Part I

Physics as global human enterprise for understanding Nature

IOP Publishing

EPS Grand Challenges

Physics for Society in the Horizon 2050

Mairi Sakellariadou, Claudia-Elisabeth Wulz, Kees van Der Beek, Felix Ritort, Bart van Tiggelen, Ralph Assmann, Giulio Cerullo, Luisa Cifarelli, Carlos Hidalgo, Felicia Barbato, Christian Beck, Christophe Rossel and Luc van Dyck

Chapter 2

Physics bridging the infinities

Freya Blekman, Angela Bracco, Nelson Christensen, Emmanuel Dormy, Patrick Eggenberger, Claus Kiefer, Franck Lépine, Mairi Sakellariadou, Emmanuel N Saridakis, Jochen Schieck and Claudia-Elisabeth Wulz

2.1 Introduction

Mairi Sakellariadou[1] and Claudia-Elisabeth Wulz[2]

[1]King's College London, London, UK

[2]Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, Austria

At the horizon 2050, our physics textbooks will have to be rewritten.

When the Higgs boson was found in 2012 at CERN's Large Hadron Collider (LHC) in Geneva, history was made. This particle, and its associated field, is the reason why atoms, stars, galaxies, and you, reading this book, are tangible entities. In addition, a tiny asymmetry between matter and antimatter that developed soon after the Big Bang made it possible for us to exist at all. Without the Higgs boson and this asymmetry, only radiation would permeate the Universe.

Detailed studies of the Higgs boson, at current or future colliders, as well as precision measurements of the properties of matter and antimatter at a multitude of different experiments will reveal how the Standard Model of particle physics has to be amended. That it needs to be extended is evident. It does not contain dark matter, whose existence was already manifest decades ago. Another phenomenon, only discovered in 1998 through the study of the brightness of supernovae as a function of their distance, is dark energy, which makes our Universe expand in an accelerated fashion. Known or 'visible' matter, the so-called baryonic matter, only accounts for 5% of the Universe, and is well described by the Standard Model of particle physics. The rest are dark matter (27%) and dark energy (68%). We have hardly any clues, but many ideas of what they could be.

doi:10.1088/978-0-7503-6342-6ch2

Although postulated already in 1930, neutrinos are another category of particles that are still mysterious. It was only ascertained in the 1990s that they have mass, in contrast to the assumption in the Standard Model of particle physics, and that they come in different flavours that can transform into each other. There might even be more varieties—sterile neutrinos—which do not interact through the forces described by the Standard Model, but only through gravity.

Although the latter is so present in our everyday life and the movements of objects in the cosmos, it is the least understood force, and is not part of the Standard Model of particle physics. We do not even have a quantum-mechanical formulation of the theory of gravity yet, which would allow us to describe this force down to the smallest scales of the Universe. Our current understanding is based on Einstein's theory of general relativity, which however breaks down at the centre of black holes, for example. Everywhere else, it has so far been proven to be perfectly descriptive and accurate. The spectacular direct discovery of gravitational waves—ripples in space time predicted to arise from violent events in the cosmos such as mergers of black holes or neutron stars—in 2015 confirmed it once more. This discovery has further opened up a new field called multimessenger astronomy. We are no longer limited to observing the sky with our eyes or with telescopes detecting light or other electromagnetic waves, but we now also have gravitational waves, and neutrinos, at our disposal as messengers from cosmic sources. We can study all kinds of signals in a coordinated fashion in experimental facilities around the globe and even in space.

There are many bridges between the smallest and the largest scales. Nuclear physics, with its quest to understand the origins of known matter, from the primordial soup made of quarks and gluons, the protons and neutrons, the atomic nuclei, to the formation of the heavy chemical elements in explosions of stars, connects these infinities. It also has a large potential for technological spin-offs such as nuclear fusion to ensure the supply of electric power and medical applications such as cancer therapy, as well as efficient and affordable isotope production for diagnostic purposes. For the latter, imaging techniques using artificial intelligence and other means are drivers for improving diagnostic accuracy, rapidity, and the comfort of patients. Astrophysics and high-energy particle physics are also connecting the scales, and have given rise to the new field of astroparticle physics.

Cosmology, with its quest to understand the largest scales and nothing less than the fate of our Universe, needs information about its smallest components. Amongst others, measurements by space observatories such as Planck operated by the European Space Agency have helped establish the now widely accepted Standard Model of cosmology. For the time being, we are at a turning point in our knowledge of the future of our Universe. Soon we should know more about its evolution, and in particular, whether it will be confirmed to expand forever, or to rip apart, or even contract again.

The stars, the Sun, and the planets, including our own, still have secrets themselves. Their formation and evolution are vibrant research areas, tackled through computations and observations, and exploiting a multi-disciplinary approach. The study of exoplanets has also become a central subject in astrophysics. More down-to-earth, geophysics addresses topics that can affect us all, such as volcanic eruptions, earthquakes, or even changes in the Earth's magnetic field. The understanding of these phenomena can help predict their occurrence, for the benefit of mankind.

Will all or many of the open questions be answered at the horizon 2050? There is justified hope, supported by a plethora of theoretical developments and experimental facilities on Earth and in space. Will new questions arise? You bet.

2.2 Particle physics: physics beyond the Standard Model

Freya Blekman[1,2]
[1]Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[2]Institut für Experimentalphysik, Universität Hamburg, Hamburg, Germany

2.2.1 Particle physics: the extremely small connects to the big scientific questions

Particle physics explores Nature at the tiniest scales by studying the properties and interactions of elementary particles. Many of these properties and interactions between elementary particles are predicted in a theoretical framework known as the Standard Model (SM) (figure 2.1).

The predictions of the SM have been experimentally confirmed to an extraordinary degree of precision by a series of historic particle accelerators and most recently using the collisions at the LHC at CERN. Understanding the behaviour of these elementary particles has consequences to other fields of physics and beyond. For example, knowledge of the SM is necessary to obtain insight into the Universe just after its creation during the Big Bang. Another of these questions is that only about 5% of all matter and energy that constitutes our environment that we experience can be described with the SM, meaning that the remaining almost 95% is presently unaccounted for by our current knowledge. The remaining 'dark' energy in the Universe is referred to as Dark Matter and Dark Energy.

Particle physics does not only deal with the fundamental constituents of matter but also with the processes of how particles interact. While in our macroscopic world, an almost infinite number of forces and interactions can be experienced, on the microscopic scale of particles, these can be reduced to only four known types of interactions:

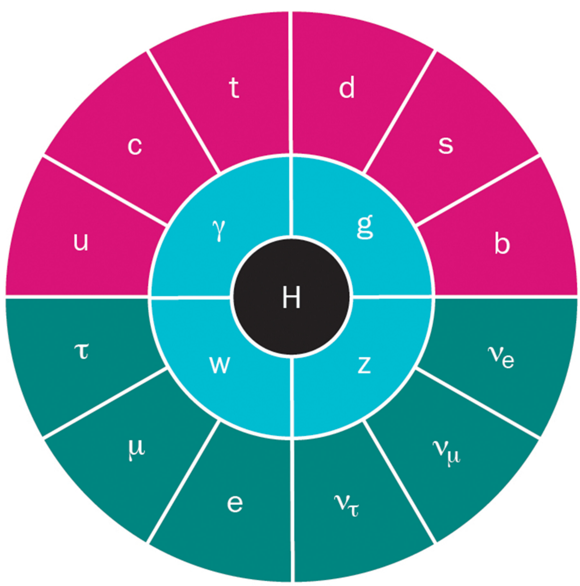


Figure 2.1. The Standard Model, as represented in [1].

gravitational, weak, electromagnetic, and strong. Gravitation is the force that people may be most familiar with, but it is counter-intuitively the least understood when dealing with particles. The weak interaction is responsible for, for example, the radioactive interactions in stars like our Sun. Electromagnetism, responsible for most of our daily life including light, magnetism, and the stability of molecules and the orbits of electrons in atoms, is deeply connected to the weak interaction and shares many of its properties. Finally, the strong interaction is the mechanism that ensures that the particles inside the atomic nucleus remain stable. The weak, electromagnetic, and strong forces are all included in the SM. The weak, electromagnetic, and strong interactions in the SM are mediated through particle interchanges between the quarks and leptons, the constituents of matter. The most commonly known quantum mediator is the one associated with electromagnetism, the photon. For the strong and weak forces, other equivalent particles exist that are called the gluon and the $W^{\pm}$ and Z bosons, respectively. Substantial work both on the theoretical side and extraordinary experimental proof to confirm any such new theory will be necessary to achieve the inclusion of gravity in the SM or in a Grand Unification Theory, one of the noble goals of physics. For example, particle physics uses conservation laws equivalent to the well-known rule of conservation of energy in the form of symmetries that a quantum field theory can describe. This framework is the SM.

In total, the SM describes the behaviour of all known elementary particles. The quarks can be used to create composite particles such as protons and neutrons. There are also three charged leptons such as electrons, and three neutrinos. These fermions are grouped into three so-called ‘generations’ of increasing mass. The origin of the generational structure is currently not understood. The forces are represented by the photon, $W^{\pm}$ and Z massive gauge bosons, and the gluon. The Higgs boson completes the present picture of the SM by inducing mass to the elementary particles, which would otherwise be mathematically massless in the SM.

It is essential to be aware that there are critical weak points in the SM as presented in the previous paragraphs. Additional unexplained arguments that the SM is not the complete picture of particles at the smallest scale include the apparent conflict in the equal treatment of matter and antimatter in the SM. After all, we live in a world dominated entirely by matter, a fact that is largely unexplained by the SM. Also, the SM does not explain why particles have different masses; mathematically, it allows all quarks and leptons to have equal mass. The proton would not be stable at that point, and atoms and molecules would not be stable as we know them to be.

The SM importantly also relies on neutrinos being massless. However, it was proven in the early 2000s [2] that neutrinos have a tiny but non-zero mass. Not only does the SM not need the neutrinos to have mass, but the neutrino mass also creates problems. The mass of neutrinos cannot be easily explained the same way as other particles and naturally suggest that the SM needs to be extended with more symmetries, forces, or interactions that solve the neutrino conundrum.

This means that physicists look for more explanations to create a better description of Nature than the SM. The diverse Beyond-the-Standard-Model (BSM) physics theories typically involve a new mathematical formalism that controls the known types of interactions and particles and almost always predict the existence of additional undiscovered particles. The search for evidence of the (in) direct production and existence of these particles and mechanisms is an important research challenge in fundamental physics.

A large number of theoretical extensions to the SM fix one or many of the previously mentioned problems. Unfortunately, it is currently impossible to identify which of these extensions is most likely to solve the problems of the SM. However, certain physicists indeed have a preference. Particle physics is at the moment very much a data-driven science, meaning that without further empirical input from experiments, it seems to be impossible to solve these questions. The goal of the programme of the LHC now that the Higgs particle has been discovered is to probe the particle world at the smallest possible scale to find answers to these urgent fundamental questions. While the LHC runs at energies around 13–14 TeV, LHC measurements access higher scales accessible in specific cases where the particle production involves intermediate undiscovered particles. The search for new particles is an important aspect of particle physics.

The link between particle physics and gravitation is one of the large open questions in physics in general, as it, to the best of our knowledge, would imply combining Einstein's theory of relativity with the quantum physics of the elementary particle world.

2.2.2 The LHC and the discovery of the Higgs boson

The LHC started its first preparation and feasibility studies in the 1980s. Still, construction began in strides when the previous accelerator in the same tunnel, the Large Electron–Positron collider (LEP), ceased operation in the early 2000s. In 2008 the LHC had its first collisions, and in 2010 the physics quality data started to be collected. Accelerators such as the LHC shoot bunches of tens to hundreds of billions of protons on each other. When bunches are shot at each other, the chance that a proton collision occurs is still relatively small. The number of these bunches inside the accelerator, how many protons each are filled with and how closely they can be packed, and how the beams are aimed at each other determine how many collisions occur. Improving each of these is extremely challenging and requires substantial and detailed work by the accelerator physicist teams at CERN. To implement substantial improvements in beam intensity work is necessary on the cutting-edge CERN magnet and accelerator system, which can take years. To facilitate data taking with similar conditions the LHC operation is organised in Runs. Run 1 took place between 2009 and 2012, Run 2 started in 2015 and lasted until 2018, and Run 3 is has started in 2022. For some reference, Run 2 provided approximately four times more data than Run 1, and Run 3 is expected to produce yet again double the datasets compared to Run 2. The substantial data increase is expected to occur in Runs 4 and later, at which point the so-called *High-Luminosity LHC* will produce datasets that are a factor ten larger than what was collected in the 15 years of Runs 1 through 3.

