e, then you would expect to see only the first process and not the second.

At this stage, accelerators once again come to the fore. A synchrotron can accelerate protons to, say, 30 GeV which, upon a collision with a fixed target, create a beam of pions. 25 meters down the road, these pions decay into a beam of muons and neutrinos. At this point, you need to put some shielding in place to remove the muons, leaving behind just the neutrinos. In the original experiment, performed at Brookhaven in 1962 by Lederman, Schwartz, Steinberger, and others, this shielding was achieved by 5,000 tonnes of steel from a decommissioned battleship (often falsely claimed to be the USS Missouri, which seems unlikely given that this ship, although mothballed in the late '50s, was subsequently reactivated and even served in the gulf war).

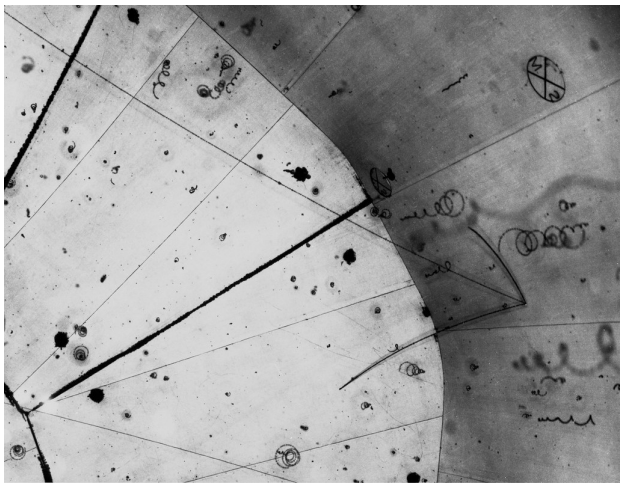


Figure 43. Muonic inverse beta decay, seen in a bubble chamber. The neutrino entered from the right where it hit a proton at the vertex where three lines meet. The proton is the short line moving up and to the left; the pion the slightly longer line moving down and to the left. The longest line of all, exiting top-left, is the muon. This image was taken at Argonne National Laboratory.

Waiting on the other side of the battleship, was a novel kind of detector known as *spark chamber*. This consists of an array of plates, sitting at a high voltage, with gas between them. When a charged particle passes through it ionises the gas which, when it hits the plates, leaves a trail of visible sparks. One can easily distinguish the track a muon from the much lighter electron.

One advantage of using accelerators is that the neutrino cross-section increases with energy. As Pais puts it, while you need a lightyear of lead to stop an MeV neutrino, a few million miles will do the job for a GeV neutrino. And, indeed, the 30 GeV machine was just sufficient to see the desired result: after several months of experiments, starting with roughly 10^{17} protons hitting the target, 29 anti-muons were seen and no positrons. This means that the subscript matters: ν_μ is indeed different from ν_e.

In subsequent years, muonic inverse beta decay could be seen in bubble chambers. The first such example is shown in Figure 43.

D.2 Not P and Not CP Either

We started Chapter 4 with a description of parity violation, since this sets the scene for much of the subsequent structure of the Standard Model. As we reviewed there, parity violation was first seen in the realm of atomic physics through the beta decay

of cobalt atoms. But, prior to this, there was a hint of parity violation in the world of particle physics.

Heavy mesons provide the setting to see violation of both parity and of CP. Here "heavy" means mesons that include strange, charm or bottom quarks. The lightest of these is, of course, the strange quark and this is where the phenomena were first observed.

The story starts with what we now call the charged kaon, $K^+ = \bar{s}u$, which weighs in at 494 MeV. Back in the 1950s, before quarks were understood, this particle was seen in experiments. But it had an unusual property: sometimes it decayed to two pions, and sometimes it decayed to three pions. In fact, this difference was so striking that, at the time, it was assumed that the experiments must be seeing two different particles. They gave these particles names θ and τ. Both names have since been retired and, in the case of τ, upcycled into the name of a lepton, but back in the 1950s these were two of the most interesting particles around. They decayed as

$$\begin{aligned} \theta^+ &\longrightarrow \pi^+ + \pi^0 \\ \tau^+ &\longrightarrow \pi^+ + \pi^+ + \pi^- \end{aligned}$$

But then there was a surprising coincidence that needed an explanation: as far as the experiments could tell, θ^+ and τ^+ had exactly the same mass and lifetime! Why on earth would that be? This was known as the *theta-tau puzzle*.

Of course, we know now the resolution to the theta-tau puzzle: it's that both particles have the same mass because they're actually the same particle – the kaon K^+. But physicists in the 1950s were reluctant to draw this obvious conclusion because it violated one of their cherished principles of physics: that the world should be invariant under parity.

It's not so easy to explain why this is the case without going into the mathematics. But it turns out that the decay to two pions looks identical when reflected in the mirror, while the decay to three pions does not. This isn't because of any obvious reason due to the directions in which the pions fly out. Instead, it shows up only in the subtle fact that the wavefunction of three pions differs by a minus sign upon reflection.

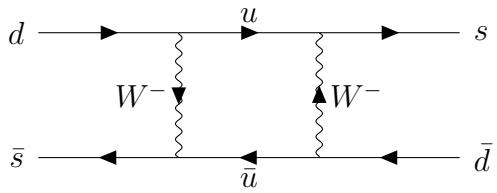
Given this experimental observation, parity could be conserved only if θ^+ and τ^+ were genuinely different particles, with θ^+ staying unchanged when reflected in a mirror, while τ^+ picked up a minus sign in its wavefunction. Conversely, if you want to identify the θ^+ and τ^+ and say that they're the same particle, then you can do so only if you sacrifice the symmetry of parity.

Needless to say, the argument above is somewhat subtle and was far from enough to convince physicists that parity was indeed broken in the weak interaction. The first physicists to take this possibility seriously were T.D. Lee and C.N. Yang (yes, Yang-Mills Yang) who, in 1956, did a systematic analysis of parity symmetry in various experiments. They came to the conclusion that, while there was overwhelming evidence for parity in the strong and electromagnetic forces, the jury was still out when it came to the weak force. They also proposed the experiment involving the decay of cobalt atoms that was subsequently performed by their colleague, C.S. Wu.

Wu's experiment showed that parity isn't just violated in the weak force a little: it is violated as much as it possibly could be. This spurred particle physicists to find other pieces of evidence. Confirmation came quickly from showing that muons arising from $\pi^+ \longrightarrow \mu^+ + \nu_\mu$ decay are polarised in a manner that can only be explained by parity violation. Moreover, the signature was so stark that even though the experiment was initiated after hearing of Wu's results, it was completed before she had finished. In the end, the two papers establishing parity violation in very different ways were published back to back.

Neutral Kaons and CP

Kaons also played a starring role in the discovery of CP violation. This time it was the neutral kaon $K^0 = d\bar{s}$ and its anti-particle $\bar{K}^0 = \bar{d}s$ that were of interest. The phenomenon of quark mixing that we met in Section 4.3.3 means that heavy, neutral mesons of this type have an interesting property: over time, the particle can change into its anti-particle and then back again! This happens through a Feynman diagram of the form



As we learned when discussing quark mixing, those intermediate u and $\bar{u}$ quarks could also be replaced by c and $\bar{c}$, or t and $\bar{t}$.

This means that as a K^0 wends along its merry way, it's constantly switching between a K^0 and $\bar{K}^0$. The actual particles that we observe are some mix of the two. But what mix? Here's where things get interesting.

Under CP symmetry, a K^0 switches with a $\bar{K}^0$ (because particles swap with anti-particles). So if the kaon was to arrange itself in accord with CP, the right mix would simply be

$$K_1^0 = K^0 + \bar{K}^0$$
$$K_2^0 = K^0 - \bar{K}^0$$

Then K_1^0 reflects into itself under CP, while K_2^0 reflects into minus itself. (These words make more sense when written as quantum mechanical equations! The right statement is that the two combinations above are CP eigensates.)

Moreover, just as parity conservation (if it existed!) would dictate the possible decays of particles, so too would CP conservation. CP symmetry means that we should see the two different species of kaons decay in different ways

$$K_1^0 \longrightarrow \pi^0 + \pi^0$$
$$K_2^0 \longrightarrow \pi^0 + \pi^0 + \pi^0$$

So do we?

The K^0 and $\bar{K}^0$ do indeed mix into two different combinations and, experimentally, these are distinguished by their lifetime. There is a long-lived neutral kaon that is called K_L and a short-lived neutral kaon that is called K_S. At first glance it seems as if all is good, since these two different combinations decay as

$$K_S^0 \longrightarrow \pi^0 + \pi^0 \text{ in around } 10^{-10} \text{ s}$$
$$K_L^0 \longrightarrow \pi^0 + \pi^0 + \pi^0 \text{ in around } 5 \times 10^{-8} \text{ s}$$

This makes it look like we can identify $K_S^0 = K_1^0$ and $K_L^0 = K_2^0$, in which case CP would be preserved.

However, a careful examination shows that this isn't quite the case. In 1964, Christenson, Cronin, Fitch and Turlay created a beam of neutral kaons, and let it travel for 18 m. Travelling at close to the speed of light, it takes around 6×10^{-8} s to travel 18 m, meaning that the short-lived K_S^0 kaons had long since departed and the beam contained only K_L^0 kaons. The goal was to see if they did, indeed, all decay to three pions as expected by CP.

They did not. Of the roughly 23,000 decays, they found 45 decaying into two pions (both $\pi^0 + \pi^0$ and, more commonly, $\pi^+ + \pi^-$). This is only 0.2% of the sample, but was enough to show that CP is not a symmetry of our world. The kaon states that have a definite mass are not quite the K_1^0 and K_2^0 states compatible with CP, but instead take the form

$$K_S^0 \approx K_1^0 - 0.002\, K_2^0$$
$$K_L^0 \approx K_2^0 + 0.002\, K_2^0$$

That tiny extra 0.002 piece is is the sign of CP violation.

The signature of CP violation in quarks is remarkably subtle. The effect is stronger in meson involving bottom quarks, but the essence of the idea remains the same. Recall from Section 4.3.3 that among the indirect implications of this result is the statement that the laws of physics are not invariant under time reversal. It seems surprising that the observable consequences of something so profound can only currently be seen in something so subtle and seemingly insignificant as the decay of certain mesons.

D.3 The Bosons of the Weak Force

Before the discovery of the W-boson, there was beta decay. Before the discovery of the Z-boson, there were *neutral currents*. This is the name given to the process in which a neutrino sneaks in, gives a charged lepton a gentle kick, and then quietly leaves again. For example,

$$\nu_e + e^- \longrightarrow \nu_e + e^-$$

Since we don't actually see the neutrino, the signal is rather subtle: the electron simply flies off without warning.

The first detection of neutral currents came in 1973 at CERN. A proton synchrotron (known simply as the PS) was used to create a beam of muon neutrinos which then entered an enormous bubble chamber, given the poetic name Gargamelle. You can see the discovery picture in Figure 44.

By this time, the theory of the weak force, with its W- and Z-bosons, was in place and it was possible to get a ballpark figure for their mass. But to find them required a machine that could reach the required energy. Something just shy of a 100 GeV or so should do it.

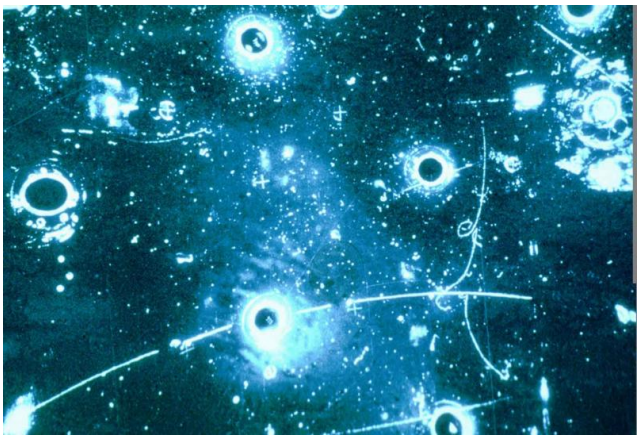
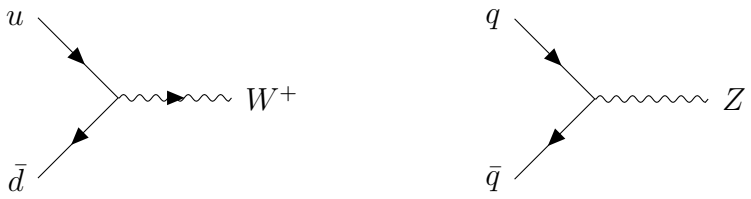


Figure 44. The first detection of neutral currents. The neutrino came in somewhere at the top and sent an electron barreling downwards, creating a shower of e^+e^- pairs, with their tell-tale spiral tracks. A black and white version of this appears in the discovery paper "Search for Elastic Muon-Neutrino Electron Scattering". The same journal contains a second paper with evidence for neutral currents from hadrons.

The W and Z Bosons

By 1976, CERN had turned on its latest machine, uninspiringly named the Super Proton Synchrotron or SPS. The proton beam reached 300 GeV before colliding into a fixed target. Sadly, as we mentioned before, much of the energy in fixed target machines ends up in the kinetic energy of the final product so only $\sqrt{2m_pE}$ of the energy, or about 25 GeV in this case, is available to do something useful. And that's not enough.

In the same year that the SPS turned on, Rubbia, Cline and McIntyre wrote a paper proposing to add a counter-rotating beam of anti-protons to CERN's newest accelerator. The centre-of-mass energy is then $300 + 300 = 600$ GeV. It turns out that around half the momentum in a proton is carried by the gluons, with the remainder split evenly between the three valence quarks. That means that each valence quark carries roughly 1/6 of the energy, so 600 GeV should be sufficient to create both W and Z-bosons through Feynman diagrams of the form



where q and $\bar{q}$ can be any quark-anti-quark pair.

Not everyone in CERN was overjoyed at the prospect of their new toy undergoing such an extensive overhaul so soon after turning on but Carlo Rubbia was, by all accounts, a persuasive if not particularly likeable man. The argument that he might move to Fermilab and do the experiment there was perhaps the decisive one and the SPS was upgraded to the SppS, or Super Proton-Anti-Proton Synchrotron. A key part of the design was an idea to focus the beam known as *stochastic cooling*, introduced by Simon van der Meer.

The first collisions occurred in December 1981 and were recorded in two detectors known as UA1 and UA2, the "UA" standing for "underground area", referencing the fact that both beam and detectors sit 100m below the surface. The goal was to see the decay products.

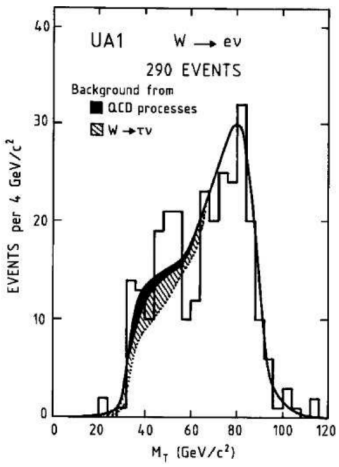
First, the W-boson. This decays about 70% of the time to quark-anti-quark pairs, such as $W^+ \to u\bar{d}$, and these appear in the detector as jets. But jets are ten a penny in $p\bar{p}$ collisions. For this reason, you need to focus on the 30% of the time that the W-boson decays as $W^+ \to \bar{l}\nu$ or $W^- \to l\bar{\nu}$ with the lepton l either an electron or muon.

If the W-boson was sitting at rest when it decayed, we would get a back-to-back leptons with momentum $\mathbf{p}_l = -\mathbf{p}_\nu$. Of course, you miss the neutrino in the detector, so you should see the electron or muon fly off with its kinetic energy[17] given by $|\mathbf{p}_l| = M_W/2$, half the mass energy of the W-boson.

In reality, the W-boson usually isn't at rest when it decays, but travelling at some unknown velocity aligned with the direction of the beam. Furthermore if the lepton also travels in the same direction of the beam, then you're simply going to miss it: there's too much going on down there and no detector. The trick, therefore, is to look at the *transverse momentum* $\mathbf{p}_T$ of the lepton, meaning the momentum orthogonal to the beam. This obeys $|\mathbf{p}_T| \leq M_W/2$, with equality occurring only when you're lucky, and the W-boson was at rest *and* the lepton emerges perpendicular to the beam. However, one can show that the expected distribution of $|\mathbf{p}_T|$ from many W-boson decays takes a specific shape, peaking at $M_W/2$ and then quickly dropping off.

[17]For slowly moving particles, where you can ignore relativity, the kinetic energy is quadratic in the momentum: $E = \frac{1}{2}mv^2 = \frac{1}{2}p^2/m$ where the momentum $p = mv$. But this formula is no longer right when particles move fast. The correct relativistic formula combines the rest mass energy and kinetic energy as $E = \sqrt{m^2c^4 + p^2c^2}$. When particles move extremely fast, with $pc \gg mc^2$, we can neglect the rest mass energy and the kinetic energy is approximately linear in the momentum: $E \approx pc$. This is the formula we've used in the text, with units $c = 1$.