The Higgs particle was discovered by the ATLAS and CMS Collaborations in 2012. This elusive particle was proposed initially in 1964 to solve some of the inconsistencies in the theory that occurred when combining the weak and electromagnetic interaction (which is why particle physicists talk about the electroweak interaction). Generalising, the Higgs boson is the result of the same mechanism that gives particles mass. While the discovery of the Higgs boson solved the general question of the origin of mass, it raised more questions. For example, it does not explain why different particles have different masses or why the range of masses in the SM is as extreme as it is, ranging from feather-light neutrinos to enormously massive top quarks and anything in between (figure 2.2).

Fortunately, the mathematical machinery of the SM does predict very accurately how the Higgs particle will interact with particles with different masses, so an essential part of the study of the Higgs boson has moved from discovery to important consistency checks of the SM. This is particularly relevant as the SM predicts that the frequency that the Higgs boson decays to certain particles is directly connected to the masses of those particles. As the range of particle masses in the SM is so extensive and unexplained, these studies offer an important avenue into probing any consistencies in our understanding of the origin of mass. In addition, these decay probabilities are difficult to measure for lighter particles just because these particles will be produced less by the Higgs boson. At the Higgs boson discovery in 2012, the decay to Z and W bosons and to two photons[1] were the only sensitive channels. Since then, the Higgs boson decays to two tau leptons, two bottom quarks, and two muons has been observed. The associated production of the Higgs boson together with two top quarks allows studies of the interaction between top quarks and Higgs boson.

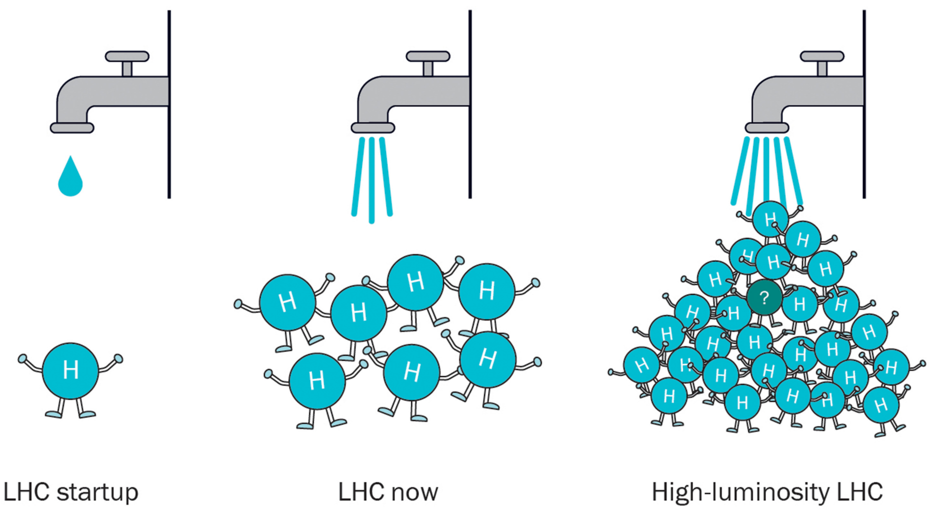


Figure 2.2. The number of Higgs bosons produced at the LHC has already increased by a factor of about ten since the discovery in 2012. But to get very large samples of Higgs bosons and examine if they are truly consistent with the Standard Model, the High-Luminosity LHC or even future colliders will be necessary.

[1] Hey, but the photon is massless, and the Higgs boson decays to mass, that's not right? Indeed, the Higgs boson decay to photons is a complicated exception, where the Higgs boson first decays to heavy particles such as top quarks and W/Z bosons. After that, those *virtual* particles create the two photons. These virtual particles become important in the next section.

After the discovery in 2012, it also became possible to measure the mass of the Higgs boson. At that point, it became increasingly clear that the value of the Higgs boson mass, about 125 GeV, or the approximate equivalence of the mass of an Iodine atom (that obviously contains many protons and neutrons), is not light enough to immediately confirm that new particles need to be present to make the SM work, but *also* not heavy enough to decidedly indicate that the SM is complete. It is in a metastable *grey zone* where the only way is to confirm which of those two is true. This makes the study for the study of the Higgs boson particularly relevant also in the context of, for example, understanding the evolution of the Universe and explanation for the age and evolution of the Universe [3].

Each of these observations at discovery has been refined over time. With more data, the connections between the Higgs boson and the heavier particles in the SM have now been confirmed to be consistent with the Standard Model's predictions to within, on average, about 10%. The majority of the lighter SM particles has not been confirmed to interact with the Higgs boson yet. To further improve and confirm if the Higgs mechanism is truly consistent with great accuracy, the ten-times larger datasets of the High-Luminosity LHC or even future colliders will be necessary.

Studying new rare decays of the Higgs boson (to the lighter quarks and leptons and invisible particles) is the highest priority for the LHC and one of the driving topics for future colliders. Another important topic is examining the already observed interactions of the Higgs boson as precisely as possible to identify if they also are consistent with the SM at the sub-percent level. Both these endeavours require a vast number of Higgs bosons. Producing these large samples of Higgs bosons is the goal of the High-Luminosity LHC and one of the metrics used to evaluate future colliders' scientific potential.

2.2.3 The challenge of breaking the Standard Model

One of the reasons why the SM is one of the most successful scientific theories is because it can make accurate predictions. The Higgs boson is only one of the SM particles, and internal consistency checks of the SM go much further than the precise examination of the Higgs boson properties and interactions described in the previous paragraphs.

One interesting aspect of the quantum physics of elementary particles is that it is possible to produce *any* particle for a very short time as long as that particle can interact with other particles. Due to the Heisenberg uncertainty principle, it is even possible to produce very heavy particles for an infinitesimally short time. These virtual particles cannot be directly observed but affect the behaviour of the other, observable, SM particles. This means that the effect of undiscovered particles can already be observed indirectly in the behaviour of well-understood particles. Figure 2.3 represents this as the known particles behaving as if they are shielding the undiscovered world. Still, it is a shield that can be understood by measuring smart and precisely. This specific quality of indirect measurements of particle physics was why there were already indirect constraints on the Higgs boson and top quark properties, many years before they were discovered by direct production.

These subtle changes in behaviour can be observed if the properties, kinematic behaviour, and relative production rates are measured accurately. A massive

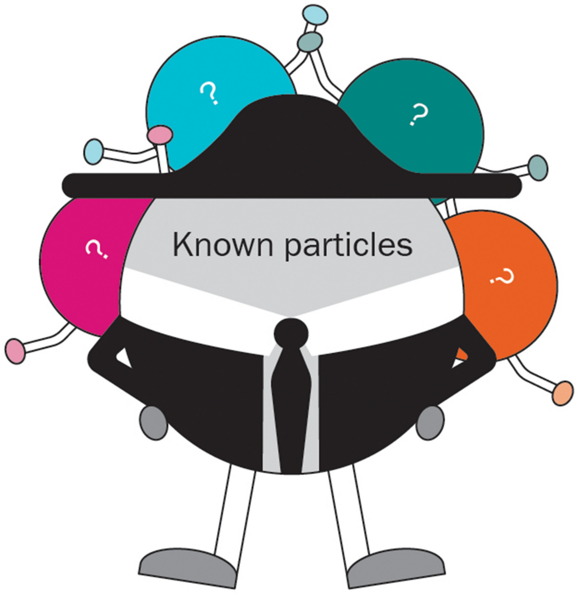


Figure 2.3. The detailed behaviour of known particles can be used to indirectly study undiscovered particles.

challenge for these precision measurements is that when the SM is challenged at such precisions, the uncertainties due to the experimental knowledge, the simulations, and the theoretical calculations that are providing the prediction become relevant. Testing and identifying the subtle changes that virtual particles introduce is another one of the main challenges in particle physics, as these precise measurements require combinations of extremely detailed understanding of the detectors, backgrounds, and SM predictions, together with the humongous datasets produced by extremely high-intensity colliders.

A great example of the power and challenges of these precision measurements is in the extremely challenging measurement of the μ g-2 experiment at Fermilab [4]. This experiment measures the muon's magnetic moment, a property that can be extremely precisely predicted by the SM. The magnetic moment is the strength of a muon's interaction with a surrounding magnetic field, where virtual particles (so that means at least all the other SM particles, and if other particles exist, those as well) modify the behaviour of the muon. This makes this experiment a great test for telltale signs of new particles and forces. To do so, the knowledge of the detectors and particularly an unprecedented knowledge of the exact value of the magnetic field that the muon is moving in is essential. In the spring of 2021, the first result from this experiment confirmed previous measurements of the magnetic moment are off from the SM prediction by a precision of about one part per million. The eventual goal of the experiment is to test the Standard Model's predictions of this value by measuring the precession rate experimentally to a precision of one-tenth part per million, and new data is currently being collected and analysed to do so. However, it is crucial to be aware that the measured numbers have to be compared to a SM prediction.

Predicting something that accurately is not easy, and the theoretical calculations by hundreds of theoretical physicists are another aspect of this test that receive serious scrutiny before it is definitive that there is an inconsistency and the SM is incomplete and in need of revision.

There are similar examples from the LHC; recently, the LHCb experiment measured unexpected discrepancies between the decays of beauty and charm quarks to the different leptons. In the SM, the electron, muon, and tau particles all behave identically except for their mass. Only virtual particles that interact differently with the different leptons can easily be used to explain such discrepancies. The ATLAS and CMS collaborations, and in Japan the Belle2 collaboration, are now studying similar decays to shed light on whether these flavour anomalies are reproducible. Results are expected in the coming years, and the study of these flavour anomalies is a very actively developing field of study within experimental particle physics. It is important to note that several boundary conditions need to be achieved to make progress in these precise measurements. For example, to understand the difference in behaviour between two similar particles, such as the muon and electron, the experimental measurement of both particles needs to be exquisitely understood. In addition, not all experiments were specifically designed to measure these signatures. Still, with the fast improvements in both detector knowledge and statistical analysis techniques such as deep learning, it turns out even *general-purpose detectors* such as ATLAS and CMS can be expected to make contributions to the study of these flavour anomalies.

To achieve these precision tests of the SM, two other requirements are essential that will provide challenges in the near future: the simulation that in such analyses is used needs to be accurate enough to represent the detectors and SM to the best degree possible, and the theoretical predictions that these precise measurements will be compared to will need to be of similar precision as the experimental results. For both of these, substantial difficulties will need to be overcome. In the case of simulation, this is partially a scientific computing challenge, as the number of simulated collisions will need to increase substantially, and this will put a substantial strain on the software and storage solutions used by the collaborations. In addition, such detailed simulation requires that no corners be cut as far as the modelling of tiny differences in different detector components and regions, making the simulation substantially slower. There currently are multiple innovative efforts ongoing that try to address these challenges, both as far as using different hardware, using smarter software solutions, including modern computer science and data science techniques that will be tested to their limits on data that may reach the size of up to exabytes. The simulation also needs to be sufficiently realistic as far as the computation of the SM, implying that substantial improvements will need to be made in the precision of the physics in the simulation, which is now typically performed at leading order or maybe next-to-leading order precision in perturbation theory. This challenge to increase in accuracy even more affects the SM calculations, which will need to be performed as a reference that measurements use to compare in hypothesis tests. In recent times, multiple precision measurements that implied discrepancies with respect to the SM were proven to be correct but still consistent with the SM when it was calculated to higher precision. To reliably and effectively improve the higher-order calculations, substantial strides in theoretical physics, both as far as computational techniques and mathematical methods, will be necessary.

2.2.4 The challenge of searching for new particles

There is obviously a more direct way to prove that the SM is incomplete, and this is to identify new particles in the collisions at the LHC. There are also a substantial number of smaller, dedicated experiments that may not even need a collider, and that aim to identify such signatures; the direct and indirect experiments looking for dark matter and searches for axions and sterile neutrinos are good examples.

Before the LHC startup, there was overwhelming positivity that new particles would appear as soon as the accelerator was turned on. By now history (and the Standard Model) has taught physicists some more humble attitudes. Twelve years after the first collisions of the LHC there are no significant signs of new particles, and this is driven partially by the fact that about 5% of the lifetime data has been collected, but more importantly by the fact that the focus up to now was mostly on the easy, *low-hanging fruit* new physics scenarios. It is, however, useful to consider how the searches programme, which is responsible for about half of the research output of the ATLAS and CMS collaborations, is performed and what the current and future challenges are for the direct search for new particles.