Using this method, the first clear signs of the W-boson were seen in January 1983. The discovery was announced soon afterwards. The figure on the right shows that data collected by UA1 between 1982 and 1985. The data is plotted using a variable related to $\mathbf{p}_T$ called *transverse mass*, a slightly odd name when you first hear it given that mass has no direction.



The search for the Z-boson is cleaner because this time there's no neutrino to miss and you see back-to-back electron-positron pairs. However, the events are rarer, with a frequency of around 10% of the W-bosons decay. For this reason it took a few months more before the Z-boson was also discovered.

While W- and Z-bosons were first discovered at a $p\bar{p}$ collider, their detailed study came with the next upgrade at CERN. This was the Large Electron Positron Collider, or LEP, a 200 GeV machine built in a new tunnel, some 27 km in circumference. While hadron collisions are a mess, e^+e^- collisions allow for exquisitely precise measurements. It is from this experiment, with its four detectors Aleph, Delphi, Opal, and L3, that we have our most accurate understand of the weak force. The current values of the masses are now known to be

$$M_W = 80.379 \pm 0.0021 \text{ GeV} \quad \text{and} \quad M_Z = 91.1876 \pm 0.0021 \text{ GeV}$$

with comparable precision on their decay widths.

However, there was one last part of the Standard Model that was tantalisingly just out of reach of the LEP machine. This was...

The Higgs Boson

By the end of the last century, the LEP data told us that the Higgs boson must be heavier than 114 GeV. But how much heavier?

To answer this, LEP was dismantled and the Large Hadron Collider, or LHC, constructed in its place, designed to collide two proton beams with a centre of mass energy of 14 TeV. The full CERN accelerator complex is shown in Figure 45. There are a number of detectors around the ring, the most important of which are ALICE, used for heavy ion collisions, LHCb, used to study B-mesons, and two multi-purpose detectors called ATLAS and CMS. These latter two detectors were the ones to find the Higgs.

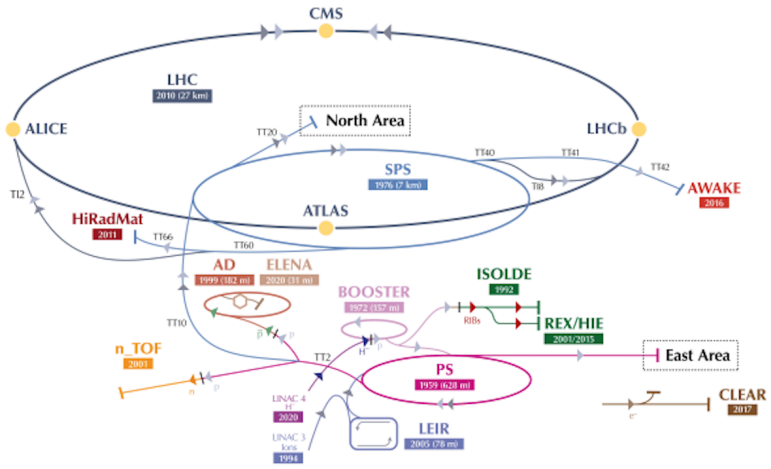
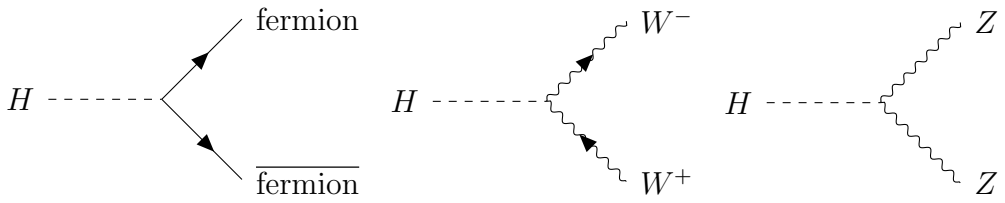


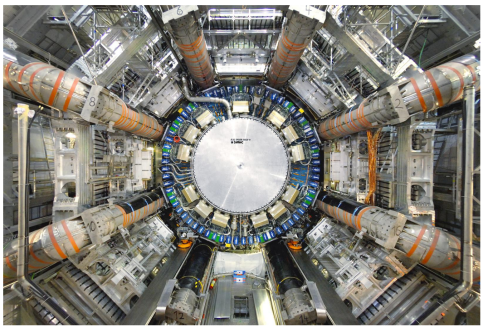
Figure 45. The CERN accelerator complex. You can see the older PS and SPS accelerators which now act as feeders for the LHC.

The Higgs, like the top quark of Interlude C.3, is detected through its decay products. It can decay either into fermions, or into W and Z bosons through diagrams like this



As we explained in Section 4.3, the strength of the interaction is proportional to the mass of the particle, because it comes from the same physics that gives fermions and W and Z bosons a mass in the first place. This means that the Higgs preferentially decays into the particles with higher masses.

The coupling to the top quark is the strongest, but because the top is heavier than the Higgs that decay route woud appear to be ruled out. That means that the decay $H \rightarrow b\bar{b}$ is the most common, but it is difficult to distinguish the resulting two jets from the background of a pp collision. The next two decay modes, $H \rightarrow W^+W^-$ and $H \rightarrow ZZ$, are significantly cleaner since the both W and Z can subsequently decay to leptons through $W \rightarrow l\bar{\nu}$ and $Z \rightarrow l\bar{l}$, giving two lepton and four lepton events respectively where the leptons are either electrons or muons. (Remember, no one sees the neutrino.) The decay through Z-bosons to four leptons is particularly clear and sometimes referred to as the golden channel.



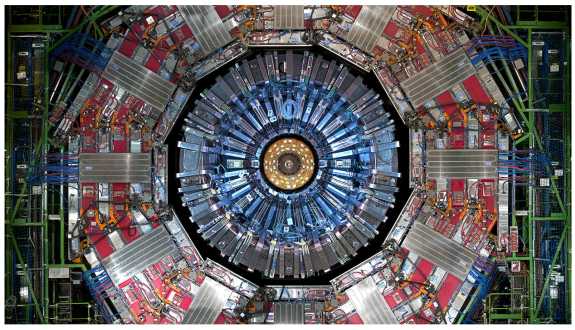
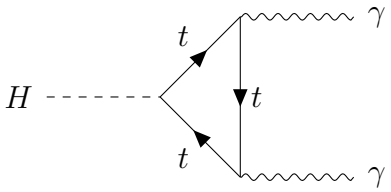


Figure 46. Iconic images of the ATLAS detector, on the left, and CMS, on the right, not to convey any information but simply to let you stare in awe at one of the great engineering feats of all time.

Given that the Higgs decays preferentially to massive particles, it is somewhat ironic that one of the best signals comes from the emission of two photons. These arise from top quarks which, although too heavy to appear directly as decay products, couple so strongly that the one-loop Feynman diagram is comparable to other tree level processes:



On July 4$^{\text{th}}$, 2012, ATLAS and CMS held a joint seminar in which they announced the discovery of the Higgs boson, the final piece of the Standard Model jigsaw. There is nothing as pretty as a bubble chamber photograph to show, merely a small bump above the background for the emission of two photons shown in Figure 47, revealing that the Higgs boson has mass

$$M_H = 125.10 \pm 0.14 \text{ GeV}$$

Since then, a number of different Higgs decay channels have been seen, all impressively (but, to some, disappointingly) in perfect agreement with expectations from the Standard Model.

There is much that we still don't understand about the Higgs, not least its self-coupling and more detailed information about the shape of the Higgs potential $V(\phi)$. There is hope that a future e^+e^- collider – perhaps the proposed ILC – could change

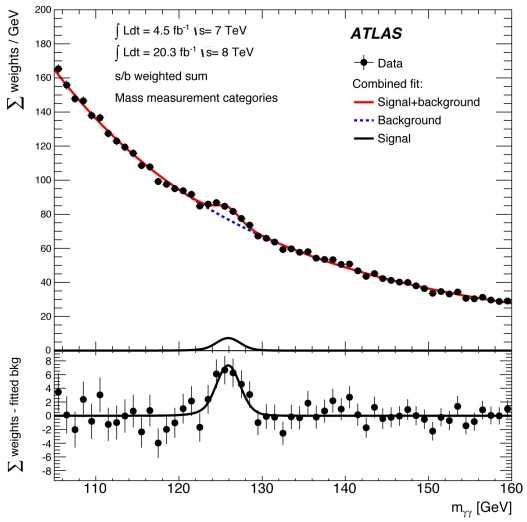


Figure 47. The tiny bump that revealed the Higgs, taken from the ATLAS experiment.

this, allowing us to understand the Higgs sector with the same precision that LEP explored the rest of the electroweak sector.

D.4 Neutrino Oscillations

We end this interlude the same way in which we began: with neutrinos. We saw how neutrinos were first discovered in Interlude D.1. Now we will learn about their mass, a long and involved story that, as we will see, is still far from complete.

Solar Neutrinos

The first hint that something was awry came when neutrinos failed to escape the Sun.

In the late 1960s, Ray Davies built the world's first solar neutrino detector, buried deep underground in the Homestake gold mine in South Dakota. The detector consisted of a large tank filled with a dry-cleaning fluid that, importantly, was rich in chlorine. The neutrinos were detected by a neutrino capture process, which is like stimulated beta decay

$$\nu_e + {}^{37}\mathrm{Cl} \longrightarrow {}^{37}\mathrm{Ar} + e^-$$

These argon atoms were then counted and used as a proxy for the original neutrino.

Neutrinos have the ability to induce such an interaction provided that their energy is greater than ~ 800 keV. However, they do so with extraordinarily small probability. We now know that around 10^{11} neutrinos that originate from the Sun stream through a surface area of one square centimetre here on Earth every second. Most, it turns out, aren't powerful enough to induce the reaction above, but there's still about 10^8, per square centimetre, per second, that can do the job. Yet the number of reactions observed in 600 tonnes of cleaning fluid was just a few a day. (Neutrino experiments use the unit of SNU, where 1 SNU means 10^{-36} interactions per target atom per second. The Homestake experiment detected about 2.5 SNU of solar neutrinos.)

There was a problem however. The observed solar neutrinos were a factor of 3 too small. Detailed calculations of the reactions in the Sun, largely performed by John Bachall, showed that the expected flux was around 8 SNU. Where did the other neutrinos go?

The reaction of the larger scientific community to this puzzle was mostly to shrug and move on. To particle physicists, the Sun looked like a ridiculously messy and complicated object; surely those astrophysicists had screwed up the calculations. Meanwhile, the astrophysicists were bewildered at the possibility that you could reliably detect neutrinos; surely the particle physicists had got something wrong in their experiment. After all, could we even be sure that it was detecting neutrinos from the Sun, and not some other source?

However, as time went on the problem became more urgent. Bachall's theoretical models of the Sun passed many impressive tests, leaving little doubt of their accuracy. Meanwhile further experiments confirmed and refined the Homestake results. In Japan, Kamiokande, and later Super-Kamiokande, detected neutrinos at higher energies through Cerenkov radiation emitted in scattering process

$$\nu_e + e^- \longrightarrow \nu_e + e^- \tag{D.1}$$

Importantly, this gave directional information about the incoming neutrinos and showed clearly that they were coming from the Sun. Meanwhile GALLEX in Italy, and SAGE in Russia, both repeated the experiment with gallium rather than chlorine, now relying on the neutrino capture process,

$$\nu_e + {}^{71}\mathrm{Ga} \longrightarrow {}^{71}\mathrm{Ge} + e^-$$

The advantage is that the threshold is somewhat lower than the chlorine reaction, needing only around 200 keV, meaning that many more of the Sun's neutrinos can

partake. Indeed, the number of events seen was significantly higher, at around 75 SNU, but still below the theoretical prediction of 130 SNU. Curiously, the shortfall in these experiments is only around 40%, compared to the 70% seen in the chlorine experiments. This is telling us that the loss of neutrinos is energy dependent. Collectively, these discrepancies between experiment and theory went by the name of the *solar neutrino problem*.

As the years went on, it became increasingly clear that the resolution to the solar neutrino problem must be found in neutrino oscillations which, as we have seen in Section 4.4, requires that the neutrinos have mass. The question was: how to test this? This requires us to count not just the electron neutrinos emerging from the Sun, but also the muon and tau neutrinos. And this is significantly harder. We can't rely on the obvious neutrino capture process

$$\nu_\mu + n \longrightarrow p + \mu^-$$

because the incoming neutrinos have energies less than the 150 MeV needed to create a muon.

An unambiguous resolution to the solar neutrino problem was provided by the Sudbury Neutrino Observatory (SNO), based in the Creighton nickel mine in Ontario, Canada. One novelty was that their tank was filled with heavy water, D_20, where the hydrogen is replaced by deuterium D. It doesn't take much to split the deuterium nucleus apart; just 2 MeV of energy is enough. Moreover, neutrinos can knock apart a deuterium nucleus in two different ways. A weak interaction involving an intermediate W-boson does the job through a neutrino capture process analogous to those that occur in chlorine or gallium,

$$\nu_e + d \longrightarrow p + p + e^-$$

Only electron neutrinos contribute to such processes. However, the neutrinos can also split the deuterium through a weak interaction involving a Z-boson,

$$\nu + d \longrightarrow n + p + \nu$$

This time there is no charged lepton created, meaning that all three kinds of neutrinos, ν_e, ν_μ and ν_τ contribute.

In addition, SNO measured neutrino scattering events of the form $\nu + e^- \to \nu + e^-$. Electron neutrinos undergo such scattering events through both W-boson and Z-boson interactions, but muon and tau neutrinos only scatter off electrons when an intermediate

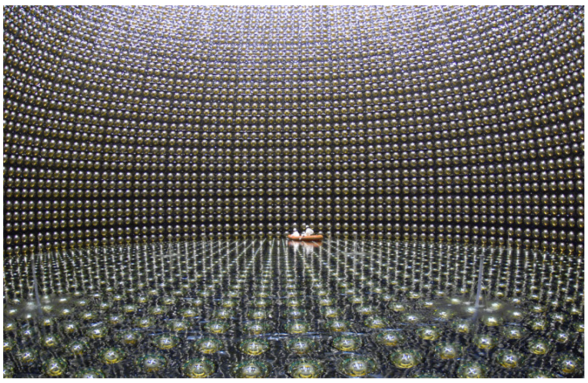
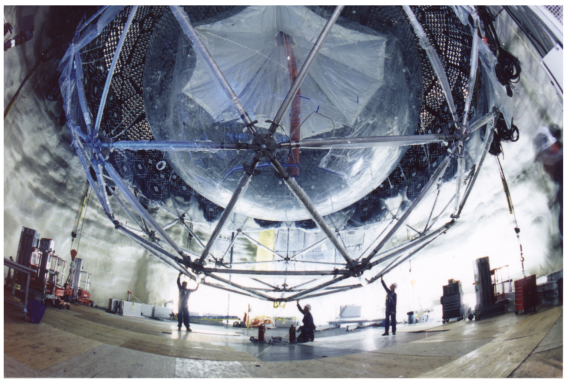



Figure 48. Neutrino detectors tend to look like the lair of a James Bond villain. On the left is a boat cleaning the Super-Kamiokande photosensors as the tank slowly fills up. On the right is the SNO tank, filled with heavy water.

Z-boson is involved. This means that the total rate of such events depends on some combination of the flux of electron, muon and tau neutrinos. (For what it's worth, it turns out to be ν_e flux + 0.15 times ν_μ and ν_τ flux.)

The upshot is that SNO was able to see everything – electron, muon and tau neutrinos. And once you see everything, nothing is missing. The end result agreed perfectly with theoretical expectations of the nuclear reactions inside the Sun. The electron neutrinos missed by previous experiments had transmuted into muon and tau neutrinos, incontrovertible evidence for neutrino oscillations.