Many of the new physics predictions for the LHC are relying on assumptions that are inspired by the general idea of supersymmetry. Supersymmetry relies on the introduction of a new symmetry to the SM, which implies that every known particle has a partner particle. The idea of supersymmetry is definitely not a new one, but the Higgs boson mass of 125 GeV is very consistent with the preference for low mass Higgs bosons in even the most general versions of supersymmetric models. As of today, no signs of supersymmetry have been spotted in the LHC data, and the experimental searches are now focusing on the more challenging scenarios where supersymmetric particles may still survive undetected.

It is important to realise that the number of possible signatures that supersymmetric models is so large that after initial data taking, the LHC experiments focused on a strategy where signatures were investigated instead of specific models. In general, the number of varied parameters is substantially reduced and only very simple *Minimal Supersymmetric Models* are considered. The *simplified model spectra* of these models that are being used tend to assume that there is only one supersymmetric particle that can be produced at the LHC (instead of the full spectrum of a doubling of the number of particles). In addition this single supersymmetric particle then only decays only in one (or few) ways, creating a very specific signature that can easily be searched for. This well-defined signature does come at the sacrifice that the connection to realistic collections of supersymmetric particles are produced together. This means that when statements are made regarding 95% confidence level excluded mass ranges of supersymmetric particles (e.g., for the gluino, the supersymmetric partner of the gluon), the constraints will be given for a 100% production of gluinos with no other particles being accessible, and a 100% assumption on the gluino decaying to one specific particle (e.g., a gluon and an undetectable neutral particle). If these assumptions are loosened, the excluded mass range of gluinos may be reduced substantially and many scenarios where the particle decays to multiple different particles at some probability are still possible and not considered. The interpretation challenge of simplified models is important to be considered when drawing broad conclusions as far as whether supersymmetry is still accessible at the LHC; particularly in summaries this fact is regularly not neglected in the discussion of supersymmetry as still being a viable model to use for inspiration of searches for new physics.

Even in the rudimentary world of simplified model spectra there are still substantial fractions of supersymmetry scenarios for which the LHC does not yet have sensitivity. These tend to come in two general categories: scenarios where the particles for some reason are extremely rarely produced (these would mostly be the supersymmetric partners of the W, Z, and Higgs bosons and leptons), or the considered masses are such that the supersymmetric particles are kinematically suppressed (this is commonly referred to as *compressed spectrum* supersymmetry).

Beyond the simple supersymmetry predictions and searches a whole diverse world of *exotic* other models opens. There are many possible ways to combine multiple supersymmetric extensions, also with extra dimensions, with gravity and so forth. The famous representation in figure 2.4 is a very good way to represent the huge number of different scenarios that non-minimal supersymmetry models predict. There are many different scenarios that predict new particles that would create resonances in di- and tri-particle invariant mass in the SM continuum; for example, these tend to be related to various models predicting new bosons, extra dimensions, and string theory-inspired new symmetries that introduce one or more new mediator particles.

There are as many extensions to the SM as there are enthusiastic theoretical physicists, so at an experimental level the challenge is typically to identify a signature

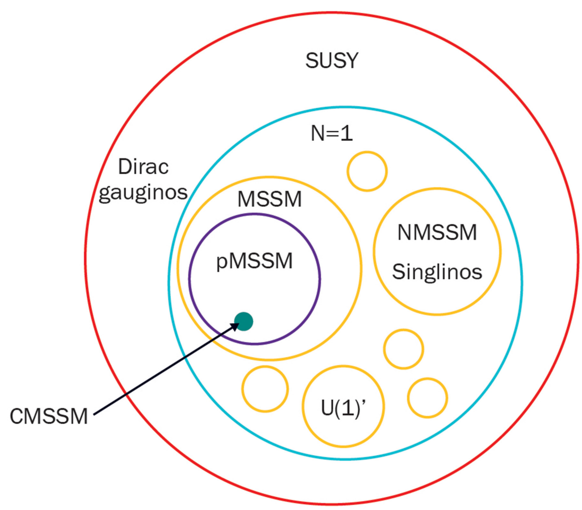


Figure 2.4. The green dot represents what the typical LHC experiments consider when searching for supersymmetry. Obviously many other scenarios are possible, and many of these are considered at the LHC as well but are grouped under the exotic nomenclature [5].

that does have either a very low or very well-understood background from the SM. However, one of the weaknesses of searches for such resonances is that they are driven predominantly by the collider energy, so additional data due to increased accelerator intensity do not necessarily increase the mass sensitivity of searches for resonances but do continue to provide sensitivity to smaller and smaller cross-sections. The topological choices tend to focus on signatures such as:

- New resonances, where practically every combination of jets, including flavour tagged jets, are considered, as are all charged leptons. Such resonances on backgrounds that are typically following simple power laws are extremely striking.
- Other striking signatures, such as deviations in event mass and energy distributions. These are predicted by a wide range of new physics models, such as microscopic black holes, and undiscovered new particles at energies not directly accessible by the LHC energy that do modify the behaviour or the amount of SM particles produced in general.
- Searches for invisible objects, such as direct production of dark matter particles, tend to focus on the production of an additional particle together with the invisible particles. The visible, from SM processes such as initial state radiation, in that case recoil against invisible momentum. Again, such mono-jet signatures are by now also examined with jet flavour identification tools and for mono-W/Z/Higgs bosons with jet substructure techniques.
- New fermions that decay in some way to known SM particles. Examples are new quarks (e.g., vector-like quarks), but also leptoquarks, particles that have both leptonic and strong interacting properties and that can be used to explain some tantalising deviations in SM precision measurements in the flavour sector, by, for example, the LHCb experiment.

As the LHC datasets increase, new signatures continue to become accessible. This is particularly relevant in searches for new physics theories that might have suppressed couplings in the resonance or missing energy signatures, but would create deviations from the SM in other signatures that are so rare that at LHC startup there was not enough data to examine them. With the ever-increasing improvements in the understanding of the detector, some other searches are also being developed, using techniques that were not even predicted when the LHC was designed.

2.2.5 Leaving no stone unturned

In recent years the searches at the LHC have become more creative. This development is partly driven by the lack of observed excesses, innovation on the experimental side, and the next generation of reconstruction algorithms becoming available. Improved understanding of the detectors has led to opportunities to reconstruct jets and their structure better. It is now possible to identify hadronic decays of Higgs bosons, W and Z bosons, and top quarks inside jets. These substructure tools have made enormous leaps in the last years, are now widely available, and perform at similar usability levels as, for example, b quark jet flavour tagging. This innovation at the reconstruction level has created a whole new class of possible analyses. ATLAS and CMS have developed effective research programs using substructure tools to study the SM precisely and search for new particles that decay to Lorentz-boosted objects.

Another avenue that has recently become very fruitful is searching and identifying long-lived signatures for physics beyond the SM. The LHC experiments were designed to identify particles that are produced close to the collision point. Still, over the last years, it has become clear that many new physics models have a parameter space where particles are limited in their decay somehow (usually either through phase space or coupling considerations) which creates displaced signatures. Examining these non-prompt signatures is possible as the experiments can detect such decays but previously rejected them. Experimental challenges tend to be driven by the fact that the backgrounds can be much more challenging to study or identify potential background sources. The methods to compare simulation to data become more complex. However, the diversity of physics models and signatures that can be probed by studying long-lived signatures is so bountiful that this has by now been established as a very effective LHC program, including links beyond the main LHC experiments. Long-lived particles have also inspired an LHC-wide collaborative community that includes a very effective collaboration between theoretical and experimental physicists [6].

The understanding of the LHC detectors is continuously improving still, and this also continually opens new potential signatures. Currently, the experimental innovation beyond substructure and long-lived signatures focuses on improving background rejection for lower-energy lepton signatures and improved performance in high jet multiplicity regions. Both are only marginally explored up to now; the existing methods focused on the low-hanging fruit of leptons coming from electroweak boson decays for which the detectors were designed. Still, the previously mentioned flavour anomalies have motivated strong drives for the experiments to go as low as possible in transverse momentum and examine more and more rare signatures at high multiplicity that were not sufficiently populated in smaller datasets to study in more detail than would be possible in simple counting experiments.

At hadron colliders, traditionally, only a small fraction of the data is analysed. At the same time, the rest of the data is rejected to save disk space. Strategies to collect more difficult-to-access data are also being implemented. The LHCb experiment plans to implement an online selection strategy where almost all of the data will be selected by reducing the individual event size so that a larger fraction of events are saved and analysed [7]. Similar strategies are also being implemented in CMS and ATLAS, such as the *scouting* and *parking* efforts by CMS [8].

2.2.6 Beyond the general-purpose experiments

The two giant experiments at the LHC, ATLAS and CMS, are by no means the only endeavours to break the SM. Even the collisions at these so-called general-purpose experiments are examined by smaller experimental collaborations that aim to detect dedicated signatures, such as potential detection of very long-lived particles and detection of anomalous behaviour of neutrinos at tens or even hundreds of metres from the collision point [9–11]. In addition there are multiple very forward detectors that study LHC collisions where the protons only barely collide, which provide important measurements and potential tests of particularly the strong force [12–14]. At the LHC ring, there are two more collision points which host dedicated experiments, such as the already mentioned LHCb experiment that was originally predominantly designed to study the properties of beauty and charm quarks, including any potential differences between quarks and antiquarks and other properties of bound state of these quarks. LHCb has since changed its development into a broad and internationally leading collaboration that also studies many of the same questions addressed by ATLAS and CMS, often using complementary techniques and kinematic properties. At other colliders, similar experiments are coming online, such as the Belle2 experiment that is ramping up to perform an essential precision physics programme in the heavy flavour sector using electron–positron collisions at the SuperKEKB accelerator complex in Japan. The expected precision of measurements that are expected to be performed soon by Belle2 will create a very dynamic period in the study of heavy flavour quarks, particularly as many of the current results by LHCb that show some tension with respect to SM predictions can be performed independently at this facility.

2.2.7 Beyond the LHC

Beyond collider physics, there is a wealth of other particle physics experiments that also examine the SM. The study of neutrinos, the most abundant particles in the Universe, has a vibrant program that complements the study of the SM at colliders. Some may argue that neutrino physics has already moved to physics beyond the SM, as it is clear the particles have a mass that is not the default in the SM. Open questions such as the number of neutrinos, their mass hierarchy, and the possibility of violation of the tenets of the SM such as CP violation and the potential unification of forces can all be examined by studying the physics of neutrinos.

There are experiments where neutrinos are studied coming from particle accelerators, from reactors, and from the cosmos. Neutrinos are notoriously difficult to detect, so most of these experiments look for very few interactions and use detectors and experimental underground locations where almost no background is present. Producing such sensitive detectors creates unprecedented experimental challenges. In the future, new experiments will come online that will be able to answer many of the open questions. For example, the Deep Underground Neutrino Experiment (DUNE), a leading experiment for long-baseline neutrino physics and proton decay studies that uses a neutrino beam between Fermilab, Illinois and Sanford lab, South Dakota in the USA. In recent years many European particle physicists have joined these efforts. The goals are to use neutrinos to study similar underlying questions as accelerators, such as the origin of matter, unification of forces, and searches for new physics. Some of the challenges in neutrino physics are different than at colliders. Still, many of the final physics goals are shared. In addition, many large neutrino detectors have the sensitivity to study particles coming from the cosmos, such as the potential observation of thousands of neutrinos from supernovas and other transient astronomical phenomena.

The hunt for a solution to the dark matter puzzle is also becoming competitive, with the Xenon-based experiments XENONnT and LUX-ZEPLIN both starting their science runs recently [15, 16]. Both experiments rely on extremely pure large quantities of liquid Xenon atoms that should create very challenging and tiny signals when dark matter particles interact with them. These experiments rely on extremely low background environments, so are housed deep underground and aim to have very pure conditions so only a few dark matter interactions could be enough to establish a potentially paradigm-shifting discovery. A large number of smaller experiments that aim at specific dark matter signatures are also running, and more are in development. In combination with the Xenon-based experiments, dark matter programmes at neutrino experiments, and the dark matter signatures that could potentially be visible at the LHC experiments, it is likely that if dark matter is a particle, it will be cornered in the coming years.