Atmospheric Neutrinos

The story of missing neutrinos was repeated when we looked elsewhere. One key clue came from neutrinos created in the upper atmosphere. Cosmic rays, mostly in the form of protons or helium nuclei, are constantly bombarding the Earth. When they hit the atmosphere they create a constant stream of $\pi^\pm$ pions. These pions decay to muons

$$\pi^+ \longrightarrow \mu^+ + \nu_\mu \quad \text{and} \quad \pi^- \longrightarrow \mu^- + \bar{\nu}_\mu$$

and the muons then quickly decay to electrons,

$$\mu^+ \longrightarrow e^+ + \nu_e + \bar{\nu}_\mu \quad \text{and} \quad \mu^- \longrightarrow e^- + \bar{\nu}_e + \nu_\mu$$

Indeed, as we saw in Interlude B, this is how both the pion and muon were first discovered. Now, however, our interest lies in the neutrino by-products. These "atmospheric

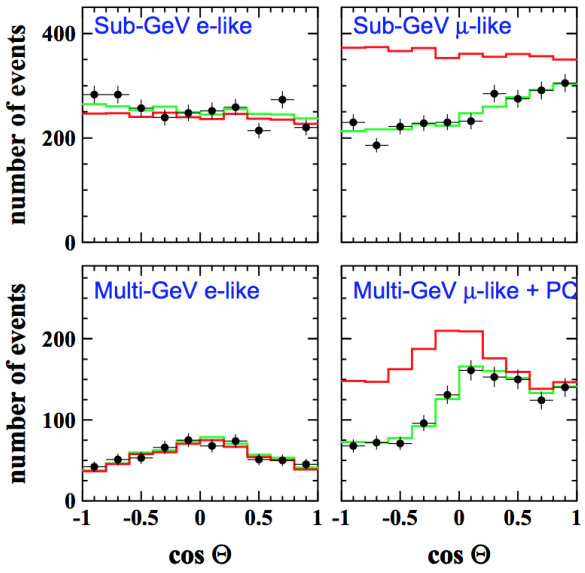


Figure 49. The observed flux of electron neutrinos (on the left) and muon neutrinos (on the right). The top boxes show low-energy neutrinos; the lower boxes high-energy neutrinos. The red line is the theoretical expectation without neutrino oscillations, and the black boxes the data.

neutrinos" have significantly higher energies than solar neutrinos; often around a GeV or higher. Given the decay processes described above, each collision should result in a two muon neutrinos (strictly one ν_μ, one $\bar{\nu}_\mu$) for every electron neutrino. The question is: can we find them?

The answer, given by Super-Kamiokande, is interesting and shown in Figure 49. These show plots of the neutrino flux (on the vertical axis) against the angle at which the neutrinos come into the detector (on the horizontal axis). An angle $\cos\theta = 1$, on the far right, means that the neutrinos come directly down. An angle $\cos\theta = -1$, on the far left, means that neutrinos come up, through the Earth.

The data on the left two boxes is for electron neutrinos, both for low-energy events (shown in the top box) and high-energy events (in the bottom box). The red line is the theoretical expectation; the black dots the observed flux. We see that the agreement between experiment and theory works well.

The story is more interesting for muon neutrinos, shown in the two boxes on the right. The number of neutrinos coming straight down agrees perfectly with what we expect, but there's a clear deficit for those that come up through the Earth. Why?

For any other particle, you might think that the Earth is simply getting in the way. But neutrinos pass right through the Earth without any difficulty. (Remember the picture of the Sun at night in Figure 39.) Besides: theorists aren't stupid and had taken the presence of the Earth into account when computing the red line! Instead, the key point is that the muon neutrinos have travelled further, and so had more opportunity to convert into other neutrinos, in this case tau.

Importantly, the atmospheric neutrinos clearly show us that neutrino oscillations depend on the length L that neutrinos travel. We have

$$\begin{aligned} \text{Straight down: } L = 15 \text{ km} &\quad \Rightarrow \quad \text{No oscillations} \\ \text{Straight up: } L = 13000 \text{ km} &\quad \Rightarrow \quad \nu_e \text{ unaffected, but } \nu_\mu \to \nu_\tau \end{aligned}$$

Meanwhile, for solar neutrinos we have $L = 150$ million km. As we'll now see, this collection of data goes a long way to allowing us to determine neutrino masses.

Getting a Handle on the Masses

Nature is kind, and gives us two sources of neutrinos: solar and atmospheric. As we've seen, these clearly show that something fishy is going on, as neutrinos appear and disappear. But to be sure, we should design our own experiments here on Earth, in which a source of neutrinos is created and subsequently detected elsewhere.

There are now a number of these experiments up and running, with results consistent with the oscillations seen in solar and atmospheric neutrinos. The best results so far have come from the T2K experiment, which directs a beam of muon neutrinos created in Tokai, Japan to the Super-Kamiokande detector located 300 km away. They clearly see ν_e neutrinos in a beam that, at its origin, consisted purely of ν_μ neutrinos. Meanwhile, the OPERA experiment in the Gran Sasso lab in Italy has directly detected ν_τ neutrinos in a ν_μ neutrino beam from CERN, 730 km away.

From this collection of experiments, together with results from solar and atmospheric neutrinos, we can put together the picture of masses and mixing angles that we described in Section 4.4.2.

5 What We Don't Know

It is often said that each new discovery opens up many more questions than it answers. That's not the case for the Standard Model. The collection of interlinked ideas, bound together in the Standard Model, has brought a synthesis that is unprecedented in science, bringing order to many, seemingly disconnected phenomena and leaving very few threads hanging as a result.

Furthermore, the current experimental situation is one of remarkable harmony. Wielding a broad brush, it is not too inaccurate to say that the Standard Model predicts the correct answer to each and every one of the thousands of particle physics experiments that we've performed.

That's not to say that everything is perfect. There is, as we have recounted in Section 4.4, much still to learn about the neutrino sector. Moreover, if you look in finer detail, then there are a handful of experimental anomalies that seemingly cannot be described by the Standard Model. The most longstanding of these is the magnetic moment of the muon. Recall that the magnetic moment describes how strongly a particle couples to a magnetic field. Our best theoretical result for the muon is

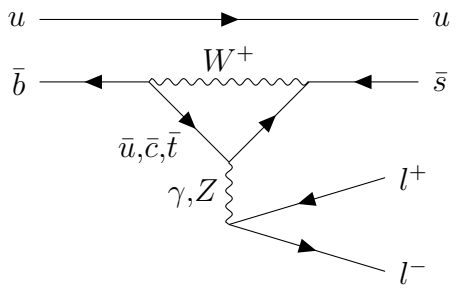
$$g_{\text{theory}} = 2.00233183602$$

while the experimental result is

$$g_{\text{expt}} = 2.00233184122$$

As you can see that, by the time you get to the 9^{th} decimal place, things don't quite match. In any other area of science, you wouldn't care less about a discrepancy in the 9^{th} decimal place. But here it matters. Taken seriously, the deviation between theory and experiment is at the level of 4.2 sigma. Optimistically, this discrepancy might be pointing to extra corrections to g_{expt}, beyond those of the Standard Model. However, there are reasons to be cautious. In particular, the theoretical result involves very difficult numerical simulations to determine contributions from the hadrons and there is some controversy over the accuracy of these results.

There are a number of further niggles too. In nuclear physics, the lifetime of the neutron seems to be slightly different depending on the way in which its measured. In particle physics, a collection of results around B-mesons seem to be slightly discrepant from Standard Model predictions, hinting that electrons, muons and tau leptons may differ in their interactions. The most prominent of these comes from looking at the decays of B-mesons to kaons and a lepton-anti-lepton pair, through a diagram like the following:



These are known as penguin diagrams. (As with the constellations, to see the resemblance you have to squint and reach into the depths of your imagination, before giving up and wondering what these people were smoking. I like to think of the lepton pair as the penguin's beak, but apparently they're the legs.) The quark running in the loop is either (anti) up, charm, or top, while the neutral boson is either a photon or Z. Finally, the end product is, in addition to the kaon, a lepton-anti-lepton pair where either $l = e$ or $l = \mu$.

Since the bottom quark is so heavy, the mass of the leptons is largely irrelevant. This is important because, if we ignore their mass difference, the electron and muon have identical couplings to the weak and electromagnetic forces, a fact that is sometimes called *lepton universality*. This means that the probability to decay to an electron should be the same as the probability to decay to a muon.

The best current measurements suggest that lepton universality is *not* respected in this decay: there is a preference to decay to electrons over muons. Taken at face value, it appears that this because something untoward is going on with muons, rather than electrons.

If these anomalies hold up to further analysis, then they are telling us something extremely important: the Standard Model needs replacing. They may, however, be due to random fluctuations and will disappear as the data improves. The B-meson results are currently 3.1 sigma from Standard Model expectations, somewhat short of the 5 sigma gold standard necessary to claim a discovery. At any given time in history their are always number of such mismatches between theory and experiment. Since the Standard Model was put in place, they have nearly all evaporated upon closer inspection.

These anomalies not withstanding, the current state of affairs is that the Standard Model works extraordinarily well. You have to look very very hard — like the 9^{th} decimal place! — to find clear disagreement between experiment and theory. However, at the same time, it is overwhelmingly clear that the Standard Model is not the last

word in physics and the purpose of this chapter is to describe some of the questions that remain, together with some speculative suggestions for how they may be resolved.

These issues fall into different categories. First, there are a number of unexplained aspects of the Standard Model itself, and these will be described in Section 5.1. Moreover, there is one part of physics that the Standard Model ignores completely: gravity. We will describe how this fits in to the bigger picture in Section 5.2. Finally, if you want some incontrovertible observational evidence that there are things not described by the Standard Model, then we should turn to heavens. In Section 5.3 we describe some of the many puzzles that come from cosmology.

5.1 Beyond the Standard Model

Nearly all the unanswered questions about the Standard Model come from looking more closely at the various constants of Nature and asking why they take the particular values that they do. Here the "constants of Nature" are the parameters that we input into the Standard Model.

For some of these parameters, our understanding is as good as it could possibly be. As explained in Section 4.1.3, the electric charges (or, equivalently, the hypercharges) of the various fermions simply can't be any other way. They are fixed to their values by the stringent requirements of quantum anomaly cancellation. What we would love is to have a similar level of understanding for the other parameters of the Standard Model. Sadly, as we will see, we are a long way from that.

So what are these parameters? Roughly speaking, they fall into three classes (although one could certainly make a more refined classification, especially for the third class). These are:

- Three constants that specify the strengths of the three forces. These are the fine structure constant and its counterpart for the strong and weak force. We'll discuss these in Section 5.1.1.
- Two parameters that specify the Higgs potential. These can be thought of as the mass and expectation value of the Higgs boson. We'll discuss the issues surrounding these in Section 5.1.2.
- Loads of parameters that specify the way the Higgs field interacts with various fermions, usually referred to as the flavour sector. These are the topic of Section 5.1.3.

For quarks, the parameters are six masses (or, equivalently, Yukawa couplings) and a further ten mixing angles that sit in the CKM matrix. (These ten then split further into 9 mixing angles and the phase that gives CP violation.) There is, in addition, one further parameter known as the QCD theta angle that we haven't yet mentioned because, as far as we can tell, it is zero. Nonetheless, zero is a number too and it deserves an explanation.

For leptons the counting a is a little more fuzzy. If the neutrinos get a Dirac mass, so that $B-L$ symmetry is preserved, then we again have six masses, or Yukawa, parameters and ten mixing angles in the PMNS matrix. If, however, neutrinos have a component that is a Majorana mass then there are additional parameters (and phases) to specify.

We'll now look at each of these classes of parameters in turn.

5.1.1 Unification

Perhaps the most important fact about all the constants of Nature is that they are very poorly named. They are not constant. Instead, the phenomenon of renormalisation means that the "constants" depend on the energy scale at which you do your experiment. We described this in some detail back in Section 2.3, and again in Section 3.1, when we explained how the fine structure constant, and its counterpart for the strong force, change with energy.

The energy dependence of coupling constants brings important clues when we come to better understand their origin. In particular, we understand well how the coupling constants vary on scales that we've tested – say, up to 10^3 GeV. But we could then extrapolate to further energies. Of course, we don't know what lies ahead at further energies, but to get the ball rolling we could simply assume that there's nothing other than the Standard Model, and then see what we find.

What we find is extremely interesting and shown in Figure 50. First let's explain what we're looking at. The forces of the Standard Model are summarised by $U(1)\times SU(2)\times SU(3)$. Corresponding to each of these is a coupling constant α_i where $i=1,2,3$. Note, in particular, that α_1 is the hypercharge coupling, rather than the fine structure constant of electromagnetism which emerges at low-energies from a combination of hypercharge and the weak force. On the horizontal axis is the energy, here called Q, plotted on a logarithmic scale. The part of the graph that we've measured experimentally is way over to the left, with $\log_{10}(Q/\text{GeV}) \lesssim 3$. Everything else is extrapolation.

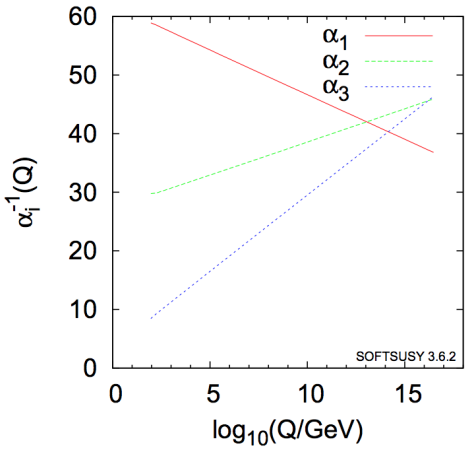


Figure 50. The running of the coupling constants. α_1 is hypercharge; α_2 the weak force and α_3 the strong force. This plot was made by Ben Allanach and taken from the PDG review of GUTs.

On the vertical axis of Figure 50 is the inverse coupling, α_i^{-1}. The red line decreases with energy, while both the blue and green lines increase. This is telling us that α_1 increases with energy, while both α_2 and α_3 decrease, and this is the expected behaviour given our discussion of asymptotic freedom in Section 3.1.

The most striking aspect of the plot is the way the three lines cluster together as we approach higher energies. Obviously, they don't precisely meet, but nonetheless they lie in the same ballpark. This would appear to be hinting that the three forces are not as different as they appear in our world: perhaps, at a much higher energy scale, they are all unified as one. This idea is know as *grand unification*. The weak and strong forces meet at a scale

$$M_{\text{GUT}} \approx 10^{16} \text{ GeV}$$

which is know as the grand unified scale. At this point, the three coupling constants all converge on a value somewhere around

$$\alpha_{\text{GUT}} \approx \frac{1}{40}$$

Even the location of this almost-meeting is important. First, α_{GUT} is nice and small at this point, telling us that the calculation to extrapolate the lines is at least consistent. Second, the scale M_{GUT} lies just below another important scale in nature, namely the

Planck scale

$$M_{\rm pl} \approx 10^{18}\ {\rm GeV}$$

(This is sometimes referred to as the reduced Planck scale to distinguish it from another contender that differs by a factor of $\sqrt{8\pi}$ and so is closer to 10^{19} GeV.) The Planck scale $M_{\rm pl}$ is where the effects of gravity become important in the quantum regime. We'll have more to say about this later. For now we just mention that the fact $M_{\rm GUT} < M_{\rm pl}$ is important. Had it turned out to be the other way round, then there would be no reason to think that $M_{\rm GUT}$ is interesting: it is a scale that was derived by neglecting the effect of gravity, and that's only an acceptable thing to do at energies less than $M_{\rm pl}$.

If we take the existence of three lines not-quite meeting as evidence of unification, what can we do about it? Is it possible to write down a grand unified theory, or GUT, in which the three forces are unified? The answer is yes. In fact, in many ways the Standard Model is just crying out to be packaged into something simpler! Recall that the mathematical way of describing the three forces is as

$$G_{\rm SM} = SU(3) \times SU(2) \times U(1)$$

where each of these can be roughly thought of as a electric and magnetic fields whose individual elements are themselves matrices: 3×3 for the strong force, 2×2 for the weak force and just usual numbers for $U(1)$ hypercharge. But all of these can be packaged nicely inside a bigger matrix. (Or, more precisely, a bigger group.) For example, you can put all of them inside a 5×5 matrix with

$$G_{\rm GUT} = SU(5)$$

Alternatively, they can be packaged into a 10×10 matrix, but where the matrix is now based on real numbers rather than complex numbers

$$G_{\rm GUT} = SO(10)$$

Other options are also available.

In all of these cases, there are new gauge bosons that come as part of the unified force. These are usually called *X-bosons*, in analogy with the W and Z bosons of the weak force. There are also new scalar fields that condense, giving rise to a Higgs-like mechanism. Unlike the Higgs mechanism of the weak force, this conjectural GUT-Higgs should get an expectation value at scale of around $M_{\rm GUT}$. Correspondingly, the X-bosons are heavy with a mass also somewhere in the vicinity of $M_{\rm GUT}$. Needless to say, we would not expect to discover such X-bosons in collider experiments any time soon.

It's not just the forces which have to unify. The matter particles and their other interactions must too. Here too things start off looking rosy. Recall that, including the count over colours, there are 16 fermions in one generation of the Standard Model. It must be possible to package these into the groups mentioned above that unify the forces. (Mathematically, we're looking for *representations* of $SU(5)$ or $SO(10)$ and these come only in special numbers, special like the 8 and 10 of the eightfold way are special.) It turns out that the particles of the Standard Model are tailor made to be put together in this way. It seems like everything fits like a glove.

Interestingly, for $SU(5)$ GUTs, the right-handed neutrino remains an outsider, not coupling directly to the forces. Meanwhile, for $SO(10)$ GUTs, the right-handed neutrino also is brought into the fold and, at least at the fundamental level, sits on the same footing as all the other particles.

So far, so good. The last part of grand unified theories is to find a way that the flavour sector drops out nicely, with the Yukawa terms and mixing matrices all falling into place. Here things are less rosy. It's possible, but it's not pretty, typically involving the introduction of yet further fields put together in a fairly baroque way. As with so many other things in the Standard Model, the flavour sector is the one we understand least.

Tweaking the Lines

The coupling constants in Figure 50 get close, but fail to actually meet. However, this plot was made under the assumption that there's nothing new to be found between the energy 10^3 GeV that we've probed experimentally and the GUT scale 10^{16} GeV. And that seems unlikely.

If there are new particles to be found, then they contribute to the renormalisation of coupling constants and so change the way the lines run. One obvious suggestion is that these new particles may correct the lines in such a way that they do, in fact, meet after all.

For this to happen, the new particles shouldn't be too heavy otherwise they come in too late to make a difference. One possibility that has been greatly studied is a theory called *supersymmetry*. We'll say more about this in Section 5.1.2, but for now we'll simply mention that we introduce a bunch of new particles at, say 5×10^3 GeV. This tweaks the running of the couplings, to give the result shown in Figure 51. And ... ta-da. The lines now meet perfectly.

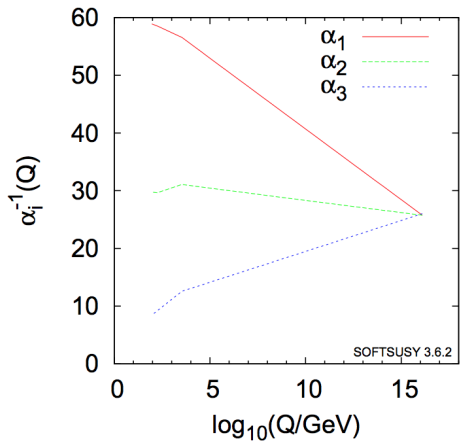


Figure 51. The running of the coupling constants if new supersymmetric particles exist at a low mass. This plot was made by Ben Allanach and taken from the PDG review of GUTs.

However, there's a major problem with this particular supersymmetric scenario. The new particles are so light that some hint of them should have shown up at the LHC by now. They haven't. It's possible to write down theories where supersymmetry still does the job of unification while just evading detection but they look increasingly unlikely.

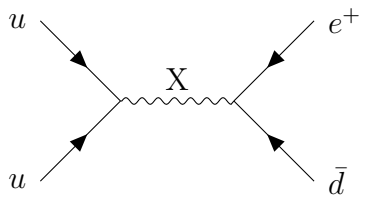
Nonetheless, the idea that there may be further particles out there which point clearly to unification is a tantalising one, and the convergence of the three coupling constants remains one of major clues we have about physics beyond the Standard Model.

Proton Decay

Although the grand unification scale $M_{\rm GUT}$ is way beyond what we can study experimentally, the idea of grand unification still gives an observable consequence. That is proton decay.

We already discussed proton decay in Section 4.3.5 in the context of the conservation laws. Recall that, within the Standard Model the proton should be absolutely stable. There is a rather exotic process — known as electroweak instantons — that allow three baryons to decay to three leptons but this has never been observed and is unlikely to be any time soon given that calculations give a lifetime of around 10^{173} for a helium nucleus!

However, proton decay is necessarily a consequence of any grand unified theory. This is because the X-bosons gives rise to a Feynman diagram like those of weak decays, but now linking quarks and leptons like so



where, in this example, the X-boson carries electric charge $+4/3$. The $\bar{d}$ quark then combines with the d quark in the proton to form a pion, hence $p \to e^+ + \pi^0$.

Because the mass of the X-boson is so large, the lifetime of the proton is long. But not all that long! Most GUTs predict a lifetime between 10^{31} and 10^{36} years. Current experimental bounds tell us that the proton lifetime is longer than 10^{34} years, already ruling out the simplest GUTs.

Magnetic Monopoles

All magnets are dipoles, with a north pole and a south pole. Cut a magnet in two, and you'll end up with two dipoles. It's impossible to get, say, a single north pole on its own. If one could find such an object – known as a *magnetic monopole* – it would have a distinctive radial magnetic field of the form

$$\mathbf{B} = \frac{g}{4\pi}\frac{\hat{\mathbf{r}}}{r^2} \tag{5.1}$$

where g is the magnetic charge.

At first glance, there seems to be little reason to think that magnetic monopoles exist. Indeed, there is even a law of physics that forbids monopoles! One of the Maxwell equations (2.3), governing the theory of electromagnetism, reads

$$\nabla \cdot \mathbf{B} = 0 \tag{5.2}$$

and its sole purpose is to disallow any solution of the form (5.1).

This makes it somewhat surprising that, as the laws of physics evolve beyond electromagnetism, magnetic monopoles re-emerge as one of the most likely candidates for new particles, finding interesting and creative ways to evade the seemingly insurmountable obstacle (5.2). By the time we get to grand unified theories, monopoles become obligatory, appearing as an entirely novel kind of particle known as a *soliton*. The mass of these magnetic monopoles is

$$M_{\text{monopole}} \approx \frac{M_{\text{GUT}}}{\alpha_{\text{GUT}}}$$

putting them somewhere in the range of 10^{17} GeV, well out of reach of current colliders.

Unlike all heavier particles that we've discussed in these lectures, monopoles would be completely stable. Ironically the same Maxwell equations which once seemed to forbid magnetic monopoles, now forbids them from decaying since they imply the conservation of magnetic charge. Monopoles can only vanish by annihilating with anti-monopoles. This leaves open the possibility that we may, once again, turn to the skies and search for monopoles among cosmic rays. So far, none have been found[18].

It's not just GUTs that give rise to magnetic monopoles. Instead, pretty much any theory that goes beyond the Standard Model will contain magnetic monopoles. They are one of the very few robust predictions for new physics. If you want to learn more about monopoles then you've come to the right place. You can read about their subtle interplay with quantum mechanics in the lectures on Solid State Physics, about their role in quantum field theory in the lectures on Gauge Theory, and about some of their more mathematical aspects in the lectures on Solitons.

5.1.2 The Higgs Potential

Our next pair of parameters are associated to the Higgs potential. Recall from Section 4.2.1 that the Higgs potential $V(\phi)$ determines whether the Higgs boson condenses. In the Standard Model, it takes the very simple form

$$V(\phi) = a|\phi|^2 + b|\phi|^4 \tag{5.3}$$

The two fundamental parameters are a and b.

If $a > 0$ and $b > 0$ then it looks like the graph on the left-hand side of Figure 52. If $a < 0$ and $b > 0$ then it looks like the right-hand side of Figure 52. (If both a and b are less than zero then the potential has no minimum the Higgs scalar runs away to infinity unless we include further $|\phi|^6$ terms of higher.)

As we've seen, the Standard Model has a potential with the shape on the right, meaning that $a < 0$ and $b > 0$. The values of a and b determine both the mass of the Higgs boson m_H and the Higgs expectation value $\langle\phi\rangle$. Roughly speaking, the relationship between these two scales and the parameters in the potential is

$$m_H^2 = |a| \quad \text{and} \quad \langle\phi\rangle^2 = \frac{a}{2b} \tag{5.4}$$

[18] A more correct statement is that exactly one has been found! On Valentine's day, 1982, a single event consistent with a magnetic monopole was observed. Nothing similar has been seen since. Given the importance of replicating scientific results, it's difficult to view this as anything more than a tantalising footnote (literally here) in the story of the monopole

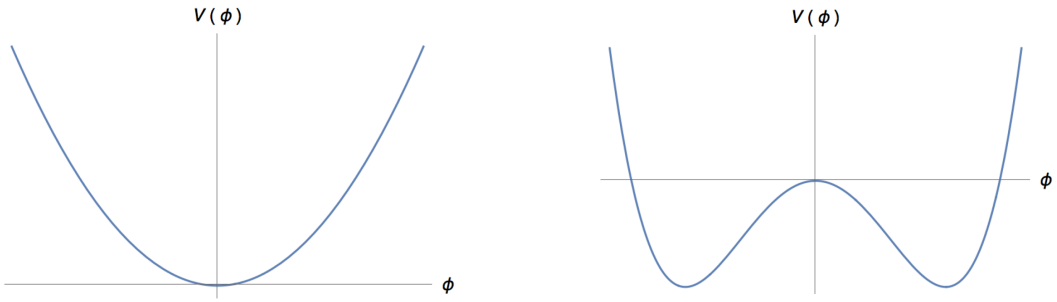


Figure 52. Two possible shapes for the Higgs potential for a scalar field. With $a, b < 0$ we get the shape on the left; with $a < 0$ and $b > 0$ we get the shape on the right.

Much of this section will be devoted to explaining further what the words "roughly speaking" in the previous sentence actually mean. For now, we'll just roll with the equations above and see what they tell us. Experimentally, we know that

$$m_H \approx 125 \text{ GeV} \quad \text{and} \quad \langle\phi\rangle \approx 246 \text{ GeV}$$

This means that the parameters in the potential must take the values

$$a \approx -(125)^2 \text{ GeV}^2 \quad \text{and} \quad b \approx 0.13$$

For now, the important point to note is that a has dimension of energy-squared, while b is dimensionless. Our interest here lies in a. This is the scale of the Higgs sector which, through Yukawa interactions subsequently sets the mass of all the elementary fermions.

As we've stressed earlier in these lectures, most of the mass of the proton and other hadrons doesn't come from the Higgs sector, but instead from $\Lambda_{\rm QCD} \approx 200$ MeV. But this sits on a different footing because $\Lambda_{\rm QCD}$ was, itself, a derived scale: it is the energy at which the dimensionless coupling of the strong force becomes $\alpha_s \approx 1$. This means that $\Lambda_{\rm QCD}$ should be thought of as an emergent energy scale.

In contrast, there is no such story for $\sqrt{|a|} = m_H$. This is an absolute energy scale. In fact, rather remarkably, it is the only fundamental parameter of the Standard Model that is not dimensionless! As we now explain, this means that it comes with certain baggage.

The Higgs Mass

Now we come to the crux of the matter, one that revolves around the "rough speaking" that lead us to the equations (5.4). To understand what the issue is, we need to think a little more deeply about the kind of parameters that make sense in a fundamental theory and how these change at various energy scales due to renormalisation.

To set the scene, I'll introduce some analogies. To this end, here are three objects that should not be viewed as a "fundamental parameters" in a theory.

There sits, in a vault in Paris, a platinum-iridium cylinder that, until 2019, was used as the definition of the kilogram. Clearly it would be ludicrous to take the mass of this object as any kind of input in a fundamental theory of Nature. Even putting aside the facts that the kilogram is very much a human construct, and that the mass of the cylinder is changing over time as it slowly erodes, it is simply silly to think that a huge, complicated object with roughly 10^{24} constituent particles should be important on the tiny distance scales at which our fundamental theory of physics is defined.

This latter criticism can also be levelled at other objects which, at first glance, might appear to be more suitable candidates for fundamental parameters. For example, for many decades the mass of the proton was thought to be a fundamental scale in nature. Of course, we now know that the proton, like the Parisian cylinder, is a horribly complicated object. In particular, the mass of the three valence quarks — two up and a down — contribute a negligible amount to the total mass of the proton. The full mass is often attributed to collection of gluons and sea of quark-anti-quark pairs but, for the purposes of this analogy, we will be better served if we remember what this really means: the mass comes from the wild thrashing of the quantum fields that are excited in complicated ways inside the proton. The mass of the proton is in the ballpark of a few times $\Lambda_{\rm QCD}$ but its exact value is something emergent that depends on lots of messy dynamical processes. For this reason, the mass of the proton is, like the Parisian cylinder, an emergent scale. Neither are good candidates for parameters in a fundamental theory.

And this brings us to our third example of an object which is *not* a good candidate for a fundamental parameter. This is the mass of the Higgs boson!

Quantum fluctuations mean that, even though the Higgs boson is a fundamental particle, its mass depends on all sorts of complicated and messy dynamics and is ultimately determined, like the mass of the proton, by the behaviour of other quantum fields. In particular, if the fundamental theory has a parameter like a in the potential

(5.3), then its relationship to the mass m_H of the Higgs boson is nothing like as simple as the $m_H^2 = |a|$ equation that we used above. Instead the relationship between the two is much more complicated.

To understand what's going on, first recall our discussion of renormalisation in Section 2.3. There we learned that all parameters in a quantum field theory depend on the distance scale or, equivalently, the energy scale at which an experiment takes place. Moreover, we stressed that all quantum field theories should come with a health warning: there is a minimum distance scale, or maximum energy scale, beyond which they shouldn't be used. This energy scale is called the *UV cut-off* and we will denote it as Λ_{UV}.

You should think of the UV cut-off as the energy scale at which a given quantum field theory is defined. When you specify the parameters of the theory, you should specify their value at the scale of the cut-off. This is really just the statement of reductionism in physics: small things determine the behaviour of larger things, and a fundamental theory should be defined on the smallest distance scale at which it applies. As you then look at lower energies – or longer distances – you can use renormalisation to figure out how these parameters change.

Now let's return to the Standard Model. It definitely works up to energy scales of 1 TeV so let's be pessimistic and, with the expectation that the Standard Model will cease to give the right answers very soon, take the cut-off to be $\Lambda_{UV} = 1$ TeV. That means that we take the Standard Model to correctly describe the dynamics of quantum fields down to distance scales as small as 10^{-19} m. The fields may well have fluctuations on scales much smaller than that, but we will just admit ignorance about these and proceed.

What now happens if we put a Higgs boson into the mix. Naively the mass of the Higgs boson is given by the formula that we used previously: $m_H = \sqrt{|a|}$. But the Higgs then gets surrounded by a swarm of quantum fluctuations and these change the mass. The upshot is that if a is the fundamental parameter of the theory, then the mass of the Higgs boson that we measure is actually something like

$$m_H^2 \approx \left|a + \mathcal{O}(\Lambda_{\rm UV}^2)\right| \tag{5.5}$$

where the $\mathcal{O}(\Lambda_{\rm UV}^2)$ means a contribution that is roughly around $\Lambda_{\rm UV}^2$, but where the exact coefficient (including its sign) depends on the nature and dynamics of all the other fields.

Here's a rather nice analogy. Take a ping pong ball and submerse it in water. What's its mass? According to the International Table Tennis Federation, a ping pong ball should have a mass of exactly 2.7 grams. But the federation rarely promotes matches that take place underwater and there is no suggestion on their website that they understand renormalisation. You can use Newton's equation $F = ma$ to experimentally determine the mass of a ping pong ball in water: just apply a force and measure the acceleration. You'll find that the ping pong ball appears to be roughly 11 times heavier than in air, so closer to 30 g. The reason for this is intuitively simple: when the a ping pong ball moves in water, it drags a body of fluid with it. This increases its inertial mass.

This same effect is at play for the Higgs boson. Its mass gets a contribution from the fluctuations of all other quantum fields. This additional contribution is of order $\Lambda_{\rm UV}$ which is much larger than the mass that we actually observe.

What are the consequences of this? The formula (5.5) tells us that the fundamental parameter a in the Higgs potential is not equal to the mass of the Higgs boson. Instead, it too must be something on the scale of the UV cut-off $\Lambda_{\rm UV}$. We call the parameter a the *bare mass* (squared), while m_H is physical mass of the Higgs boson that we observe. To get the observed physical mass $m_H \approx 125$ GeV we must add two contributions, a and $\Lambda_{\rm UV}$, both of the same order of magnitude, which then cancel out to leave behind the physical mass.

So what? Well, if we take the cut-off of the Standard Model to be $\Lambda_{\rm UV} \approx 10^3$ GeV, then this seems eminently reasonable. We're just adding two numbers, both of order 1000, to get something or order 100.

But what if the Standard Model holds to higher energy scales? Suppose that it is trustworthy to energies of order $\Lambda_{UV} \approx 10^4$ GeV. Now we add two numbers of order 10,000 to get something left behind that's order 100. Still not preposterous, but you might start to feel a little uneasy.