2.2.8 Challenges

There are also technical and theoretical physics challenges that are vital to make progress beyond the challenge of identifying potential new physics that were described in all its facets previously. These are partially technical, such as the genuinely overwhelming data volumes that the LHC will produce in the future and the improvement of detector technologies to robustly detect these large volumes, but also address the more detailed understanding of the SM in enough detail to continue testing it with better accuracy.

2.2.9 Challenges: the data challenge

In the coming years the amount of data produced by particle physics experiments will increase exponentially. In addition there are challenges in producing the simulated samples that are necessary to analyse this data. To successfully analyse the data produced by the high luminosity LHC, the computational tools available will have to improve substantially, both as far as being able to process data volume, data processing speed, and simulation methods.

Each of those poses unique challenges that also have links to other big challenges in broader scientific context and in society. After all, the big techniques from particle physics have by now made large changes in other scientific fields possible, and with the large majority of the world's population using social media, the data volumes produced by humanity are, just like the data produced in particle physics, only expected to continue growing. It is also clear that the evolution of computing technology will alone not cover the increase in data resource needs. Improvements in computational techniques including modern artificial intelligence, but also potential new hardware developments and new computational methods like quantum computing and quantum sensing are expected to open opportunities to reduce the expanding cost of collecting, processing, and storing the data.

2.2.10 Challenges: the accelerators of the future

Since the dawn of particle physics in the early 1950s, particle physicists have relied on colliders that collide either (anti)protons, or electrons and positrons. Historically, these machines have played different roles; due to lack of synchrotron radiation, hadron collider machines can reach typically higher energies but with substantial backgrounds that create experimental challenges. In contrast, the lepton machines are extremely powerful at performing precise measurements with substantially reduced interference from other SM processes. One crucial piece of knowledge that the LHC has taught us is that with modern analysis techniques, hadron machines can also be partly used as precision machines. This precision functionality is limited chiefly to measurements that test the SM through ratios, where many uncertainties are reduced and the SM tends to be well understood, and the case where the backgrounds are understood to such a level that SM processes can be very precisely measured.

The particle physics community is currently performing an international procedure examining the potential, feasibility, and support of future colliders. Broadly, multiple machines are under consideration, divided into two categories: Higgs boson factories and search machines. The Higgs boson is still hiding many of its properties and has only been tested up to about 10% in its most sensitive scenarios and will be tested up to a few percent at the end of LHC running. Such a machine has already been deemed the highest priority by the European physics community [17]. Questions, however, remain as to how such a machine will be realised. It is important to have an international consensus and unanimous support on what machine be built and potentially where. The huge number of different options, including even a muon-antimuon collider, makes this decision making a daunting task. One of the considerable organisational challenges of the field will be to achieve the completion of such a machine and a collider physics program that will reach well into the 21st century.

2.2.11 Challenges: timing detectors and detector innovation

As the data volumes increase and the collisions become more busy, it becomes also more and more challenging to precisely detect interesting collisions. A vibrant and innovative instrumentation effort is part of the particle physics community, and creative new ideas that will allow high fidelity detectors that will perform well under an ever-increasing number of particles in a single collision.

Detector techniques that use not only spatial or energy measurements are of particular promise, as with accurate time detection it is easier to disentangle otherwise overlapping particles that did arrive tens of nanoseconds apart. In the future, such timing technology is expected to play an important role for detecting the collisions at future colliders, and the collaborations at the LHC are already developing first versions of these detector technologies that will be available for high-luminosity LHC.

There are of course many challenges in the developments of all these detectors; particle physics experiments have unique needs as far as the uniformity of the individual detection elements but also as far as radiation hardness and material budget (so indirectly power-usage). Many of these aspects are still actively being researched and are expected to continue to improve in the years to come. Innovation and strengthening of existing instrumentation efforts are critical to maximize the scientific outcomes of future accelerators; after all, the investment of these machines makes it paramount that powerful detectors are present in these precious collisions, and to push forward the innovation in non-accelerator-based experiments as well. In addition, a strengthening of the link between detector innovation and collaboration with industry is one that is expected to become more and more important in the future.

2.2.12 Challenges: the theory challenge

With ever-increasing experimental data, challenging the SM more and more precisely, it becomes also ever-more important that the SM can be calculated more accurately so that the hypothesis testing as the classical experimental method described can continue to flourish. Theoretical physics calculations are an essential part of this endeavour, where potential new lines of research can be identified and the precise tools are made that allow to fully exploit experimental data.

A broad programme of theoretical research that covers the full spectrum of particle physics is essential to achieve these goals. This includes both exploratory new ideas, foundations in mathematical physics and links to cosmology, astroparticle physics, and nuclear physics. As the calculations become more and more complex, the tools to do so also become an important part of the research. For example, to fully exploit the study of the Higgs boson at the high-luminosity LHC or at future colliders, the experimental precision is expected to very quickly not be the most important uncertainty, and uncertainties due to theoretical assumptions that are in principle reducible through challenging intellectual innovation will become dominant for many scenarios.

2.2.13 What is the ultimate status of the constants of Nature?

Particle physics is a field where that has the sensitivity for probing the most basic concepts of Nature. To do so, solutions to complex challenges both on the detector and on the accelerator side are necessary. Further challenges in computing, simulation, and data processing are also creating challenges to make full use of the available data. The field benefits from a fertile cross-pollination between theoretical physics and experiment. The apparent weaknesses of the SM, in combination with a wealth of potential extensions of the SM, put the fields in an exciting but very precarious position. Inconsistencies of the SM in the neutrino sector, the dark matter puzzle, the fact that the SM may very well be incomplete at higher energies or is at least not able to describe all processes reliably. These and many other arguments all point to the fact that the SM is incomplete and creates the clear potential for new particles to be discovered at the LHC in direct production or indirect measurements of SM properties.

On the other hand, the Standard Model's predictions tend to agree to very high accuracy in most measurements, and direct searches for low-hanging fruit at the start of the LHC data taking have not produced any convincing sign for a new particle beyond the discovery of the Higgs boson in 2012. This creates a potentially exciting but extremely challenging dichotomy where the focus has moved from predictions to searching for deviations for unpredicted (or at least, not overly model-dependent) signatures. The transition from a more theory-prediction-oriented field into an experimentally driven research program is currently in full swing in this field. Only time will tell what the future has in store for particle physics, but considering the wealth of experiments, the large number of extremely fundamental questions that are corroborated with data and that the SM cannot answer, and the fact that the flagship LHC has only collected one tenth of its data, the future is very bright.

2.3 The origin of visible matter

Angela Bracco[1]
[1]Università degli Studi di Milano, Milano, Italy and INFN

Nuclear physics is the science of the atomic nucleus and of nuclear matter. The atomic nucleus is the dense core of the atom and is the entity that carries essentially all the mass of the familiar objects that we encounter in Nature, including the stars, the Earth, and indeed human beings themselves. The quest for the origin of visible matter requires very good knowledge of nuclear matter and reactions under a large variety of conditions up to the most extreme ones. The physics of nuclei, starting from the hot dense soup of quarks and gluons in the first microseconds after the Big Bang, up to the formation of the chemical elements drives intense and innovative research in both experiment and theory.

The experimental investigations of nuclear properties require the use, and the continuous development, of an arsenal of experimental techniques and require different types of research infrastructures and theoretical approaches. The technical developments for the nuclear research are in large part also applied to the benefit of our society.

The distinctive feature of nuclear physics research is the breadth of the topics it covers. Here these topics are grouped under two titles, 'The fundamental description of the heart of matter' and 'Nuclear structure and reactions for the origin elements'.

The progress made in these research areas and the future perspectives are briefly presented and discussed. It is clear from the short narrative in the following pages that the physics of atomic nuclei and their constituent components is rich, varied, and extremely complex at many levels. Moreover, research in the next decades will be challenging and hopefully full of surprises.

2.3.1 Introduction

Understanding the origin of visible matter is a very complex but fascinating problem that drives experimental and theoretical research involving strongly the physics of nuclei.

From the hot dense soup of quarks and gluons (the latter carrying the interaction among quarks) in the first microseconds after the Big Bang, through the formation of protons and neutrons to the evolution of the chemical elements, the physics of nuclei is fundamental to our understanding of the Universe and, at the same time, is intertwined in the fabric of our lives. Nuclei also constitute a unique test bench for a variety of investigations of fundamental physics, which in several cases are complementary to elementary particle physics, but use different approaches. Nuclear physicists and chemists are creating totally new elements in the laboratory and producing isotopes of elements that have previously only existed in stellar explosions or in the mergers of neutron stars. For nuclear physics research new tools like accelerators and detectors are developed that often find broad applications in industry, medicine, or national security.

The overarching goal of present-day nuclear physics research is to explain the origin of visible matter and for this one needs to study in-depth several different physical phenomena involving nuclei and their constituents (see references [18–20]). Some central questions in this connection are:

- How is the mass generated and what are the static and dynamical properties of nuclei and of their constituents?
- How does the strong force in nuclei emerge from the internal structure of its constituents, the protons, and neutrons?
- What are the properties of nuclei and strongly interacting matter as encountered shortly after the Big Bang, in catastrophic cosmic events, and in compact stellar objects?
- How does the complexity of nuclear structure arise from the interactions of protons and neutrons in the nuclei?
- What are the limits of nuclear stability?
- How and where in the Universe are the chemical elements produced?

The experimental investigations addressing these questions require the development and skillful use of a variety of experimental techniques and theoretical approaches. Consequently, there are different types of research infrastructures where nuclear research is carried out. The distinctive feature of the research in nuclear physics is the presence of different research areas which can be grouped together under two titles: 'The fundamental description of the heart of matter' and 'Nuclear structure and reactions for the origin of elements'.

In Europe there are several laboratories (see, e.g., references [21–34] and references therein) that are either mainly devoted to nuclear physics or are carrying out a rich scientific program in this area. In figure 2.5 'Open access' laboratories are indicated in the map of Europe while other smaller sized, often university-based, laboratories are not shown. These smaller laboratories are good niches for specific timely activities and contribute very much to R&D and training. The scientific activities in nuclear research are distributed over different facilities which are not necessarily placed at the same location (see also references [35–40] and references therein). These facilities conduct research at the forefront, have minimal overlap among themselves, and are well coordinated via European projects. In addition, in Europe, for more than 30 years, an expert committee, NuPECC, including a representative of the Nuclear Physics division of EPS, has overseen research programs and prepares strategies for the field every 5–7 years, which are reported in published Long Range Plans.

2.3.2 The fundamental description of the heart of matter

It was just about hundred years ago that Lord Rutherford uncovered the existence of the proton as one of the basic building blocks of the atomic nucleus, and a decade later the other major building block, the neutron, was discovered. After that, many features of the force that binds protons and neutrons (the nucleons) within an atomic nucleus were established, and because of its strength this force is called 'strong

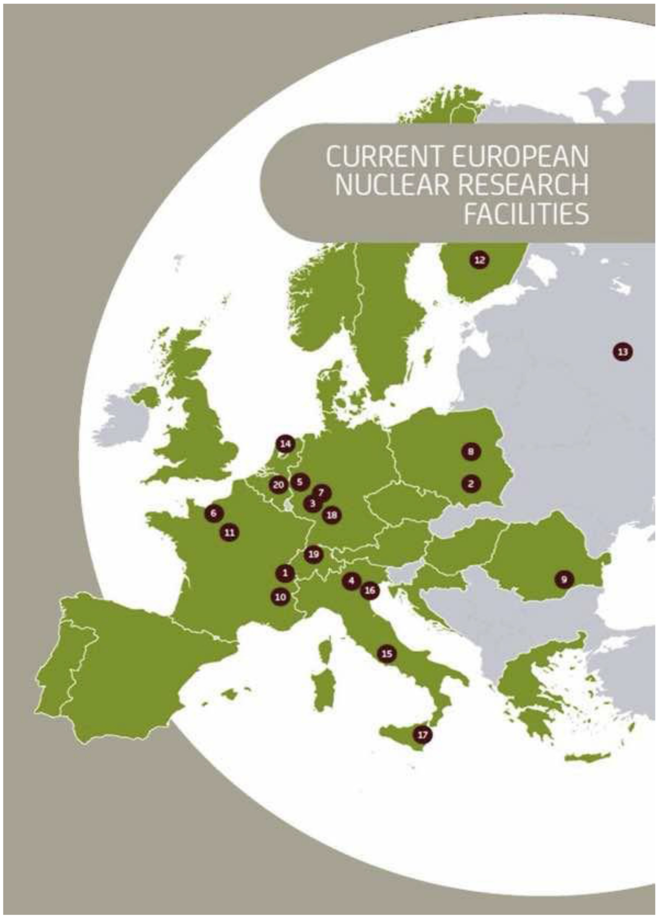


Figure 2.5. The current 'open access' nuclear research facilities in Europe. Other existing smaller size facilities, which are mainly university based, are not indicated in this map, although they contribute to the European research in this field, particularly in the area of applications. The facilities in this figure are: (**1**) CERN—European Organisation for Nuclear Research (ALICE, AD, COMPASS, and ISOLDE), Genève, Switzerland; (**2**) CCB—Cyclotron Centre Bronowice (IFJ PAN) Kraków, Poland; (**3**) ELSA—Electron accelerator, Bonn, Germany; (**4**) ECT*—European Centre for Theoretical Studies in Nuclear Physics and Related Areas, Trento, Italy; (**5**) FZJ—Forschungszentrum Jülich (COSY and HPC), Jülich, Germany; (**6**) GANIL—Grand Accélérateur National d'Ions Lourds (SPIRAL and SPIRAL2), Caen, France. (**7**) FAIR and GSI—Helmholtzzentrum für Schwerionenforschung GmbH, Darmstadt, Germany; (**8**) HIL—Heavy Ion Laboratory, Warsaw, Poland; (**9**) IFIN-HH Horia Hulubei—National Institute of Physics and Nuclear Engineering, Bucharest, Romania; (**10**) ILL—Institut Laue–Langevin, Grenoble France; (**11**) IPN—Institut de Physique Nucléaire, Orsay, France; (**12**) JYFL—Accelerator Laboratory, University of Jyväskylä, Finland; (**13**) JINR—Joint Institute for Nuclear Research, Dubna, Russia; (**14**) KVI-CART—Kernfysisch Versneller Instituut, Groningen, The Netherlands; (**15**) LNF-INFN—Laboratori Nazionali di Frascati of INFN, Frascati, Italy; (**16**) LNL-INFN—Laboratori Nazionali di Legnaro of INFN, Legnaro, Italy; (**17**) LNS-INFN—Laboratori Nazionali del Sud of INFN, Catania, Italy; (**18**) MAMI—Mainzer Microtron, Institute for Nuclear Physics, Mainz, Germany; (**19**) PSI—Paul Scherrer Institute, Villigen, Switzerland; (**20**) SCK*CEN—Mol, Belgium. These are the facilities with scientific programs dealing with the topics addressed in this chapter 'Origin of visible matter'. Adapted from http://www.nupecc.org/pub/lrp17/nupecc_lrp_brochure_2017.pdf [19] courtesy of NuPECC (nupecc.org).

force'. Of particular interest is the strong interaction of nucleons that are relatively close to each other, so that their internal structure might be expected to come into play. This 'short-range interaction' is responsible for the fact that most nuclei have roughly the same density, which ultimately determines the stability and size of neutron stars. While nucleons mostly move at moderate speeds inside nuclei (up to 30% of the speed of light), their short-range encounters can impart significantly higher momentum to each of them and make them move swiftly in opposite directions. Such high-momentum correlations, long predicted by nuclear theory, provide an excellent way to study the short-range part of the nucleon–nucleon interaction and have now been observed.

In spite of the progress made to establish the features of the nucleon–nucleon interaction and of the multitude of other strongly interacting particles (named hadrons) a fundamental understanding of the underlying laws of physics was missing until quantum chromodynamics (QCD) was recognized as the fundamental theory governing nuclear matter.

2.3.2.1 The structure of nucleons and other hadrons and the theory of quantum chromodynamics

According to the theory of QCD protons, neutrons, and all other hadrons are made from quarks, their antimatter siblings (antiquarks), and particles called gluons, which carry the force that binds quarks to each other (see, e.g., references [41–45]). Protons and neutrons can be thought of as containing three so-called valence quarks, immersed in a shimmering cloud of quarks, antiquarks, and gluons, all continually winking into and out of existence according to the laws of quantum mechanics. Quarks and gluons are the building blocks of all hadrons. No single quark or gluon, however, can be observed in isolation and thus one says that they are confined within a hadron. This implies that any process by which one tries to rip a quark out of a proton or neutron makes new hadrons, without ever isolating a single quark. Therefore, the strong interaction between quarks decreases at short distances and increases at large distances (in contrast to quantum electrodynamics where the interaction has an opposite behavior), and at very short distances the quarks are free (this property is named 'asymptotic freedom'). Gluons carry the force between quarks in much the same way that electromagnetic forces are carried by the photon, but in contrast with photons that do not interact between themselves, gluons do interact with each other.

It turns out that these unusual features of QCD imply that the intrinsic mass of the three valence quarks in the nucleons gives rise only to a small fraction of the proton and neutron masses and hence gluons also contribute to it. Since protons and neutrons account for nearly all the mass of atoms, almost all of the mass of the visible matter in the Universe is due to these seemingly exotic QCD effects. And while these general features of nuclear matter are well established, a detailed understanding of how this originates from QCD is only now emerging. Understanding the structural complexity of protons and neutrons in terms of quarks governed by the laws underlying the QCD theory is one of the most important challenges facing physics today, in spite of the impressive progress made so far.

Recent advances in computational power now allow for precise calculations of the masses which at the same time predict novel exotic particles to be identified experimentally.

The significant increase in the capabilities of large-scale supercomputers in conjunction with the enormous progress made in developing efficient simulation algorithms have allowed huge increases in the accuracy of first principles predictions (e.g., of the mass difference between the proton and neutron). Nevertheless outstanding challenges remain, among which is the understanding of how the proton gets its magnetic property, namely its spin (due to a rotation of a particle on an intrinsic axis), which is measured to have the value 1/2 (in units of $\hbar$, which is the reduced Planck constant of fundamental importance in quantum mechanics). Figure 2.6 is a pictorial representation of possible mechanisms that produce the spin of the proton. The individual spin of the valence quarks inside the proton accounts for only 1/3 of the spin of the protons. Theorists predict that gluon and a 'sea' of transiently existing quarks also contribute together with magnetic effects due to the orbital motion of the quarks.

Only with new measurements and new calculations of how quarks and gluons can combine into new particles with quarks of different types one can learn more regarding these fundamental questions. Because there are six quarks, the number and variety of hadron permutations possible is large, and laboratory experiments are devoted to creating and studying these combinations. In addition, one needs to study the spectra of hadrons, their internal structure and interactions—particularly those composed of the heavier quarks. Such hadrons are rare and exotic, yet tell us a lot about how the strong force works.

Similarly to atomic spectroscopy, which has been a crucial tool for studying the electromagnetic interactions that bind electrons to the nucleus, hadron spectroscopy experiments are carried out to illuminate the interaction that binds quarks. While the proton and neutron contain the two lightest valence quarks (up and down), other hadrons composed of these light quarks or of more massive quarks (strange, charm, bottom, and top) and their corresponding antiquarks can be created in energetic collisions produced by particle accelerators. Once produced, these hadrons decay promptly, allowing one to measure only a few of their properties, such as their mass, charge, and angular momentum.

Studying patterns of hadrons classified by their properties provides insight into the theory of QCD. The observed patterns of states suggest that almost all hadrons

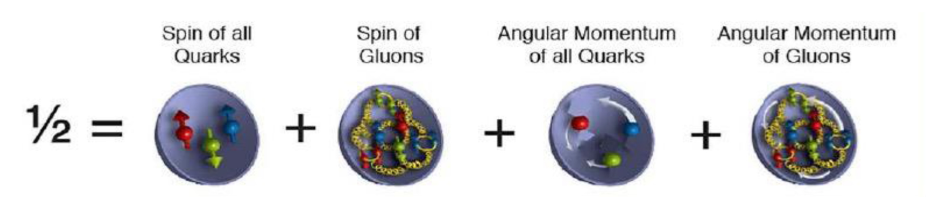


Figure 2.6. Schematic view of the proton and the potential contributions to its spin. Adapted from https://science.osti.gov/~/media/np/nsac/pdf/2015LRP/2015_LRPNS_0918155.pdf [20] courtesy of U.S. Department of Energy (energy.gov).

fall into two classes: baryons that contain three valence quarks, like the proton and neutron, and mesons that contain a valence quark and a valence antiquark. In principle, the QCD theory allows also hadrons made of two quarks and two antiquarks (tetraquarks), four quarks, and an antiquark (pentaquarks), and infinitely many other more complex configurations. Recently, physicists studying the spectrum of heavy mesons formed with charm and bottom quarks have uncovered evidence that supports the existence of tetraquark and pentaquark hadrons. Understanding the properties of these new states of QCD may provide insight on why nature prefers hadrons with relatively few quarks.

It is clear that the theory of QCD has to be tested at different energies so that experiments addressing the present open problems in this research area are conducted not only at the LHC accelerator at CERN but a large fraction of experiments are performed using accelerators providing particle beams with energy from few GeV to few hundreds of GeV. These particle accelerators are located at CERN and in other different places in Europe (see figure 2.5). In addition, several European research groups, of quite a large size, collaborate and give important contributions to experiments carried out in laboratories outside Europe, in particular at the Thomas Jefferson Laboratory in Virginia (USA).

The type of beams used for experiments addressing the problem of the strong force and of the structure of hadrons are hadron particles (protons, pions, and kaons), but also electrons and muons that interact via the electromagnetic force.

2.3.2.2 Symmetries and fundamental interactions

The presently known fundamental interactions governing Nature and the Universe from the largest to the smallest distances display symmetries and symmetry-breaking effects. High-precision studies allow tests of our understanding of Nature that are complementary to experiments at the highest energies and sometimes offer higher sensitivities to new effects beyond the SM of particle physics.

Nuclear Physics has played a major role in finding and establishing the laws that govern physics at the most fundamental level, with the nucleus being a very useful test bench for the weak, electromagnetic, and strong interactions. One of the most notable examples is the maximal violation of spatial inversion symmetry (named parity) in the weak interaction that was measured in the beta decay of unstable nuclei. Shortly after the discovery of parity violation other symmetries were found to be broken, and this triggered intense research on symmetry violations. Experimental activities include parity violation studies on atoms, ions, and molecules and searches for time reversal violating electric dipole moments of particles, including neutrons.

High-precision tests of fundamental symmetries and the SM with low-energy antiprotons are conducted at the facility AD/ELENA at CERN. This research has progressed considerably, and in figure 2.7 two examples are illustrated schematically, one showing the goal of comparing the gravitational mass of proton and antiproton and the other had the aim of comparing the mass of the antiproton with that of the electron (see also references [46, 47]).

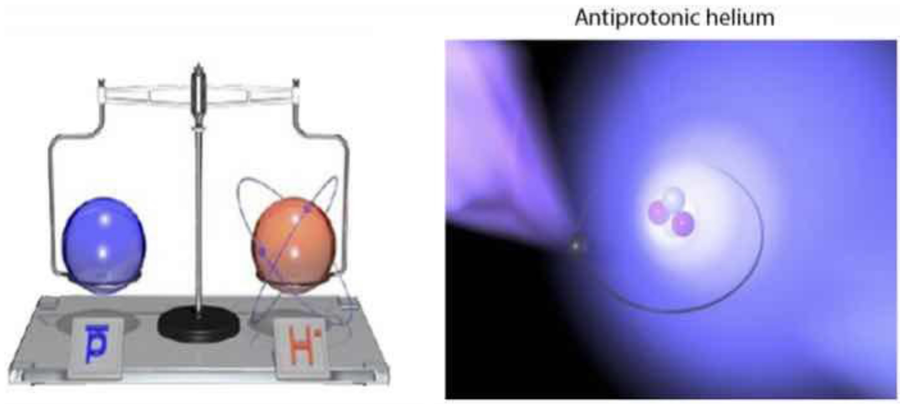


Figure 2.7. Left: Artistic views of the comparison of the gravitational mass of an antiproton and of a proton. Right: Artistic view of the antiprotonic helium. These researches are performed at CERN. Adapted from http://www.nupecc.org/pub/lrp17/lrp2017.pdf [18] courtesy of NuPECC (nupecc.org).

Today, many experiments on fundamental interactions bridge the areas of nuclear, atomic, particle, and astrophysics, and major advances in technology have made novel approaches feasible.