And what if we really push things? Suppose that the Standard Model actually holds all the way up to the GUT scale of $\Lambda_{UV} \approx M_{\rm GUT} \approx 10^{16}$ GeV. Now we must take the fundamental parameter in the Higgs potential to be $a \approx (10^{16}\ {\rm GeV})^2$ so that this cancels almost precisely the contribution from the quantum fluctuations, leaving behind a measly Higgs boson mass of 125 GeV. If you were to change a by just one part in a billion, you would end up with a Higgs boson mass that is 5 orders of magnitude higher than we observe! By this stage, the situation appears to be ridiculously untenable

We have a bunch of different names for this state of affairs. We say that the parameter a in the Higgs potential should be *finely tuned*. It's a good name and describes the issue well. Alternatively, we say that the Higgs mass m_H that we observe is *unnatural*; the quantum fluctuations want to push the mass to higher and higher energy scales, but this is cancelled — to what may feel like unreasonable accuracy — by the contribution from the bare mass. This seems to be a less useful name as it's hard to see how something in nature can be unnatural. But it does at least stress how jarring the situation is.

This whole set of ideas also goes under the umbrella of the *hierarchy problem*. Why is there a large hierarchy of scales between the Higgs mass and the UV cut-off, when theory suggests that they should be the same order of magnitude?

Recall that the UV cut-off $\Lambda_{\rm UV}$ is really an expression of our ignorance: we confess that our quantum field theory — in this case the Standard Model — is incomplete and does not include relevant physics at energies beyond $\Lambda_{\rm UV}$. The hierarchy problem is really the statement that the mass of the Higgs boson is pushed by quantum corrections to whatever is the highest mass scale in the game. Yet, somehow, the observed mass remains happily at 125 GeV.

The hierarchy problem in this brutal form is only a problem for the parameter a in the Higgs potential. All other parameters in the Standard Model are dimensionless and do not suffer the same fate. They change only very mildly (logarithmically to be precise) under renormalisation. It is only the Higgs mass that is so very sensitive to physics at higher energy scales.

Solutions to the Hierarchy Problem

The hierarchy problem motivated an enormous amount of research in the 1980's, 90's and early 2000's. The favoured explanation was the obvious one: keep the UV cut-off $\Lambda_{\rm UV}$ low enough that you don't need to invoke a silly level of fine-tuning. But this, in turn, means that there should be some new physics, invalidating the Standard Model, that comes in at some low scale, like a few TeV.

This new physics can't just be anything. Add a few new particles at the TeV scale and you'll see that they just make the problem worse, adding yet more quantum fluctuations that increase the mass of the Higgs boson. Instead, you must find a way to add some new fields to the Standard Model that stabilise the mass of the Higgs somewhere in near 100 GeV, solving the problem once and for all. There are a number of ways to achieve this. Here I describe some, roughly in descending order of popularity.

- **Supersymmetry:** This is a proposed, novel symmetry of Nature in which every bosonic field has a fermionic counterpart, and vice versa, with each boson/fermion pair experiencing the same forces.

  Clearly, we don't see supersymmetry among the particles that we know. The idea is that supersymmetry is, like many symmetries, broken so that the additional fields – so called superpartners – only appear when we reach some new mass scale, say a few TeV.

  There are a number of reasons to be enamoured of supersymmetry. First, it solves the hierarchy problem by dint of the fact that bosons and fermions contribute with opposite signs to the mass of the Higgs, ensuring that all quantum fluctuations cancel above the supersymmetry scale. Second, as we saw previously, the presence of supersymmetry causes the three coupling constants to meet perfectly at the unification scale. Moreover, there are reasons to think that supersymmetry may be an important ingredient in quantum gravity, which would mean that it should certainly be present by the time we get to the Planck scale. All in all, TeV scale supersymmetry leads to a nice, comforting story.

- **Technicolour:** The Standard Model contains just one other mass scale, $\Lambda_{\rm QCD}$. But, as we described above, this is associated to the strong coupling dynamics of QCD and so doesn't suffer from a hierarchy problem. Perhaps the mass of the Higgs emerges in a similar way.

  In such scenarios, the Higgs particle is not fundamental at all, but rather appears as a composite of two, new fermions, bound together into a meson by a new force called, in analogy with strong force, *technicolour*.

- **Something Else:** There are quite alternative proposals. Among them is the idea that Higgs as a (pseudo)-Nambu-Goldstone boson (I won't explain what this means!) or, more creative suggestions, such as extra dimensions of space which manage to dilute quantum corrections at the TeV scale.

Each of these ideas provides a viable solution to the hierarchy problem, but only by introducing some observable deviation from the Standard Model at the TeV scale. But now we have a collider – the LHC – that can reach these scales. And nothing is seen. The Higgs boson appears, as far as we can tell, as a genuinely elementary particle and there is, as yet, no hint of new particles that could stabilise its mass. This makes it increasingly unlikely that any of the "natural" solutions described above are implemented in Nature. Of course, it's certainly possible that, say, supersymmetry is still out there, but just pushed up to a higher scale, with an accompanying need

for some amount of fine tuning to the Higgs mass. But the motivation starts to look increasingly shaky.

So what are we to make of this? Admittedly, the hierarchy problem has a different flavour from other major open problems in physics. Is there really anything wrong with just stating that the parameter a is (to give an extreme example) defined at the GUT scale $\Lambda_{\rm UV} = M_{\rm GUT}$ and fine tuned to 15 significant figures so that it perfectly cancels out the contributions to the Higgs mass from quantum fluctuations of order $\Lambda_{\rm UV}$? It certainly seems odd, but perhaps that's just the way it is.

We can look elsewhere in physics to get some guidance. In particular, quantum field theory isn't just useful for particle physics: it is the right language to describe large swathes of solid state physics (also known as condensed matter physics), which is the study of how various materials behave. This arena provides many hundreds, if not thousands, of examples of quantum field theories (or, relatedly, statistical field theories) where we can test our logic. In that framework, we can ask: do we find quantum materials where we have light scalar excitations, like the Higgs boson? Here "light" means with a mass significantly smaller than the UV cut-off which, in solid state physics, is usually supplied by the underlying lattice of the material.

The answer to this question is a resounding no! Or, stated more accurately, within the realm of solid state physics if there is a light scalar excitation then there is always a good reason behind it. These reasons are often similar to the ones invoked for the hierarchy problem, like the scalar is really made of two underling fermions, or it is a (pseudo)-Nambu-Goldstone boson. (I still won't explain what this last phrase means.) The upshot is that in other realms where quantum field theory is useful, the logic of *naturalness* is a very good guide: if you see a scalar field that is unnaturally light, then you should search for an explanation because it will tell you something important and interesting about the system.

At the risk of belabouring this point with a long detour, there is one particular condensed matter system that is worth looking at more closely. This is superconductivity which, as we already mentioned in Section 4.2.1, has a mathematical description that is almost identical to that of the Higgs boson. Usefully, the hierarchy problem also makes an appearance in superconductors, although it's not in the guise of a light scalar field but, as I now explain, something more subtle. As you cool a metal, it undergoes a phase transition to become a superconductor. This happens at the *critical temperature* T_c which is typically a few degrees Kelvin. The phase transition happens discontinuously, meaning that the metal changes abruptly to a superconductor just like water

changes abruptly to ice. This is sometimes called a *first order phase transition*. But theoretical expectations tell us that the phase transition should be smooth and continuous, what's called a *second order phase transition*. This isn't seen in experiment because it turns out that you have to tune the temperature T to ridiculous accuracy before you notice that the transition is actually continuous, something like

$$\frac{T - T_c}{T_c} \approx 10^{-9}$$

Why would you have to tune the temperature so close to T_c before you can see the real physics of the second order phase transition? It's a very unusual state of affairs and in most other systems you only need $(T - T_c)/T_c \approx 1$ before you can see the true nature of the phase transition. The fine-tuning of temperature needed for a superconductor turns out to be entirely analogous to that of the lightness of the Higgs boson. The 10^{-9} accuracy is a small number and deserves an explanation. And, in this case, there is a very good explanation: the would-be Higgs boson in a superconductor is composed of two electrons bound together at an energy scale which emerges mathematically in a manner that is similar to way $\Lambda_{\rm QCD}$ emerges in the strong force. You can trace the existence of the small number 10^{-9} to this underlying dynamics. It's an explanation that is close in spirit to the technicolour proposal for the Higgs boson. But, as far as we can tell, the same story is not what's going on for the Higgs. (If you want more details about the relationship between phase transitions and renormalisation group, you can read about it in the lectures on Statistical Field Theory.)

All of this is to say that our experience with quantum field theory tells us that we should be nervous about the fine-tuning underlying the Higgs. Still, there is a vast difference between meV scale at which effective quantum field theories in condensed matter are valid, and the TeV scale or higher of particle physics. And Nature seems to be telling us very clearly that there is nothing wrong with fine tuning at these very high energy scales, even though it flies in the face of how we understand quantum field theory. It seems to me, and many others, that this is a problem one should take seriously because it tells us that we're not thinking about quantum field theory in the right way. But I don't know a better one!

There is one last word on this. The word in anthropic. It gives me a slight shudder just thinking about it so I'm going to the postpone further discussion to Section 5.3.1.

5.1.3 A Bit of Flavour and Strong CP

In the late 1800's, one of the great unsolved mysteries was the spectrum of hydrogen. Only a few lines were known when, in 1885, Balmer noted that their wavelength λ

could be fitted to the remarkably simple formula

$$\lambda \approx 10^{-7} \left(\frac{1}{4} - \frac{1}{n_2^2} \right)^{-1} \text{ m}$$

with $n_2 = 3, 4, 5, 6$. Three years later, Rydberg generalised this to the formula

$$\lambda \approx 10^{-7} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)^{-1} \text{ m}$$

with n_1 and n_2 both integers. (The simplest $n_1 = 1$ lines lie in the ultra-violet and took another 20 years to discover.)

The Balmer formula provides one of the rare examples in science where fitting data, with no understanding of the underlying physics, gives a strong hint of something important underneath. Indeed, the presence of integers in Balmer's formula foreshadowed one of the greatest paradigm shifts in science: the need for quantum mechanics. Ironically, Bohr would later claim that he had completely forgotten about the Balmer formula when he constructed his quantum model of the atom, and only later realised that the two matched perfectly!

If you were to wonder what collection of today's unexplained experimental data has the best chance of a Balmer-like breakthrough, the answer is obvious: it is everything to do with flavour. There are a few dozen parameters that describe the masses and mixing angles of three generations of quarks and leptons. There are clear patterns among them but, so far, we have little idea what those patterns are telling us.

It is not for want of trying. There are many theories that have tried to explain what's observed, many of them revolving around some kind of new, approximate symmetry, sometimes referred to as family symmetries. Some of these theories are pretty(ish), some baroque. None of them are overwhelmingly compelling, in large part because there are so many options available. For this reason, I won't describe any of these in detail.

The Strong CP Problem

There is, however, one parameter in the Standard Model that deserves special attention. It is a parameter of the strong force known as the QCD theta angle, or θ_{QCD}.

It is not so straightforward to describe, at the level of these lectures, how θ_{QCD} changes the dynamics of the strong force, not least because it is a quantum parameter, in the sense that it has no classical counterpart at all. If you want all the gory details, you can read about theta angles in the lectures on Gauge Theory.

The main reason that the QCD theta angle is interesting that it doesn't exist! As far as we can tell, $\theta_{\rm QCD} = 0$. Strictly our best experimental bound is

$$\theta_{\rm QCD} < 10^{-10}$$

There are a few questions to unpack here.

First, why should we care about something that doesn't exist?! The point is that the Standard Model is a remarkably restrictive framework. We can't just write down random interactions willy nilly. There are only very special interactions that are allowed, and each of them comes with a parameter. The game that we play is to write down all possible interactions and then determine their associated parameter by experiment. This is what leads us to the couple of dozen parameters or so listed at the beginning of this section. The theta angle is special because it is the only one of these parameters that appears to vanish. And that is crying out for an explanation.

Next: what would the consequences be if $\theta_{\rm QCD}$ were not to vanish? It turns out that this is the one opportunity for the strong force to violate the symmetries of parity P, charge conjugation C and time reversal T.

Recall from Section 4.3.4 that weak force respects neither parity nor CP but, as far as we can tell, both are respected by the strong force and electromagnetism. In a world where $\theta_{\rm QCD} \neq 0$, the strong force would also break these symmetries. It does so by endowing the neutron with an *electric dipole moment*. This means that although the neutron is neutral, the distribution of the quarks inside would be shifted slightly so that one side of the neutron is slightly positively charged with a compensating negative charge on the other side. This means that the neutron carries a little arrow, pointing in the direction of the charge imbalance. This is in addition to another arrow, pointing in the direction in which the neutron spins and it turns out that these two arrows are either aligned or anti-aligned. When one acts with parity, or charge conjugation, or time reversal, one of the arrows flips, while the other does not. This means that these symmetries are broken since, for example, the neutron appears different when viewed in a mirror.

Although all three symmetries are broken when $\theta_{\rm QCD} \neq 0$, physicists tend to focus on CP. The unexplained fact that $\theta_{\rm QCD} = 0$ is known as the *strong CP problem*.

Before we turn to putative explanations for this problem, there is one final question I'd like to address. What does it have to do with flavour physics? The question of flavour is all about how fermions couple to the Higgs boson, while the strong CP

problem appears to be, as the name suggests, firmly in the camp of QCD. Here's where things get a little complicated because the story that I told above isn't entirely accurate.

The fuller truth is that in the Standard Model the masses of the six quarks are actually complex numbers rather than real numbers. Each has a magnitude and a phase

$$M_q = m_q e^{i\chi_q}$$

where the label q runs over the six different quarks. What we observe is the magnitude m_q. The complex phases χ_q have almost no effect on the physics. They do just one thing: they contribute to the QCD theta angle! In fact, what we measure is not the QCD theta angle directly, but rather

$$\bar{\theta} = \theta_{\rm QCD} + \sum_q \chi_q$$

It is this sum of angles that is observed experimentally to vanish: $\bar{\theta} = 0$. In other words, the strong CP problem involves both QCD and flavour physics, entwined in some interesting fashion.

There are a number of prospective solutions to the strong CP problem. Here I'll just mention one, which goes by the name of the Peccei-Quinn theory. The key idea is both simple and dramatic: one takes the parameter $\theta_{\rm QCD}$ and promotes it to a dynamical field.

This means that we no longer get to chose the value of $\theta_{\rm QCD}$. Instead its value can fluctuate. This is only progress if we can explain why we don't observe the fluctuations and why, moreover, the field prefers to sit at the value such that $\bar{\theta} = 0$.

The new field $\theta_{\rm QCD}$ is, like the Higgs boson, a scalar field and its preferred value will be set by some potential, analogous to the $V(\phi)$ potential that we saw before. Here there is a very nice story. It turns out that the rest of the Standard Model fields have something to say about this potential. In particular, the dynamics of the gluons generates a potential for $\theta_{\rm QCD}$. And this potential has the property that its minimum sits at the place where $\bar{\theta} = \theta_{\rm QCD} + \sum_q \chi_q = 0$. In other words, if $\theta_{\rm QCD}$ was dynamical, it would want to arrange itself so that there is no CP violation in the strong force. This, it would seem, is a very compelling solution to the strong CP problem.

What about the second issue? Now that $\theta_{\rm QCD}$ is dynamical, why do we not see its fluctuations? Here things are less rosy. First, it's not clear that the simple dynamical mechanism described above is really sufficient to set $\bar{\theta} = 0$ to one part in 10^{10} as required by experiment. Second, with a new field comes a new particle which, in the case of $\theta_{\rm QCD}$ is referred to as the *axion*. There have been many decades of searches for the axion, with nothing yet seen. There are many further experiments underway, aiming to improve the bounds the axion mass and interaction strength or, in the dream scenario, actually find the thing!

While the Peccei-Quinn theory, and the accompanying axion, remains the most popular explanation for the strong CP problem, the lack of clear experimental support means that it is not the only game in town. Whatever the ultimate reason, the fact that the strong force prefers to respect the discrete symmetries of our world, even though it has a clear opportunity to violate them, is one of the key clues for physics beyond the Standard Model.

5.2 Gravity

There is one force that is obviously missing from the Standard Model. This is gravity, both the most obvious force at play in the world around us and yet, in many ways, the one we understand least.

There are two goods reasons why gravity is not included in the Standard Model. The first is that the force of gravity is entirely inconsequential for anything to do with particle physics. To get some sense for this, we can look at the theory of gravity first written down by Newton. Any two objects, with masses m_1 and m_2, sitting a distance r apart, will feel a force

$$F_{\rm Newton} = \frac{Gm_1m_2}{r^2} \tag{5.6}$$

where G is Newton's gravitational constant

$$G \approx 6.67 \times 10^{-11}\ {\rm m^3\,kg^{-1}\,s^{-2}} \tag{5.7}$$

It is a remarkable fact that this takes exactly the same form as the Coulomb force between any two electrically charged objects that we met earlier in (2.4),

$$F_{\rm Coulomb} = \frac{Q_1Q_2}{4\pi\epsilon_0 r^2}$$

This gives us an opportunity to compare the strengths of gravity and electromagnetism on the subatomic scale. For example, a hydrogen atom consists of a single electron

orbiting a proton, held in place by the Coulomb force. The size of the hydrogen atom is about 5×10^{-11} m, known as the *Bohr radius*. But there should also be atoms held together by the gravitational force. We could consider an electron orbiting a single neutron, held in place by Newton's force (5.6). The question is: how big would this gravitationally bound atom be?

The answer is pretty stunning: a gravitationally bound atom would have a size that is substantially larger than our observable universe. Gravity isn't just a little weaker than the other forces. It's much much weaker than the other forces. Another way of saying this is to take two electrons and compare their gravitational attraction to their Coulomb repulsion. The answer is

$$\frac{F_{\rm Newton}}{F_{\rm Coulomb}} \approx 10^{-43}$$

No one cares about the gravitational force acting on a single elementary particle.

However, gravity has a trick up its sleeve. All the other forces of nature have both positive and negative charges which cause particles to attract into effectively neutral objects on small distance scales. This happens most dramatically for the strong force, where quarks are confined into protons and neutrons. It then happens again for electromagnetism where electrons and protons are bound into neutral atoms. This means that by the time you get to the macroscopic world in which we live, these forces have done their job and their effects are no longer manifest. In contrast, there is no negative mass and so nothing to shield us from the effect of gravitation. As you pile more and more particles together, the Coulomb force becomes diluted, while the gravitational force only grows. This is why gravity is the force that seems to dominate our lives.

We do have a good understanding of why gravity is special in this way. The three forces in the Standard Model are associated to spin 1 fields. (The Higgs force, of course, is associated to a spin 0 field.) Gravity is unique because it is associated to a spin 2 field. And in contrast to all other forces, spin 2 fields can only have "positive charge", where the charge is what we call mass.

Spin 2 fields are, it turns out, special in many ways. While we could conceive of theories with many different spin 1 forces, that's not possible for spin 2 fields. There is just one way to introduce a fundamental spin 2 field into the laws of physics and it is utterly remarkable. The spin 2 field, it turns out, must be spacetime itself!

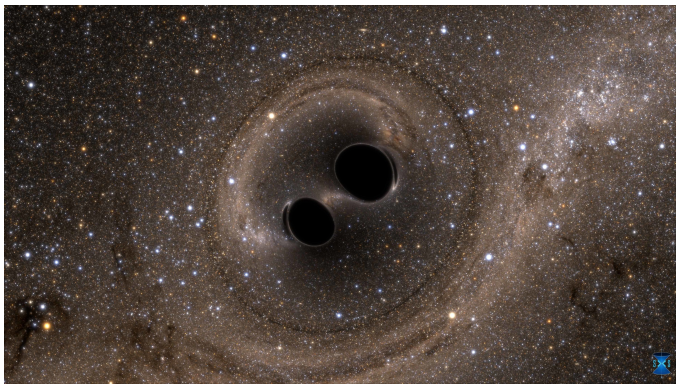
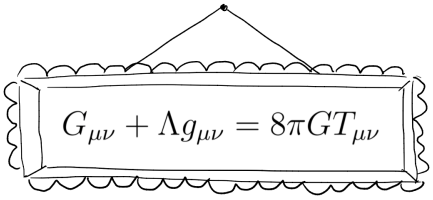


Figure 53. A computer simulation of two black holes, about to collide. The background field of stars is distorted by the curvature of spacetime in the vicinity of the black holes.

The connection between space, time and gravity is Einstein's great insight and the resulting theory goes by the name of General Relativity. It would take us too far from our main narrative to describe general relativity in any great detail, but if any equation deserves being placed in a picture frame, it is Einstein's:



This equation replaces Newton's gravitational force law (5.6). It relates the curvature of spacetime, captured in the $G_{\mu\nu}$ and $g_{\mu\nu}$ on the left-hand side, to the energy distribution of other fields, captured in the $T_{\mu\nu}$ on the right-hand side. There are two gravitational constants in the Einstein equations. The first is Newton's constant G, which retains its place a fundamental constant of nature in general relativity. The second, Λ, is known as the *cosmological constant*. We will have more to say about this in Section 5.3.1.

The Einstein equations replicate the familiar results of Newton's force law, from apples falling from trees, to the orbits of the planets around the sun. But they do so much more besides. They tell us how light bends as it passes heavy objects, giving rise to the distortion of the background field of stars, how black holes form as the density of matter becomes too great, and how collisions of these black holes can cause detectable ripples of the spacetime continuum known as gravitational waves. Furthermore, the Einstein equations provide, for the first time, a framework in which we can think about

the dynamics of the entire universe, giving rise to the field of Cosmology. For any physics on the very largest scales, the Einstein equations hold sway.

5.2.1 Quantum Gravity

Above I said that there are two reasons that gravity is not included in the Standard Model. The first is that gravity is inconsequential for particle physics. The second is that we don't really know how to do it.

General relativity is a classical theory. It does not incorporate the effects of quantum mechanics. In contrast, everything else that we've described in these lectures is all about quantum mechanics. The question is: can we incorporate general relativity into the quantum world? This is the problem of *quantum gravity*.

It's usually stated that naive attempts to combine quantum mechanics and general relativity fail miserably. That's not quite true. Naive attempts to combine the two theories actually work very well. Up to a point. For example, any simple minded approach to quantum gravity clearly shows that, on some level, spacetime acts just like any other quantum field. Small fluctuations of spacetime are tied up into little knots by the framework of quantum mechanics, resulting in a new kind of massless particle known as a *graviton*. Just as light waves are made of many underlying photons, so too are gravitational waves made of gravitons.

We've never detected an individual graviton. Indeed, because gravity is so very weak we only succeeded in detecting gravitational waves in 2015. Building an experiment to see individual gravitons is not technologically feasible in the forseeable future. To put this in perspective, we first detected light when eyes evolved. Fossil records clearly show the existence of eyes in the Ediacaran period and from then it took something like 550 million years before we were able to detect the underlying photons. I'm not promising that it will take us a similar timescale to detect gravitons, but it's as good a guess as any.

Although it's something of an academic exercise, we can also compute Feynman diagrams for gravitons to see they interact with each other under the rules of quantum mechanics. Again, all is good, but now there's a catch. The calculations make sense provided that the energies of the gravitons are not too large.

We can actually anticipate the problem from the few facts that we know about gravity. The strength of the gravitational force is governed by Newton's constant (5.7). This is what plays the role of the fine structure constant in electromagnetism or analogous quantities for the strong and weak force. Except there's a crucial difference:

the strengths of all the other forces are governed by dimensionless numbers. In contrast, there's no way to get a dimensionless number out of Newton's constant. Instead, if we include some factors of $\hbar$ and c, we get a scale:

$$G = \frac{\hbar c}{8\pi M_{\rm pl}^2}$$

where $M_{\rm pl}$ is the *Planck mass*,

$$M_{\rm pl} \approx 10^{18}\ {\rm GeV}$$

This is telling us that the strength of the gravitational force depends strongly on the energy of the process. If two particles scatter with energy E, then gravitational effects will scale as

$$\text{Strength of Gravity} \sim \left(\frac{E}{M_{\rm pl}}\right)^2 \sim GE^2$$

In some sense, you knew this already. Newton's equation (5.6) tells us that the gravitational force between two objects scales as Gm_1m_2 which is just the formula above with the masses in place of the energy. This formula also gives another perspective on why gravitational forces are so weak in the world of particle physics, where our most powerful accelerators can reach energies of $E \sim 10^3$ GeV, many orders of magnitude below the Planck scale.

This argument means that any naive theory of quantum gravity will give sensible answers providing that we look at energies $E \ll M_{\rm pl}$, where we can use Feynman diagrams and simple perturbation theory. But, as the energy increases to somewhere near the Planck scale, the gravitational interaction becomes strong and our Feynman diagram expansion ceases to work. In this simple minded approach, quantum gravity only becomes challenging when we reach energies close to the Planck scale. In technical, and slightly old-fashioned, terms general relativity is said to be *non-renormalisable*.

Before we proceed, it's worth thinking a little more about the size of the Planck scale. Back at the beginning of these lectures, I commented that the SI unit for energy, the Joule, is not particularly useful in the subatomic realm. Instead we define the much smaller unit of an electronvolt. But, by the time we get to the Planck scale, even the Joule is too small! We have

$$M_{\rm pl} \approx 10^9\ {\rm J}$$

That's a seriously large amount of energy. It is, for example, greater than the kinetic energy of all the cars in a formula one race! In particle physics, when we talk about

reaching a certain energy scale we really mean energy density. Physics at the Planck scale involves energies of $M_{\rm pl}$ squeezed into a region of size $L_{\rm pl}^3$, where $L_{\rm pl} \approx 10^{-34}$ m. These are silly numbers. There is, as we will see below, good reason to believe this is the highest energy density allowed by the laws of physics.

An Analogy: Fermi's Theory of the Weak Force

We've encountered a situation very similar to gravity elsewhere in these lectures. Recall from Section 4.2.3 that Fermi's original theory of the weak force included a dimensionful coupling known as the Fermi constant (4.4), whose value is roughly

$$G_F \approx \frac{\alpha_W}{M_W^2} \tag{5.8}$$

where $\alpha_W \approx 1/30$ while $M_W \approx 80$ GeV is the W-boson mass. If you don't know about W- and Z-bosons, and work purely within the Fermi theory then everything is fine provided that you only look at low energy processes. But, from this perspective, the strength of the weak force depends on energy as

$$\text{Strength of Weak Force} \sim \alpha_W \left(\frac{E}{M_W}\right)^2 \sim G_F E^2$$

As you approach energies closer to M_W then Fermi's theory stops working and must be replaced by something else. That something else is, of course, the W- and Z-bosons.

If we take this lesson seriously, it suggests a similar fate for gravity, with general relativity the low-energy approximation to something more fundamental. If we were to push the analogy with (5.8) yet further, we might expect that Newton's constant arises from a combination of a dimensionless coupling $\alpha_{\rm grav}$ and a fundamental scale $M_{\rm grav}$, so that

$$G_N \approx \frac{\alpha_{\rm grav}}{M_{\rm grav}^2}$$

with general relativity breaking down as we approach energies $M_{\rm grav}$. In such a scenario, we would see some new physics emerging at the energy scale $M_{\rm grav}$, possibly manifesting itself as new particles. Note that, if this is the way things pan out, then we can't be sure about the scale $M_{\rm grav}$ because it's related to $M_{\rm pl}$ by a dimensionless coupling $\alpha_{\rm grav}$. Assuming that $\alpha_{\rm grav}$ isn't ridiculously small, it would appear natural that $M_{\rm grav}$ is somewhere in the ballpark of the Planck scale $M_{\rm pl}$.

The discussion above is, appropriately for these lectures, very much a particle physicist's perspective on quantum gravity, focussing on scattering of gravitons and what happens at high energies. But Einstein's theory is, at heart, a theory of geometry and this suggests different approaches to quantum gravity where, for example, spacetime sits in a superposition of all possible geometries, much like a particle in a double slit experiment follows all possible trajectories. This brings new conceptual issues to the table, but the problem described above remains, now in the guise of difficulties when spacetime fluctuates on very small scales,

$$L_{\rm pl} = \frac{1}{M_{\rm pl}} \approx 10^{-34}\ {\rm m}$$

It may well be that the very notion of space and time need replacing when we get to such small distance scales.

The Unreasonable Effectiveness of Classical Gravity

The analogy with Fermi's theory of the weak force gives us a useful way of thinking about what a dimensionful coupling, like G_N, means in quantum theories. But general relativity is a considerably deeper and more subtle theory than Fermi's and it has a number of tricks up its sleeve.

In particular, Fermi's theory of the weak force only works for energies $E \ll M_W$. For any energies higher than that, you need to use the gauge theory of W- and Z-bosons. But that's not the way things work for gravity. General relativity will give you the right answer to any quantum question at energies $E \ll M_{\rm pl}$. But if you throw together two particles at energies $E \gg M_{\rm pl}$, then general relativity will also give you the right answer. That's because, if you throw particles together at very high energies, then you simply form a black hole!

Black holes are nature's way of putting a limit on the amount of matter than can be squeezed into some small amount of space. They have two key features: at the centre of the black hole is the *singularity*. According to general relativity, this is a region where the curvature of spacetime becomes infinitely large. What that really means is that we shouldn't trust general relativity near the singularity and it should be replaced by a theory of quantum gravity. But, happily, we don't need a theory of quantum gravity to understand many features of black holes because the singularity is shielded from view by the second feature of a black hole: the *event horizon*. This is a surface that can be thought of as the edge of the black hole. If you venture through the event horizon of a black hole then you're in trouble: you will never escape and will be dragged inexorably towards the singularity. In this scenario, you will need to very

quickly develop a fully fledged theory of quantum gravity if you want to know what fate awaits you. If, however, you are less foolhardy and remain outside the event horizon then you can get by quite happily with our naive theory of quantum gravity.

That's not to say that there aren't interesting quantum gravity effects associated to the event horizon. If you study quantum field theory in the presence of a black hole, you find *Hawking radiation*, the process in which the black hole emits (mostly) photons and gravitons and slowly evaporates. It is a striking effect, and not without its own puzzles, but these too can seemingly be understood largely within a naive approach to quantum gravity, without the need for the full theory at the Planck scale.

The size of the event horizon is called the *Schwarzschild radius*. A black hole of mass M has Schwarzschild radius

$$R_s = 2GM \sim \frac{M}{M_{\rm pl}^2}$$

This simple equation turns one of the key ideas of particle physics on its head. Throughout these lectures, going to higher and higher energies has been tantamount to looking on smaller and smaller distance scales, with energy and distance related, in natural units, by $\lambda \sim 1/E$. But, when gravitational effects become important, we see that going to higher energies gives rise to bigger black holes. If we scatter two particles at energies $E \gg M_{\rm pl}$, then we make a black hole of size $R \sim E/M_{\rm pl}^2 \gg L_{\rm pl}$. This is how nature evades energy densities smaller than the Planck scale.

This means that, provided we don't do anything stupid, like jump into a black hole, we understand perfectly well what happens in very high energy scattering. You form a big black hole which slowly evaporates over gazillions of years. We never need any knowledge of the fundamental theory of quantum gravity to figure out the physics.

All of this is to say that the applications of a full theory of quantum gravity to scattering appear strangely limited. It is needed only in a small window of energies $E \sim M_{\rm pl}$. For energies $E \ll M_{\rm pl}$, a naive theory of quantum gravity will do the job, and for energies $E \gg M_{\rm pl}$, classical general relativity will do the job, at least until the resulting black hole has had enough time to evaporate down to size $R \sim L_{\rm pl}$. In final stages of the black hole's life, we will again need to resort to a detailed theory of quantum gravity to understand what happens.

Of course, not all of physics is to do with scattering. As we mentioned above, if we really want to understand what happens inside a black hole at the singularity then we surely need quantum gravity. Furthermore, a very similar kind of singularity arises in

general relativity at the Big Bang. If we follow the evolution of the universe back in time, then general theorems of Hawking and Penrose tell us that we will necessarily meet a singularity where general relativity breaks down. This means that if we want to answer the all-important question of how the universe began then we surely need a decent theory of quantum gravity. But it is rather hard to find any lingering effects of the Big Bang in our current universe. This is isn't simply because it happened a long time ago. Rather it is because there was a period of rapid expansion in the early universe known as *inflation* which dilutes any specific effects due to the singularity. It appears, once again, that Nature wishes to thwart us in our attempts to see directly the effects of quantum gravity.

This is not to say that quantum gravity is uninteresting. The questions of how singularities are resolved, both cosmological and those inside black holes, are important ones. Moreover, a closer look at how quantum theory meshes with gravity opens up a number of subtle conceptual issues, including how information escapes from black holes, how we should think about observers in an accelerating universe, and the seeming holographic nature of gravity. Moreover, there is the striking fact that, in string theory, we do have a fully fledged theory of quantum gravity, one that has intricate connections with various quantum field theories. All of these, however, are topics for other lectures.

5.3 Cosmology

> *There are more things in heaven and earth, Horatio,*
> *Than are dreamt of in your Standard Model.*

Hamlet was wrong about the earth. As we've seen, we are painfully short of experiments that cannot be explained by the Standard Model. But he was right about the heavens. When we look to the sky, it becomes clear that there is much we don't understand.

Our shocking lack of understanding becomes apparent when we audit the energy budget of our universe. This reveals that the particles of the Standard Model comprise just 5% of the total energy. All the things that we can see – stars, galaxies, planets, light itself – and all the things that we can detect through more indirect means – inert dust, neutrinos, gravitational waves – make up just a tiny fraction of the energy that is out there.