At the same time the most advanced theoretical and computational techniques are developed and applied by nuclear, atomic, and particle theory from low-energy particles, ultracold atoms, ions, and molecules. A number of these precision experiments use dedicated tabletop setups in small-size laboratories at universities where innovative techniques are tested before they are installed at large-scale infrastructures.

Another major class of experiments requiring high precision and the capability of detecting rare events concerns the search for the double beta decay without neutrino emission, performed worldwide, and in Europe in the underground laboratories LNGS, Modane, and Canfranc. These experiments are intended to determine the nature of neutrinos and to investigate the possible violation of conservation laws.

2.3.2.3 The phases of strongly interacting matter and the quark-gluon plasma

Nuclear collisions of heavy nuclei such as lead and gold at energies in the interval 100 GeV—few TeV per nucleon produce nuclear matter with temperatures in the trillions of degrees. The main motivation to measure these collisions is to recreate the matter in the same conditions taking place in the microseconds-old Universe after the Big Bang. It was understood in the 1970s that ordinary protons and neutrons could not exist at temperatures above two trillion degrees Celsius (see also references [48–50]).

The predicted new form of matter, which can be recreated by heating protons and neutrons until they 'melt,' was named quark-gluon plasma (QGP). In the hot soup of quarks and gluons produced by the Big Bang process the quarks are free to move within the region of the plasma over relatively large distances. A few fractions of a second after the Big Bang the temperature of the Universe dropped and the quarks became confined inside complex particles such as protons and neutrons that later formed the nuclei of the atoms. The transition between the primordial QGP and the hadron formation has, as far as we know, not left any imprint that is visible in present-day astronomical observations. Nevertheless, QGP matter is believed to exist inside the core of neutron stars, at baryon densities much higher than normal nuclear densities.

The energy or baryon densities necessary to form the QGP may be recreated in the laboratory via heavy ion collisions at sufficiently high energies, within volumes of the order of the nuclear size. The idea to collide heavy ions accelerated at ultrarelativistic energies for bringing nuclear matter into the QGP phase and studying its properties in the laboratory dates back to the early 1980s. Pioneering studies promptly demonstrated that the energy deposit and the nuclear stopping in the central part of the collision were quite large. The acceleration of heavy ions at the SPS accelerator at CERN, since 1994, and later the experiments at the BNL Relativistic Heavy Ion Collider (RHIC) energy for nucleon collisions of 200 GeV, created a system with energy density above the critical value and allowed the observation of the predicted signatures of the QGP. At higher energy, the colliding system was found to enter a new regime with an increase in the lifetime of the quark-gluon plasma.

The feature of the strength of the strong interaction decreasing at short distances has a very profound implication for nuclear matter under extreme conditions because at sufficiently high nuclear density or temperature, the average inter-parton distance becomes small, and therefore their interaction strength weakens. Above a critical energy density, of the order of 0.3 GeV fm^{-3}, a gas of hadrons undergoes a deconfinement transition and becomes a system of unbounded quarks and gluons. Predictions for this transition were obtained with simulations based on the QCD theory, in the form of a rapid increase of the entropy density around the critical energy density. The formation of QGP is also accompanied by a restoration of a particular symmetry, the chiral symmetry (a system is chiral if it is distinguishable from its mirror image).

The so-called phase diagram for nuclear matter is illustrated in a schematic way in figure 2.8, where the laboratories involved in experimental programs concerning the physics of the quark-gluon plasma are indicated by stars. The goal of the high-energy heavy-ion programs at these laboratories is to identify and characterize the properties of the QGP. These programs naturally have two steps: understanding the dynamics of heavy-ion collisions (e.g., via comparison to phenomenological models) and the extraction of fundamental properties to be compared to QCD predictions.

The heavy-ion collisions can be described as being characterized by three main stages: (i) an early non-equilibrium stage, (ii) an expansion stage, and (iii) a final freeze-out stage. An advantage of this modular structure is that it allows for the use of more or less advanced theoretical tools in each stage. In this way the modeling of heavy-ion collisions can gradually be improved and used to constrain further the properties of strongly interacting matter. This picture, and the associated phenomenology, has evolved over the last years as observables have been identified that are sensitive to specific processes in each phase.

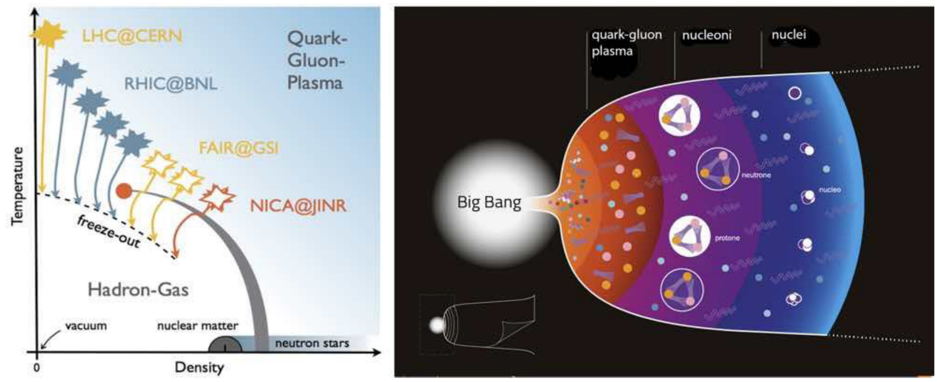


Figure 2.8. Left: Illustration of the phase diagram of quantum chromodynamics. The facilities probing the different sectors of this diagram are indicated. Right: Pictorial illustration of the Big Bang indicating the evolution from quark-gluon plasma up to the formation of nuclei. Adapted from http://www.nupecc.org/pub/lrp17/lrp2017.pdf [18] courtesy of NuPECC (nupecc.org).

The first stage, which also provides initial conditions (spatial distribution of the deposited energy and pressure, initial flow velocity) for the subsequent hydrodynamical stage, is the least known and is often described by simple geometrical models using the underlying strong nucleon–nucleon interactions. More in-depth descriptions treating the collision at quark and gluon level are being developed today. Although some productions of particular particles via electromagnetic processes that were measured by LHC experiments in lead-lead collisions (e.g., J/ψ photo-production) provide evidence for shadowing due to nuclear gluons, further efforts are required to extract its amount.

During the second stage—the fireball expansion—the bulk evolution is described by relativistic viscous hydrodynamics. Due to the near-perfect fluid nature of the QGP, the initial geometrical anisotropy is efficiently converted into momentum anisotropy of the final particles. The systematic measurement of the flow of particles at different directions (as for example the elliptic flow) provides an avenue for constraining initial-state model calculations and transport properties of the QGP. The emergent hydrodynamic properties of the QGP are not apparent from the underlying QCD theory and thus they have been quantified with increasing precision via experiments at both RHIC and the LHC over the last several years. New theoretical tools have allowed the degree of fluidity to be characterised. In the temperature regime created at RHIC, QGP is the most liquid-like liquid known, while the hotter QGP created at the LHC has a somewhat larger viscosity. This temperature dependence will be more tightly constrained by the ongoing and upcoming measurements. Moreover, this bulk evolution of the collision of ions at relativistic energy provides the substrate for the medium modifications of hard probes, although a better integration of these two aspects of the description is certainly needed.

The formation of hadron particles takes place when the system reaches the pseudo-critical temperature, and after that the scattering rate decreases quickly and a kinetic description becomes more appropriate than hydrodynamics. This kinetic description can in principle account for the decoupling of the hadrons from the fireball produced by the collision. The measured relative abundances of hadrons indicate that chemical freeze-out happens at a temperature which is very close to the hadronisation temperature and at nearly zero chemical potential. Subsequently, the hadrons continue to rescatter elastically until they reach the kinetic freeze-out temperature where they decouple and freely stream to the detectors.

The complex analysis of data collected so far has produced many interesting results addressing the bulk properties of the collisions and the details of particle production whose explanation relies to large extent on the comparison with data from proton collisions. Among the results from the data from ALICE at LHC there is the copious production of antinuclei whose investigation opens interesting perspectives in understanding the role of antimatter in the Universe, the latter based on astronomical observations.

2.3.3 Nuclear structure and reactions for the origin of elements

The subfield of nuclear structure and reactions focuses on measuring, explaining, and using nuclei to improve scientific understanding as well as for scientific applications. This research addresses the underlying nature of atomic nuclei and the limits to their existence. The complex physics of the atomic nucleus has not only shaped our Universe but also ourselves (see, e.g., references [51–55]).

Nuclear astrophysics aims to understand the evolution of matter in all its complexity across cosmological times, just after the Big Bang, during the life-cycles of stars when the primary elements needed for life were created (see schematic picture in the top-left part of figure 2.9) and in the violent cosmic events that delivered the heaviest elements (top-right part of figure 2.9). The lightest elements, hydrogen, helium, and lithium, were made from primordial Big Bang matter, before the first stars were formed. The rest of the elements are thought to be built up in stars via processes occurring in their plasma involving complex chains of nuclear reactions and radioactive decays. In the bottom part of figure 2.9 different explosive nucleosynthesis processes responsible for the creation of the heaviest elements are shown by the coloured arrows (see, e.g., references [56–59]).

During a typical lifetime of a star, its supply of hydrogen fuel is 'burnt' to helium, followed by burning to carbon, nitrogen, and oxygen. Eventually the heavier elements such as silicon and finally iron and nickel are reached. The reaction chains can be mapped on a landscape charting the nuclear species that form with increasing numbers of protons and neutrons. It climbs slowly up along the central 'valley' of stable nuclei, continuing until all the nuclear fuel of the star is entirely consumed and thus the star eventually collapses.

We know little about how the heavier elements beyond iron in the Periodic table are created. Theory suggests that many of them are generated in extremely violent environments such as a supernova explosion that ends the life of a very massive star, followed by its core collapse into a neutron star or black hole. Other stellar processes of interest are those in which a white dwarf in a binary star system is

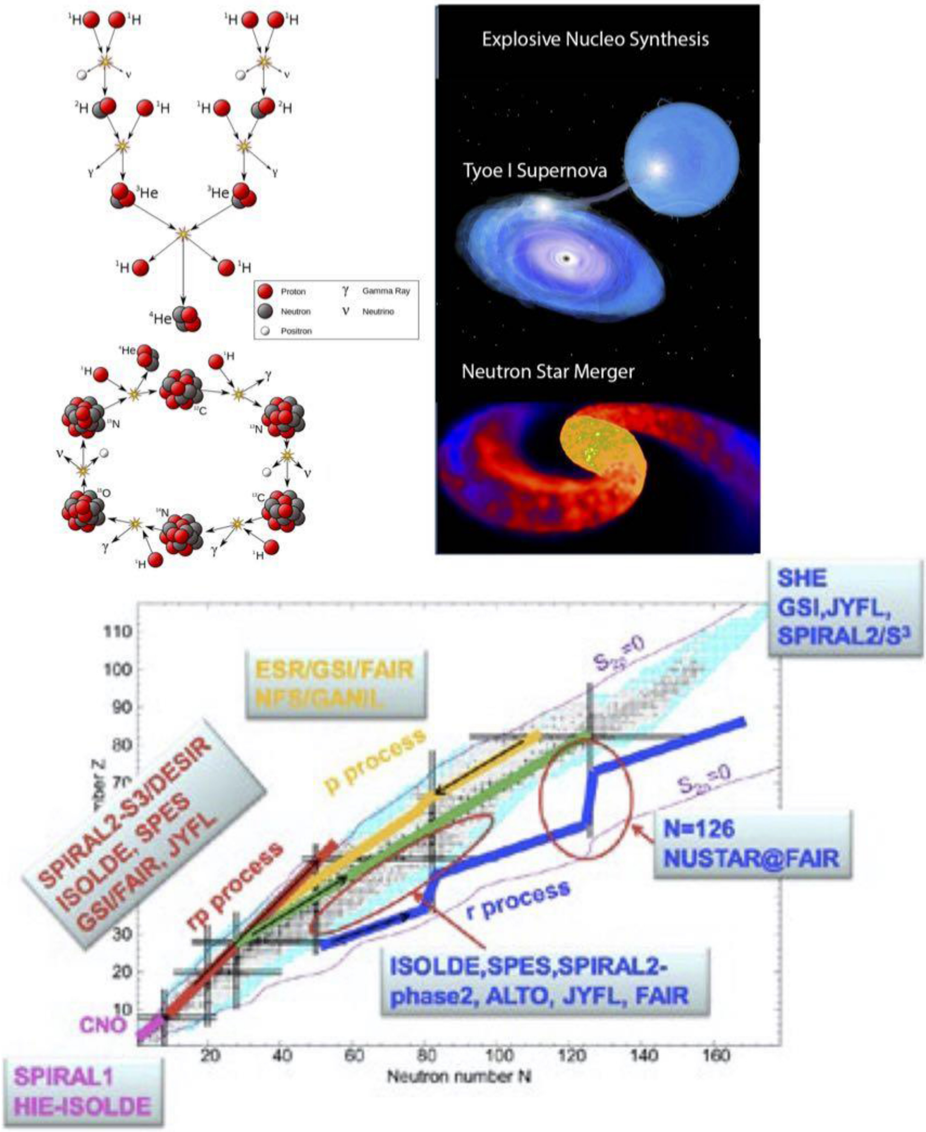


Figure 2.9. Top left: Artistic pictures of the nucleosyntheses of light nuclei. The p–p and the CNO cycles are shown. Some of the reactions involved in these cycles were measured at stellar energy in the laboratory LNGS-INFN with the experiment LUNA and the neutrinos with the experiment BOREXINO. Top right: Two pictures illustrating schematically two types of explosive nucleosynthesis (supernovae and neutron star mergers). Bottom: Schematic illustration of the nuclear chart, with the number of neutrons on the horizontal axis and the number of protons on the vertical one. The different regions relevant for the various nucleosynthesis processes in stellar explosions are indicated together with the European facilities involved in the study of these nuclei. The major facilities outside Europe addressing these issues are TRIUMF in Canada, FRIB in USA, and RIBF-RIKEN in Japan. Adapted from http://www.nupecc.org/pub/lrp17/lrp2017.pdf [18] courtesy of NuPECC (nupecc.org).