The vast majority of the energy in the universe is *dark*, which simply means that it doesn't interact (as far as we can tell) with the particles in the Standard Model. This "missing" energy can be further characterised as having two very different properties:

- Roughly 25% of the energy of the universe comprises of some new particles (or, better, new fields) that are not included in the Standard Model. This is called *dark matter*.

- The remaining 70% is a cosmological constant, sometimes called *dark energy*.

For reasons that we won't get into here, there's reason to believe that the short list above accounts for everything. There are very likely surprises in store for us in both dark energy and dark matter, but there is not some extra energy out there that we've missed completely. That's because, in the framework of cosmology, the total energy doesn't actually have to add up to 100%! Any deviation from 100% shows up as some overall curvature to the universe. But the universe is, as far as we can tell, exactly flat and that tallies nicely with our other observations which show that the energy hits the magical 100% threshold.

In this final section, we will elaborate on some of the puzzles of cosmology that are likely to have the biggest impact on particle physics. We will be fairly brief. Many more details can be found in the lectures on Cosmology.

5.3.1 The Cosmological Constant

The vast majority of energy in the universe is rather strange. It is best described as an anti-gravity force field, spread thinly throughout space, causing everything to repel everything else. We refer to this force field as *dark energy*.

We can't detect this dark energy here on earth. Nor can we see its effects within our solar system, nor even within our galaxy. Instead it is only when we look out at vast distance scales when the effects of dark energy become apparent as it manifests itself in the way the universe expands.

We've known that the universe is expanding since the 1920s. This is a straightforward prediction of Einstein's equations of general relativity, albeit one that was not embraced by theorists until the observational data made the case overwhelming. But around the turn of the last century, we learned something striking. The expansion of the universe is *speeding up* over time. We live not just in an expanding universe, but in an accelerating universe. This is despite the fact that the galaxies in the universe are all mutually attracted to each other through gravity which should cause the expansion to slow. But there is something else at play that overwhelms the natural gravitation attraction of galaxies at very large scales. This something else is dark energy.

In some sense, we understand dark energy very well. The Einstein equations of general relativity allow for exactly one additional term, with a single parameter Λ known as the *cosmological constant*. This extra term has exactly the right effect, changing the dynamics of spacetime to give the observed acceleration of the universe. All we have to do is set the cosmological constant to take the value,

$$\Lambda \approx (10^{-29}\text{ eV})^2$$

A better way of thinking about this is in terms of the density of dark energy which, in natural units (where length is equivalent to inverse energy) has dimensions of energy to the power 4. To get this, we should multiply by $M_{\rm pl}$, giving the energy density

$$\rho_\Lambda \approx \Lambda M_{\rm pl}^2 \approx \left(10^{-3}\text{ eV}\right)^4 \tag{5.9}$$

Just plug these values of the cosmological constant into the Einstein equations and you're done: the expansion of the universe then matches perfectly with what we observe.

(Before we go on, I should mention that this last sentence might not be quite true. There is currently a discrepancy between two different methods of measuring the expansion of the universe. The first, and arguably the cleanest, looks at the cosmic microwave background radiation, the afterglow of the Big Bang, and infers how the universe has subsequently evolved. The second looks at distant supernovae which are bright enough (and, apparently, uniform enough) to accurately measure both their distance and their recession velocity. These two results give answers that differ that level of 10% or so and it is difficult to reconcile them in any straightforward way. It may be, of course, that there is some unknown systematic in one of the experiments that gives a skewed result. Or it may be that there is something deep going on that we've missed. This is known as the *Hubble tension*.)

The Energy of the Vacuum

Although we have the equations to describe the accelerated expansion of the universe, there is a level of disquiet about the cosmological constant. This derives from the fact that we actually have a good understanding of the physics underlying the cosmological constant and this seems to be in conflict with the observed value.

The cosmological constant is, it turns out, something very familiar in physics: it is the *vacuum energy*. Recall that we're taught in high school that the overall value of the energy doesn't matter. Instead, only energy differences are important. That's true in all situations except for in cosmology. Gravity responds to all kinds of energy including the energy of empty space and this appears as the cosmological constant in the Einstein equations where it affects the overall expansion of the universe.

In some ways, this is a good thing. The cosmological constant isn't just some random number that we've plucked out of thin air to account for the expansion of the universe. Instead it's something that should arise naturally from our other laws of physics. And this gives us the opportunity to calculate it from what we know about particle physics.

This is where the narrative takes something of a left turn. The vacuum energy in quantum field theory is something interesting. Recall that one of the characteristic features of quantum field theory is that the vacuum isn't a dull place: the quantum fields froth with quantum fluctuations. We showed an example of the quantum field vacuum way back in Figure 4. And all of those fluctuations contribute to the vacuum energy. If the vacuum has fluctuations on some maximum energy scale E, then we typically expect a ground state energy density of order E^4.

That's problematic. We know that the Standard Model of particle physics holds up to the energy scale of a TeV or so. But this strongly suggests that the vacuum energy density should be at least

$$\rho_{\rm SM} = (10^{12}\ {\rm eV})^4$$

This is not particularly close to the observed value. It is 60 orders of magnitude greater than the observed value ρ_Λ. More generally, the contribution of any quantum field to the vacuum energy density is naturally of order $\rho_{\rm QFT} \sim \Lambda_{\rm UV}^4$, where $\Lambda_{\rm UV}$ is the UV cut-off.

In the cosmological context, a vacuum energy density of order a TeV or higher gives ridiculous results. A cosmological constant this large would give a universe that expands so quickly that it is not conducive to forming nuclei or atoms, let alone galaxies and life. The huge discrepancy between the expected value of ρ_{QFT} and the observed value of ρ_Λ is known as the *cosmological constant problem*.

The cosmological constant problem is entirely analogous to the hierarchy problem that surrounds the Higgs mass that we discussed in Section 5.1.2. In both cases, quantum corrections seem to naturally push the value of some quantity to be much higher than we observe. This happens only for the Higgs mass and the cosmological constant because these are the only two dimensionful parameters in the known laws of physics. (Strictly speaking, we also have the Planck mass $M_{\rm pl}$, but this can be thought as setting the scale relative to which all other parameters are measured.)

The cosmological constant problem is sometimes referred to as the worst prediction in the history of physics. The people making this claim clearly haven't studied the

history of physics. (Rayleigh-Jeans law anyone?) Nonetheless, the prediction is clearly nothing to be proud of. The solution to the ultra-violet catastrophe inherent in the Rayleigh-Jeans law ultimately resulted in the greatest paradigm shift in the history of science: the discovery of quantum mechanics. It's not clear if the cosmological constant problem will result in a similar upheaval to our understanding of physics or whether, with some small twist of our head, the question will unravel as we realise that we've been looking at things the wrong way. For now, it's fair to say that we have just one solution to the cosmological constant problem. This is the idea that, in addition to the contribution from quantum field fields, there is another contribution to the cosmological constant, so that the two add up to give the observed value (5.9)

$$\rho_\Lambda = \rho_{\rm SM} + \rho_{\text{something else}}$$

Clearly the additional contribution $\rho_{\text{something else}}$ must cancel the contribution from the Standard Model to 60 significant figures, leaving behind the tiny cosmological constant that we observe. This is another example of fine tuning. It is the same kind of idea that we met in equation (5.5) when discussing the mass of the Higgs boson.

It is quite possible that there is some missing principle that we've failed to grasp that makes fine tuning less silly than it first appears. The task of finding such a mechanism is made considerably harder when we realise that there have been a number of times in the history of the universe when $\rho_{\rm SM}$ abruptly changed while, presumably, $\rho_{\text{something else}}$ did not. This occurs at a *phase transition*. For example, there was a time in the early universe when things were so hot that quarks and gluons were not confined inside baryons and mesons. As the universe cooled, confinement kicked in and, at that time, the vacuum energy of the Standard Model jumped by around $\Delta\rho_{\rm SM} \sim (100\ {\rm MeV})^4$. Still earlier, the electroweak phase transition, where the Higgs boson first condensed, resulted in a change of $\Delta\rho_{QFT} \sim (100\ {\rm GeV})^4$. This means that any putative cancellation mechanism must conspire to give a tiny cosmological constant ρ_Λ at the end of the life of the universe, not at the beginning.

Before we muddy the waters yet further, it's worth mentioning that the observed vacuum energy (5.9) is pretty much in the same ballpark as the neutrino masses. Is this coincidence? Probably. Certainly it's hard to know what to make of it. But, given our evident confusion about the cosmological constant, it's worth bearing in mind.

The A-Word

As we saw above, a naive application of quantum field theory suggests a ludicrous value for the cosmological constant, one that results in an expansion so fast that not even

atoms have a chance to form from their underlying constituents. Given this, we could ask the following question: what is the maximum value of the cosmological constant that still allows complex structures to evolve? For example, what is the maximum allowed value of Λ that allows galaxies to form?

It turns out that the upper bound on Λ depends on the strength of the initial seeds from which the galaxies grew. (We'll mention these briefly later in this section.) However, if we fix this initial condition, then we can ask again: how big can the cosmological constant be?

The answer is quite striking: the scale of the vacuum energy is pretty much the maximum it could be. If ρ_Λ were bigger by an order of magnitude or so, then no galaxies would form, presumably making it rather more difficult for life to find a comfortable foothold in the universe.

What to make of this observation? One possibility is to shrug and move on. Another is to weave an elaborate story. Suppose that our observable universe is part of a much larger structure, a "multiverse" in which different domains exhibit different values of the fundamental parameters, or perhaps even different laws of physics. In this way, the cosmological constant is not a fundamental parameter which we may hope to predict, but rather an environmental parameter, no different from, say, the distance between the Earth and the Sun. We should not be shocked by its seemingly small value because, were it any higher, we wouldn't be around to comment on it. Such reasoning goes by the name of the *anthropic principle*.

The anthropic explanation for the cosmological constant may be correct. But, in the absence of any testable predictions, it is not clear what to make of it and further philosophising tends to be more of a distraction than a help.

A Rebranding: Dark Energy

Given our manifest befuddlement about all things Λ, it is prudent to wonder if perhaps the accelerated expansion of the universe has nothing to do with a cosmological constant at all! It is quite possible that the cosmological constant in the Einstein equations is $\Lambda = 0$. If this is the case, then we need to look for another explanation for the accelerated expansion. It is not difficult to find such explanations, although none of them are particularly compelling. For example, scalar field with a ridiculously low mass (around 10^{-33} eV or so), rolling down a potential can do the job should we wish.

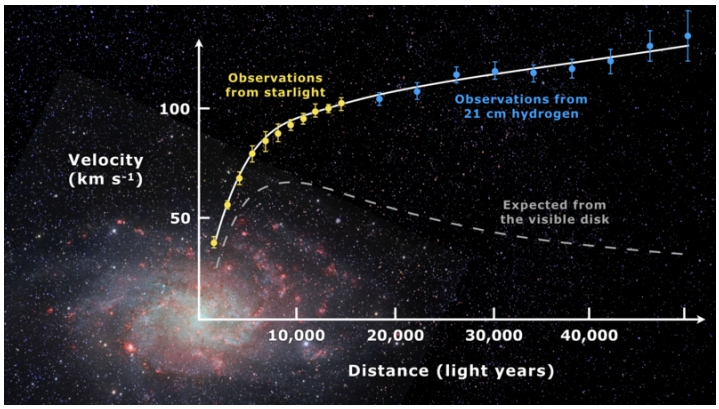


Figure 54. The rotation curve of galaxy M33. Image taken from Wikipedia.

I should stress that these new explanations in no way alleviate the cosmological constant problem. Finding a mechanism that sets the cosmological constant to zero is no easier than finding a mechanism that sets it to the observed value. Furthermore, one typically introduces many more fine tuning issues in whatever new dynamics is then introduced to drive the accelerated expansion.

Nonetheless, the history of particle physics has taught us that we shouldn't be too hasty in following our preconceived ideas about how the universe should be. To avoid committing to the cosmological constant explanation, the mysterious 70% of the energy in the universe is often referred to as *dark energy*.

5.3.2 Dark Matter

Dark matter comprises around 25% of the energy density of the universe. Unlike dark energy, it is conceptually quite straightforward: it is simply some new, heavy particles that are not accounted for in the Standard Model. Or, at a fundamental level, some new quantum fields.

We know very little about the properties of dark matter, beyond the fact that it does not interact with light. We do not know if it is a single species of particle, or many. We do not know if it consists of several decoupled sectors, or just one. Given the wonderful complexity of the Standard Model, it seems reasonable to assume that there is still rather a lot to learn about dark matter.

All we know about dark matter comes from its gravitational interactions. Yet the combined evidence is overwhelming. Here are some highlights:

Figure 55. On the left, the Abell S1063 cluster. The smeared blue lines are background galaxies, distorted by gravitational lensing. On the right, the bullet cluster.

- The beautiful spiral galaxies that we see in the sky seem to be spinning too fast! The attractive gravitational force from all the stars in the galaxy, does not come close to reproducing the necessary centripetal force to stop the galaxy from flying apart. Moreover, if you measure the spectral lines of hydrogen far from the visible edge of the galaxy, you find that it continues to rotate at a roughly constant speed for quite some distance. All of this can be explained by simple Newtonian dynamics, but only if there is much more mass in the galaxy than is visible, To account for the observations, there should be a roughly spherical cloud of dark matter surrounding the galaxy.

  The rotation curve for a nearby galaxy, together with the predicted curve if there is only the visible matter, is shown in Figure 54.

- A *galaxy cluster* is a collection of 100 to 1000 galaxies, bound together by gravity. A clever argument, known as the virial theorem, gives a relationship between the speed of the galaxies and their separation (or, more precisely, their kinetic energy and potential energy). From this, one can extract the mass of the galaxy. The answer is a couple of hundred times greater than the visible mass.

- A classic prediction of general relativity is that light bends as it passes heavy objects. Furthermore, the image gets distorted, a phenomenon known as *gravitational lensing*. Sometimes this happens in a spectacular fashion, as shown in the picture on the left of Figure 55 where the image of a background galaxy is distorted into the blue arcs by the cluster in the foreground. Even small distortions of this kind allow us accurately determine the mass of the cluster in the

foreground. You will not be surprised to hear that the mass greatly exceeds that seen in visible matter.

The bullet cluster, shown in the right of Figure 55, provides a particularly dramatic example of gravitational lensing. This picture shows two sub-clusters of galaxies which are thought to have previously collided. There are three types of matter shown in the picture: stars which you can see, hot gas which is observed in x-rays and is shown in pink, and the distribution of mass detected through gravitational lensing shown in blue. The stars sit cleanly in two distinct sub-clusters because individual galaxies have little chance of collision. In contrast, most of the ordinary matter sits in clouds of hot gas which interact fairly strongly as the clusters collide, slowing the gas and leaving it displaced from the stars as shown in the figure. But most of the matter, as detected through gravitational lensing, is dark and this, like the galaxies, has glided past each other seemingly unaffected by the collision. The interpretation is that dark matter interacts weakly, both with itself and with ordinary matter.

- The observations described above show clearly that, on the scale of both galaxies and clusters of galaxies, there is more matter than can be detected by electromagnetic radiation. This alone is not sufficient to tell us that dark matter must be composed of some new unknown particle. For example, it could be in the form of failed stars ("jupiters"). There is, however, compelling evidence that this is not the case, and dark matter is something more exotic.

  The primary evidence comes from *Big Bang nucleosynthesis*, an impressively accurate theory of how the light elements were forged in the early universe. It turns out that the relative abundance of different elements depends on the total amount of baryon matter. In particular, the amount of deuterium compared to everything else depends in a delicate way on the total amount of ordinary matter. This tells us that the total amount of ordinary matter is just a few percent of the total energy density.

- In Section 5.3.4, we'll describe the cosmic microwave background, a photograph of the fireball that filled the universe when it was very much younger. The flickering of this fireball shows that some spots were hot and others cold. As the universe evolved, these initial fluctuations provided that seeds that later grew into the clusters, galaxies and stars that we see around us.

  It turns out, however, that this cannot be achieved by ordinary matter alone. There simply isn't time, largely because ordinary matter couples to photons and

this causes a pressure which suppresses gravitational collapse. But dark matter doesn't know about the photons, so there is nothing to stop it forming gravitational wells into which visible matter can subsequently fall. The whole story of the formation of galaxies only works because of the existence of dark matter.

Moreover, the existence of dark matter also leaves an distinct imprint in the ripples of the cosmic microwave background (specifically in the relative heights of the first and second peak of the power spectrum).

Taken individually, one might have thought that there could be some alternative explanation for these pieces of evidence. For example, faced just with the galaxy rotation curves, one might try to tinker with Newton's equations of motion to get something that fits. But that will then leave us unable to explain, say, how the light elements were forged in Big Bang nucleosynthesis. And yet all of the problems above are resolved by the simple admission that there are particles (or, strictly speaking, quantum fields) in the universe that are not accounted for in the Standard Model.