explosively resurrected by drawing off gas from its companion. Even more violent events like the merger of two neutron stars or black holes may also trigger these rapid nucleosynthetic processes. They are thought to follow paths up the nuclear landscape involving thousands of unstable nuclear species on both sides of the valley of stability, around the central ensemble of stable nuclei.

The challenge of understanding the origin and evolution of the chemical elements, and the role of nuclear physics in the lives and deaths of stars, requires state-of-the-art experimental, theoretical, and observational capabilities. Nature has supplied the Earth with about 300 stable as well as long-lived radioactive isotopes that we can study in the laboratory. However, a much greater variety of unstable isotopes, those created during stellar explosions, can be produced as radioactive beams in the laboratories using different systems of accelerator facilities. Developments during the years of these facilities have provided access to an increasing number of these exotic nuclei. However, only approximately 3000, of the possibly more than 8000 different nuclei that should exist, have been produced and characterized. A large 'terra incognita' dominated by the very neutron-rich nuclei and superheavy elements has still to be uncovered.

The nuclear properties relevant for the description of astrophysical processes depend on the environmental conditions. Nuclear theory is fundamental to connect experimental data with the finite temperature and high-density conditions in the stellar plasma. Advances in the description of nuclear interactions and of nuclear structure are crucial for our understanding of the evolution and explosion of stars, the chemical evolution of the Galaxy and its assembly history. In this context, the physics of neutron stars is of particular interest to the nuclear physics community. Indeed, the structure and composition of neutron stars are uniquely determined by features of the neutron-rich matter, and moreover by the relation between the pressure and energy density of this matter. Measurements of neutron star masses and radii place significant constraints on this pressure and density relation. Conversely, the measurement of both masses and the pressure of neutrons at the nuclear surface of exotic nuclei will provide critical insights into the composition of the neutron star crust. Neutron star crust models will have to rely on a combination of experimental and theoretical data, especially modifications to masses and effects such as superfluidity and neutrino emissivity.

The features of nuclei are probed via specific nuclear reactions and the key questions are related to the evolution of nuclear structure, how the structure and the nuclear shapes change with temperature and angular momentum, and the complexity of nuclear excitations. Complex systems often display surprising simplicities and symmetries and nuclei are no exception.

The resulting emergent phenomena are the appearance of nucleonic shells, saturation of nuclear matter, rotation, superfluidity, and phase transitions. In addition, it is remarkable that a system with a hundred nucleons or more exhibits collective properties of many nucleons operating together. The detailed investigations of these varieties of nuclear features have progressed considerably, also entering into a high precision regime, thanks to powerful instrumentation, such as that for gamma-ray detection. For gamma spectroscopy the AGATA and

GRETINA arrays (in Europe and USA, respectively, see, e.g., references [60–62]) are the best-performing detector systems currently available, and there are plans to make them more powerful in the future, also in view of their great potential for applications in imaging techniques.

Extensive research is addressing the study of the neutron-rich matter which is less well known and for which very long isotopic chains are expected to exist. In heavier neutron-rich nuclei, the excess of neutrons collects at the nuclear surface creating a 'skin,' a region of weakly-bound neutron matter that is our best laboratory access to the diluted matter existing in the crusts of neutron stars. In spite of the progress made there are still mass and excitation energy regions that need to be explored to understand nuclear interactions, configurations, and nucleosynthesis processes. To access the unexplored mass regions one needs the beams that will be developed at the facilities undergoing major upgrades or under construction.

2.3.4 Challenges and opportunities up to 2050

For the next decades and in particular up to 2050 very appealing scientific challenges and opportunities exist in Nuclear Physics for both of the major research areas: 'The fundamental description of the heart of matter' and 'Nuclear structure and reactions for the origin of elements'.

At the FAIR facility, presently under construction, activities in both these research areas, together with many types of applications, will be carried out for several decades. This worldwide unique accelerator and experimental facility is composed of several research infrastructures that will allow for a large variety of unprecedented leading-edge research.

A schematic illustration of the FAIR facility, showing its four pillars, is shown in figure 2.10. These pillars are:

- The Super-FRS (super fragment separator) together with storage cooler rings and the versatile instrumentation (denominated as NUSTAR). These will allow decisive breakthroughs in the understanding of nuclear structure and nuclear astrophysics.
- The ultrarelativistic heavy-ion collision experiment CBM with its high rate capabilities. This will permit the measurement of extremely rare probes that are essential for the understanding of strongly interacting matter at high densities.
- The PANDA detector system at the antiproton storage cooler ring HESR. This facility will provide a unique research environment for an extensive program in hadron spectroscopy, hadron structure, and hadronic interactions.
- The APPA experiments will exploit the large variety of ion beam species, together with the storage rings and precision ion traps, for a rich program in fundamental interaction physics and applied sciences.

2.3.5 The fundamental description of the heart of matter: future perspectives

Our view of the structure of the nucleons has undergone a huge transformation in the last few decades. The most common picture is that the inside of the nucleon is rather a complex many-body system with a large number of gluons and sea quarks.

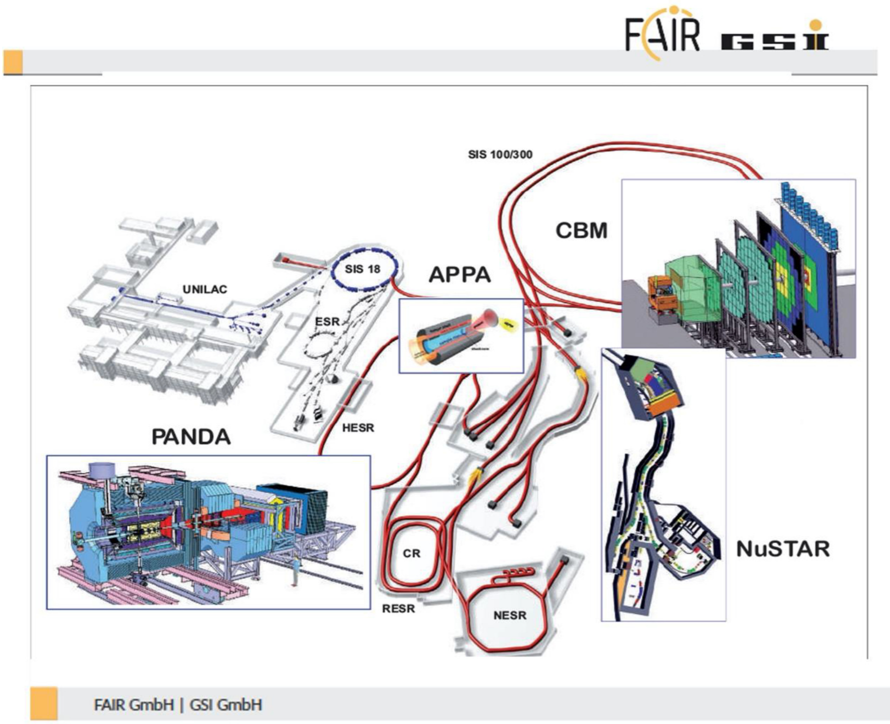


Figure 2.10. The FAIR accelerator complex. The picture illustrates schematically the accelerators and the areas in which the detector systems devoted to the experiments concerning different areas of research. PANDA and CBM are devoted to the study of the fundamental constituents of matter, with CBM specifically designed for the hot highly compressed matter; NUSTAR is for investigations of nuclear structure, reactions, and nuclear astrophysics; APPA is for applications involving also atomic and plasma physics. Adapted from [21], copyright IOP Publishing Ltd. All rights reserved.

There is unambiguous evidence that they both play surprisingly important roles for defining the structure of nuclear matter around us. Their quantitative study requires a novel sophisticated tool such as the collision of high-energy electrons and ions.

In the USA a new facility providing Electron–Ion Collisions (EIC) is under construction at the Brookhaven National Laboratory, and several groups in Europe are involved primarily in the preparation of the experiments to be made there (see, e.g., references [63, 64]). The most precise information physicists have about the internal structure comes from scattering electrons, which interact with the nucleus and transfer momentum and energy. The layers of structure exposed change as one varies the value of the transferred momentum and energy.

The electron–ion collisions at the EIC facility will allow scientists to address immediate and profound questions about neutrons and protons and how they form the nuclei of atoms. The first question is how the mass of the nucleon arises. Protons and neutrons are bound systems of very light quarks so that their mass is generated dynamically through interactions carried by gluons. These collisions will map the gluon distribution in the proton, both in space and in momentum, with unprecedented precision, using the new technique of parton tomography. These images can be used to analyse the coupling between spin and orbital angular momentum. It would be possible not only determine the distribution of gluons but also measure the distribution of gluonic energy density and pressure in the proton. Other questions to be answered concern the origin of the nucleon spin and the emergent properties of dense systems of gluons.

The force mediated by gluons is fundamentally different from the electromagnetic force, since it strengthens as the quarks get farther apart so that quarks are permanently confined in neutrons and protons. Two questions concerning the gluons arise when nucleons are combined into nuclei: how are they modified as compared with a proton and is it possible that the whole nucleus becomes a dense gluon system? This abundance of gluons provides the opportunity to address fundamental questions about nucleons and nuclei. The number of gluons grows significantly in the high-energy limit. This means that gluons must overlap in the plane transverse to the electron–ion collision. The most interesting case is when this limit can be achieved at high resolution so that the number of gluons that can be packed into the transverse area of a proton or nucleus is large. Under such conditions, a quantum state of cold dense gluonic matter may exist. Such a state is possibly analogous to Bose–Einstein condensates of clouds of cold atoms created in atomic physics laboratories.

The EIC will be able to reach unprecedented gluon densities because it can accelerate ions with high atomic weight. Figure 2.11 shows how the scientific areas of investigation map onto the required collider luminosity and center-of-mass energy. The machine must collide electrons with protons and other atomic nuclei (ions) over a range of energies. There must be enough collisions for the experiment to gather adequate data to elucidate or settle the known physics questions, and other questions that may emerge, in a reasonable time. A collider's ability to squeeze many particles of two beams into a tiny volume where they collide defines its luminosity. The luminosity ultimately required of an EIC is comparable to those of the highest performing colliders built to date, such as the LHC at CERN.