Clearly we should try to understand these fields and, ultimately, enlarge the Standard Model to embrace them. This is not so straightforward because all our evidence for dark matter comes from gravitational interactions alone. In an attempt to change this, there are many ongoing experiments designed to detect dark matter here on earth. All of these rely on the hope that there is some non-gravitational interaction between dark matter and Standard Model fields, perhaps through the weak force or perhaps through some new, as yet undiscovered force. These experiments are increasingly impressive at pushing the boundaries and one can only hope that they will one day make a key discovery.

There is one reason for optimism here. This is because the abundance of dark matter and ordinary matter is not wildly different. There is more dark matter by a factor of 5, but not a factor of 500 or 5 billion. The most plausible explanation for this is if dark matter and ordinary matter were in equilibrium together in the early universe, before they subsequently decoupled. But such equilibrium can only be maintained if there are some non-gravitational interactions between them.

In particular, there is one tantalising hint. If one assumes that dark matter has a mass of around the TeV scale associated to the weak force, and moreover interacts with the strength of the weak force, then the relative abundances come up just about right. However, if it were to interact directly through the weak force at this scale then we must ask why it hasn't shown itself at the LHC. Maybe this tantalising hint is merely a red herring.

5.3.3 Baryogenesis

The universe contains lots of matter but very little anti-matter. How did this asymmetry come to be?

One possibility is that it is an initial condition on the universe. Another is that the universe started with equal amounts of matter and anti-matter, but somehow a small dynamical shift took place that preferred one over the other. This latter process is known as *baryogenesis*.

We don't have an established theory of baryogenesis, but there are a set of three criteria that must be obeyed, known as the *Sakharov conditions*. These are:

- The first criterion is the most obvious: particle number cannot be a conserved quantity. Here "particle number" refers to particles minus anti-particles. In a symmetric universe, the total particle number would start off at zero. We want it to end up at something non-zero.

  In the Standard Model, both baryon number and lepton number are almost conserved although, as we saw in Section 4.3.5, in extreme conditions only $B - L$ is strictly conserved. The early universe certainly counts as an extreme condition. The need for baryogenesis suggests that we need interactions that break the symmetry $B - L$. For example, a Majorana mass for neutrinos will do the job.

- The symmetry CP must also be broken. As we've seen in Section 4.3.4, CP relates the behaviour of particles to anti-particles, but for baryogenesis to occur their behaviour must be different.

  We've seen that CP is violated in the quark sector, but this is not enough to give rise to the necessary level of baryogenesis. It remains to be seen whether CP violation in the lepton sector is sufficient to do the job, or whether baryogenesis requires interactions beyond those discussed in these lectures.

- The final criterion is the least obvious: the early universe must deviate from thermal equilibrium.

  A deviation from thermal equilibrium occurs when the universe undergoes a first order phase transition. (You can read more about phase transitions in the lectures on Statistical Physics and Statistical Field Theory.) The Standard Model does not appear to offer the opportunity for such a violent event at the necessary energy scales. This suggests that we should need some new physics to induce baryogenesis.

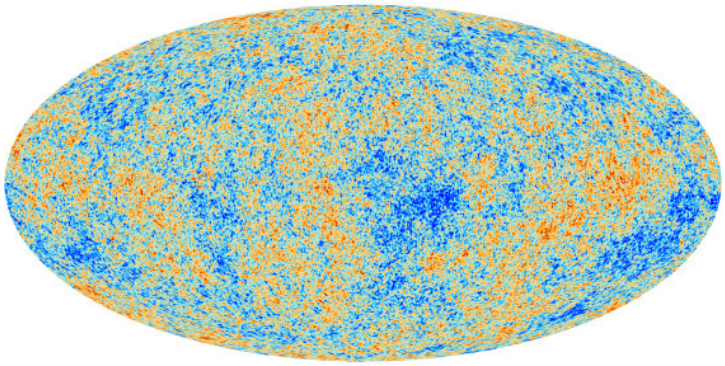


Figure 56. Look and weep, Ansel Adams.

There are many models of baryogenesis on the market, but currently no smoking gun experiment or observation that will determine which, if any, is correct.

5.3.4 Primordial Fluctuations

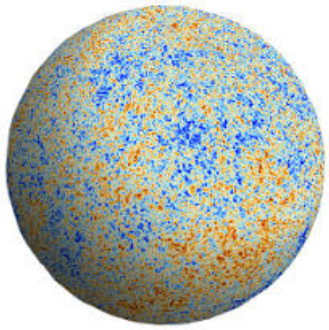
Most reasonable people agree that the greatest movie ever made is Ghostbusters. Sadly the world contains no small number of unreasonable people, those who prefer their movies to have a less intellectual bent, or those who put less stock in powerful acting performances and groundbreaking cinematography. It is difficult to argue that the opinion of these people is any less valid just because it is wrong. Art is not like science. There is no immutable, underlying truth that determines what is the right and wrong.

Until, that is, we come to photography. No one, reasonable or otherwise, can disagree about the greatest photograph ever taken. All other candidates pale into insignificance when faced with the collective endeavours of a bunch of radio horns and a handful of satellites who, between them, achieved the seemingly impossible feat of taking a photograph of the Big Bang.

First, I should tell you what the Big Bang theory entails. It is not a theory that tells us how the universe started. The question "how did the universe start?" has a very straightforward answer which is "we don't know". Instead, the Big Bang theory tells us what the universe was like when it was very much younger. The theory starts with the observation that there was a time — 13.8 billion years ago to be precise – when the universe was so hot that matter, atoms and even nuclei melted and all of space was filled with a fireball. When I say that we've taken a photograph of the Big Bang, I mean that we've taken a photograph of this fireball, capturing the light that

has travelled through the universe uninterrupted for almost 14 billion years. This light is known as the *cosmic microwave background*, or CMB for short.

The clearest photograph that we have was taken by the Planck satellite and is shown in Figure 56. This is a panoramic shot, containing information from each point in the sky which is then depicted in 2d much like a map of the Earth. A better setting for the CMB is shown in figure to the right. We sit in the centre of this sphere. If we look far enough away, roughly 20 billion light years in any direction, then we see the CMB.



In the early universe, the fireball reached extreme temperatures, almost certainly the most extreme temperatures the universe has ever seen. But, as the universe expanded, the fireball cooled and it is now a tepid 2.73 Kelvin. This temperature is almost uniform across the sky, but there are small fluctuations at the level of 1 part in 10^5. These hot and cold spots are depicted in red and blue in the photograph. These fluctuations have been imprinted in the CMB for 14 billion years, and so contain a wealth of information about what the universe was like when it was much younger. This information is usually plotted as a function of angular scale, as shown in Figure 57. From the positions of the peaks and troughs, we can determine much about the age and contents of the universe. For example, the position of the first peak contains information about the age of the universe, while the relative height of the first and second peaks contains information about the amount of dark matter in the universe.

Here, however, our interest is rather different. The question that we wish to ask is: where did those temperature fluctuations come from originally? This question for which we're fairly confident that we have the answer to this. And it is nothing short of astonishing.

Inflation

The answer involves a process known as *inflation*, a period of rapid accelerated expansion when the universe was very young. Here "very young" means when the universe was, as most 10^{-11} second old, but most likely it occurred much earlier than this. (In this counting, 10^{-30} seconds counts as "much younger" than 10^{-11} seconds!)

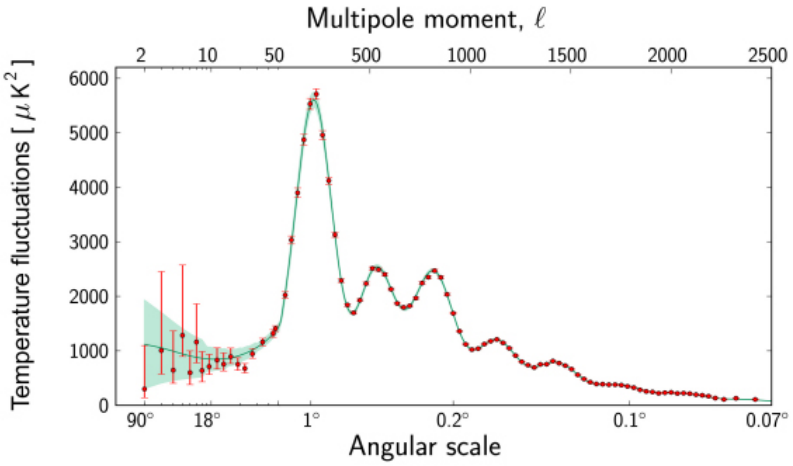


Figure 57. The dependence of the CMB temperature fluctuations with angular scale.

The original reasons for postulating the existence of an inflationary phase had nothing to do with the fluctuations. Instead, such a phase would resolve two unanswered questions about the universe we live in.

The first is: why is the universe so flat? It turns out that a flat universe is dynamically unstable, like a pencil balanced on its tip. Any small overall curvature at the beginning of the universe would have grown over time yet, after 14 billion, the universe appears as flat as a metaphorical and cosmological pancake. Why? Inflation gives an answer to this question. A brief phase of rapid expansion, stretches out any curvature that may have once existed, like pulling a membrane taut. This nicely explains why we find ourselves living in a flat universe.

The second is: why is the CMB so uniform? We can stare out at the sky in one direction and detect photons from the CMB, photons that have travelled roughly 20 billion light years uninterrupted. We can then turn around and see photons that have travelled 20 billion light years from the opposite direction. The properties of these photons are basically the same: in particular, both exhibit identical temperatures of 2.73 Kelvin. The fact that these two far flung reaches of the universe have the same temperature is a puzzle. Usually when two systems sit at the same temperature it's because they've had the opportunity to interact, exchange energy, and settle down to thermal equilibrium. But those two patches of the universe have had no such opportunity simply because they're too far away. Light from one region of space hasn't had

time to reach the other: indeed, we're sitting in the middle and it's only now that we can see both regions. So how could they possibly sit at the same temperature?

There is a more nuanced version of this second issue. This arises, somewhat ironically, when we appreciate that the CMB is not completely uniform after all but contains the tiny fluctuations shown in the photograph. There are fluctuations in both the temperature and the polarisation of light and, importantly, these two different types of fluctuations are correlated. These correlations – which go by the uninspiring name of "TE correlations" – are the kind of the type that would arise through simple and well understood dynamical processes in the early universe, such as photons scattering off electrons. But observations reveal that there are correlations over patches of the sky that were, apparently, never in causal contact with each other. That's a worry. Taken naively, it's telling us that there were dynamical processes in the early universe that occurred faster than the speed of light and that violates one of the key tenets of physics.

The TE correlations are, by far, the strongest argument for inflation. If we want preserve some of our most cherished notions of physics, like locality and causality, then it tells us very clearly that those far flung patches of the sky *must*, in fact, have been in causal contact back in the day. Inflation is the mechanism that allows this to happen. In the very early universe, two patches of space could be near. But inflation then takes those patches and stretches them to enormous distances, until they sit on opposite sides of the observable universe. Yet they retain, in the CMB, the correlation that gives the game away that they were once playmates.

The arguments given above strongly suggest that inflation happened. The next question is: what caused it? Here we're on less sound footing. The good news is that when general relativity is coupled to quantum fields, one naturally gets the rapid expansion that we need for inflation. Indeed, we've already seen how to do it: the vacuum energy, or cosmological constant, does the job. The accelerated phase of inflation in the early universe is conceptually identical to the accelerated phase that we now find ourselves in, but with two differences. The first difference is quantitative: the effective cosmological constant in the early universe must have been many many orders of magnitude larger than what we see today. The second difference is that the original period of inflation must, ultimately, have come to a halt. The real challenge in constructing a viable model is therefore figuring out how to get inflation to stop!

This, it turns out, is not so difficult. Indeed, if there's bad news in the story of inflation, it is that it's *too* easy to write down models that do the job which means that

every Tong, Dick and Harry has their own theory of inflation, with hundreds now on the market and very little to decide between them. None of these models are particularly exotic and nearly all include the same basic ingredient: a scalar field. The idea is that the inflaton, unlike the Higgs boson, had not yet settled to the minimum of its potential energy in the early universe, so its value changes over time. Correspondingly, so too does the vacuum energy that drives inflation.

None of the models for inflation stand out as being overly compelling. Indeed, in many ways they all look somewhat artificial. One might wonder if perhaps we could identify the inflaton with the Higgs boson itself, and there have been attempts to do so, but it's not a particularly natural fit. This means that while there is good evidence that the process of inflation took place, our knowledge is limited when it comes to the detailed underlying dynamics. Before we beat ourselves up about this too much, it's worth remembering that we're talking about a process that happened something like 10^{-30} seconds after the Big Bang. The fact that we haven't yet got all the details pinned down isn't terribly surprising.

Inflationary Perturbations

All of which brings us to the main topic of this section: where did the ripples in the CMB come from? These ripples contain correlations over large distance scales which means that, if they are to have a local and causal origin, then they must have been laid down during the inflationary period itself. Happily, inflation provides a remarkable origin story for these fluctuations.

The ripples arise simply from the realisation that the inflaton is a quantum field. In fact, we started these lectures by explaining that the vacuum of space is not a dull place since the quantum fields cannot stay still: they froth and bubble with quantum jitters. During inflation, the universe expands so quickly that these quantum fluctuations get caught in the act and are stretched from the microscopic scale, to distances that span the entire visible universe. These are what we see imprinted as hot and cold spots in the CMB: they are nothing less than quantum fluctuations that took place just fractions of a second after the Big Bang and are then frozen in place by the rapid expansion of the universe.

This may be one of the most extraordinary ideas in all of science, connecting our understanding of physics on the very smallest scales with that on the very largest. It passes many checks. A statistical analysis of the CMB fluctuations shows that they agree perfectly with those expected from a weakly interacting quantum field. Moreover, as we get better data on the fluctuations, so we begin to get a handle on

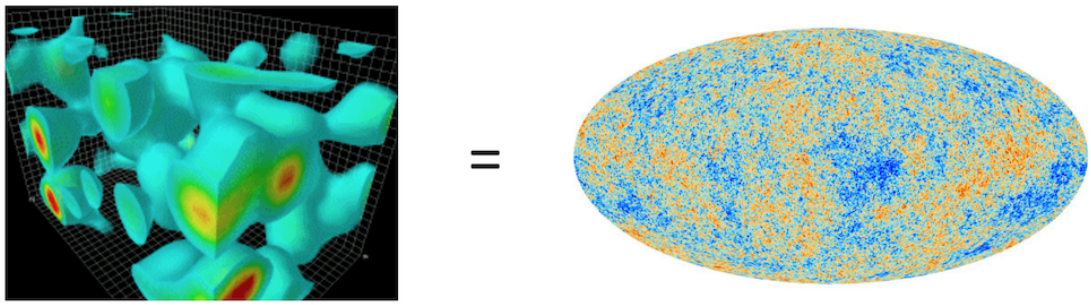


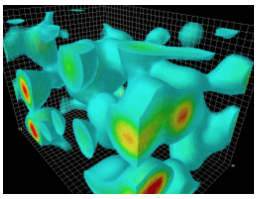
Figure 58. The ripples in the CMB are quantum vacuum fluctuations, laid down in the first few fractions of a second after the Big Bang.

the dynamics of the inflaton field in the very early universe. We're currently at a stage where many putative models of inflation can be ruled out, although there are many that still survive. The hope is that further study of the CMB will yield precious clues about the interactions of these quantum fields in the first few moments after the Big Bang.

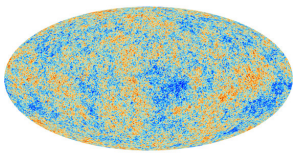
Remarkably, there is one further chapter to this story. Rather than asking where the ripples in the CMB came from, we could ask what subsequently happened to them. Here too we find an astonishing answer. The quantum fluctuations resulted in temperature variations in the CMB, with some places hotter and others colder. As the universe expanded and cooled, these hot and cold spots became the gravitational wells into which matter fell. First protons and electrons, which subsequently bound together into hydrogen dust and a smattering of other light elements. Over time, this dust gathered, and the pressure grew until finally, after 500 million years or so, these balls of dust ignited and became stars.

This means that the quantum fluctuations from when the universe was in its infancy later became the seeds from which galaxies grew. This is backed up by observation: a statistical analysis of the galaxies in our universe matches impressively with an analysis of the CMB fluctuations. For example, the large peak in Figure 57 manifests itself in a particular way in which galaxies cluster in the sky (known, boringly as "baryon acoustic oscillations").

Putting the pieces together, we can draw a direct line from this:



to this:



to this:

It is one of the most remarkable stories in all of science, but there are many details still to be written. Hopefully, in the future we will understand better how the quantum fields involved in this story fit with those of the Standard Model.