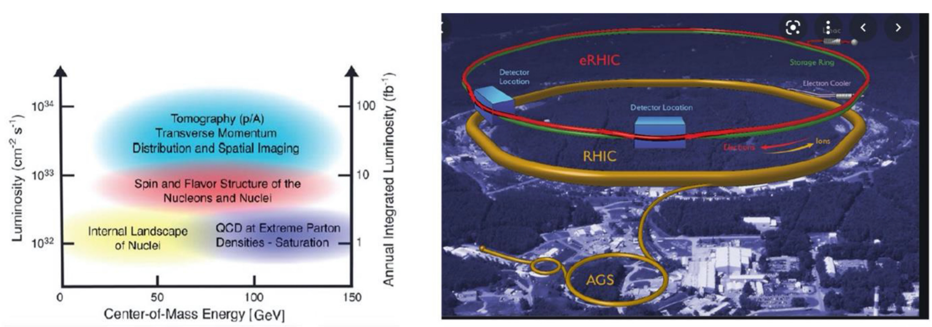


Figure 2.11. Left: The picture highlights the different scientific areas that can be addressed depending on the luminosity of the Electron Ion Collider (EIC) and the centre-of-mass energy. Right: Schematic layout of the planned EIC accelerator based on the existing RHIC complex at Brookhaven National Laboratory. Adapted from [34] courtesy of Sissy Körner.

Furthermore, given the crucial role of spin, both beams must be polarized. At present, it is anticipated that the EIC facility will begin operation in the early 2030s.

Beyond its importance for the understanding of fundamental properties of quarks and their interaction when they are in extreme conditions, the study of highly compressed nuclear matter is of great interest in the context of astrophysics, and the planned future measurements based on fixed target experiments (such as those at NICA and at FAIR) will provide highly selective data which are relevant for the physics of neutron stars.

The discovery of neutron stars with masses of about two solar masses and the observation of gravitational waves from neutron stars merging impose severe constraints on the nuclear matter in these stars. For example, it seems that the appearance of particles containing strange quarks could indicate that heavy neutron stars may collapse into black holes. While the structure of neutron stars is governed by the features at small temperature, present simulations of neutron star mergers indicate that the shocks which eject the matter relevant for formation of heavy elements (with approximately 200 nucleons) could reach temperatures of 100 MeV and densities of several times nuclear matter density. These conditions are very similar to those that are responsible for the collective expansion of the matter in heavy ion collisions that will be measured with very high rates at FAIR. High rates will allow the collision process to be probed in a selective way and thus permit a comprehensive analysis of collective flow and the particle production—quantities that are important inputs for simulations of neutron star mergers.

2.3.6 Nuclear structure and reactions and the origin of elements: future perspectives

Up to now many rare and unusual nuclei have been created and studied using accelerator facilities. The result has been the discovery of a wonderful variety of complex structures and shapes, from lightweight 'halo' nuclei in which an outer neutron orbits far from the central structure, through nuclei arranged in smaller clusters of nucleons, or in concentric shell-like structures, which when fully occupied by nucleons are very stable ('magic' nuclei), to dense, heavyweight nuclei with unusual shapes that behave more like wobbling liquid droplets. Experiments probing these nuclei attempt to understand how these various structures emerge as the numbers and proportions of protons and neutrons change across the nuclear landscape (see figure 2.12), and what controls their stability and behaviour when their energy content is varied. Of particular interest are nuclei with neutron–proton proportions that are only just bound, and are on the edge of shedding neutrons or protons (the neutron or proton 'driplines'), and 'superheavy' nuclei containing up to 300 nucleons.

The territory of neutron-rich nuclei is the most fertile ground for research in nuclear structure and further progress in this area requires measurements of key isotopic chains that encompass multiple magic numbers. With future experiments with radioactive beams, in particular those at FAIR with its suite of unique instrumentation, these chains will be accessible from the proton drip line to the neutron drip line (namely the boundary delimiting the zone beyond which atomic

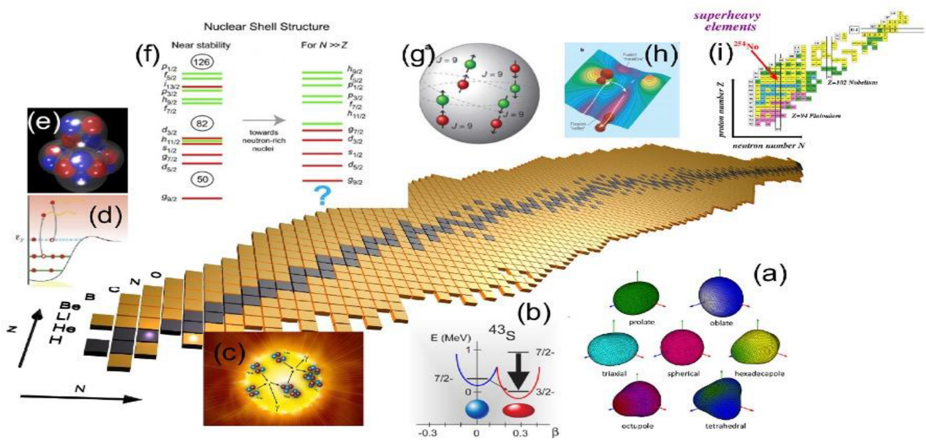


Figure 2.12. Artist's view of the nuclear landscape illustrating some of the key properties currently being studied using high-resolution spectroscopy: (a) the variety of nuclear shapes, (b) shape coexistence and isomerism, (c) reactions of astrophysical interest, (d) coupling to the continuum of unbound states, I cluster structure in nuclei, (f) evolution of the shell structure, (g) nuclear superconductivity, (h) understanding nuclear fusion and fission reactions, and (i) the journey towards the heaviest elements. Adapted from [60] CC BY 4.0.

nuclei decay by the direct emission of a proton or a neutron), permitting the study of the N/Z dependence of the nuclear force and continuum effects over very broad ranges. Such investigations will allow us to explore new paradigms of nuclear structure in the domain where many-body correlations, rather than the nuclear mean field, dominate. Single- and even multiple-neutron emissions are expected to characterise nuclei at the neutron drip line, while beta-delayed-neutron decay is prevalent among neutron-excess nuclei before the drip line is reached. Both forms of radioactivity only occur among nuclei far from stability.

To explain the nature of neutron-rich matter across a range of densities, as seen in the crust of neutron stars, an interdisciplinary approach is essential in order to integrate low-energy nuclear experiments and theory with knowledge from astrophysics, atomic physics, computational science, and electromagnetic and gravitational-wave astronomy. The nuclear input to this mix is essential. It includes the studies of nuclear matter at around both supranuclear and subnuclear densities, the analysis of high-frequency nuclear oscillations, and ab-initio approaches to the equation of state of nuclear matter.

Laboratory measurements using specific nuclear reactions induced by light and heavy ions are necessary to constrain nuclear matter compressibility and the symmetry energy. The symmetry energy is the energy difference between nuclear matter with protons and neutrons, and the pure neutron matter properties determine a range of neutron star features such as cooling rates, the thickness of the crust, the mass-radius relationship, and the moment of inertia. The precise characterisation of the nuclear symmetry energy through future experiments will be, therefore, a crucial step towards our capability of interpreting neutron star matter and its properties. Likewise, studies of masses, giant resonances, dipole polarisabilities, and neutron skin thicknesses of neutron-rich nuclei will provide key insights for astrophysics. Extending such measurements to more neutron-rich nuclei and increasing the precision, especially of neutron skin thickness measurements, will be an important component of the future studies.

The territory at and beyond the proton drip line offers instead unique opportunities to study other exotic nuclear decays and correlations, such as ground-state one- and two-proton decay, a class of radioactivity that exists nowhere else but that provides unique insight.

Superheavy nuclei will be the object of intensive research with new dedicated facilities, like at the new one at JINR which has just started its operation. It was recognized long ago that, in spite of the huge electric repulsion between all the protons in the nucleus, the binding that comes from the strong force could tip the balance in favour of the existence of superheavy nuclei. Precise calculations at the limits of mass and charge are difficult and thus the progress in this field has come largely from experimental efforts. The important steps forward in heavy-element discovery have to rely on the decay chains of alpha particles, as demonstrated by pioneering measurements in the recent past. To find the most favourable conditions for producing even heavier elements, extensive measurements with fusion reactions of different types of beams will be employed. The future on this topic seems to be bright, in particular because the new experiments will inform us on how to reach the expected region of long-lived superheavy nuclei. These nuclei will have to be identified and investigated to determine their properties, and x-ray detection will help to pin down their new high values of protons. One-atom-at-a-time chemistry studies will also expand into this region of superheavy elements.

2.3.6.1 Nuclear reactions and decays in the plasma

The availability of very intense ion beams will open new research directions concerning the formation and characterization of hot plasma at very high density, a region that needs to be explored. In such very compressed hot plasmas it is very interesting to study the acceleration acquired by nuclear particles (proton, heavy ions), the occurrence of nuclear reactions, and how nuclear decay properties change. The attraction by these issues is connected to the fact that when such plasmas are created in the laboratories, one can reproduce conditions on Earth similar to those occurring in some stars and planets.

Another efficient way to produce high-density plasmas is to use lasers (free-electron lasers and high-power lasers) which, depending on their power, can create different conditions of density and temperature. The combined and coherent efforts at the facilities either in operation or under construction or planned worldwide will allow the temperature–density plane to be mapped, where one can locate different elements of the cosmos, when characterised in relation of the features of their plasma (figure 2.13).

At FAIR research using the worldwide unique combination of intense laser radiation with beams, started at the existing laboratory GSI, will be boosted in the coming years. This combination opens the door for interesting experiments in the field of plasma physics, nuclear physics, and atomic physics. In addition, standalone

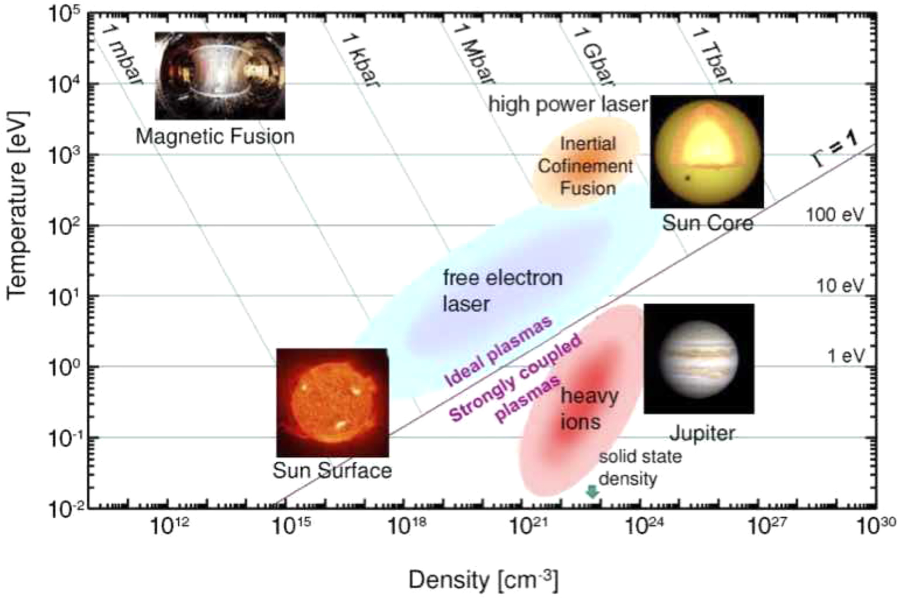


Figure 2.13. The plasma in plane of temperature and density with indicated some typical processes in the cosmos. The techniques used in laboratories to create plasma in different regions of temperature and density are also schematically indicated. Adapted from [21], copyright IOP Publishing Ltd. All rights reserved.

experiments with intense laser beams can be carried out to study proton acceleration or x-ray laser generation. These experiments complement other ones at facilities mostly devoted to develop new laser-based acceleration techniques.

A very interesting new direction in the field of nuclear physics is represented by the measurements of nuclear reactions induced by ions of different types produced in the plasma which can be created at high density by very powerful lasers. Although the concept of producing fusion reactions with the lightest nuclei (H, D, T, and He) in the plasmas is at the basis of reactors designed mainly for energy-generation purposes (as in the case of the very large infrastructure ITER), the measurements in laboratories of stellar reactions involving heavier nuclei, as in several types of stars, are very few. New efforts are made exploiting the ignition process that is used within the inertial confinement fusion research. In that context, the initiation of nuclear fusion reactions is produced by heating and compressing a fuel contained in solid targets made of different types of materials. The energy to compress and heat the fuel is delivered to the outer layer of the target using laser light beams, electrons, or ions.

It is important to point out that the high-power lasers at the new ELI International Facility (with three pillars, in Hungary, the Czech Republic, and Romania) that have recently become operational will offer the possibility to carry out interesting programs on the investigation of several different nuclear processes in the plasmas. The high-power lasers of ELI will be capable of producing high-energy

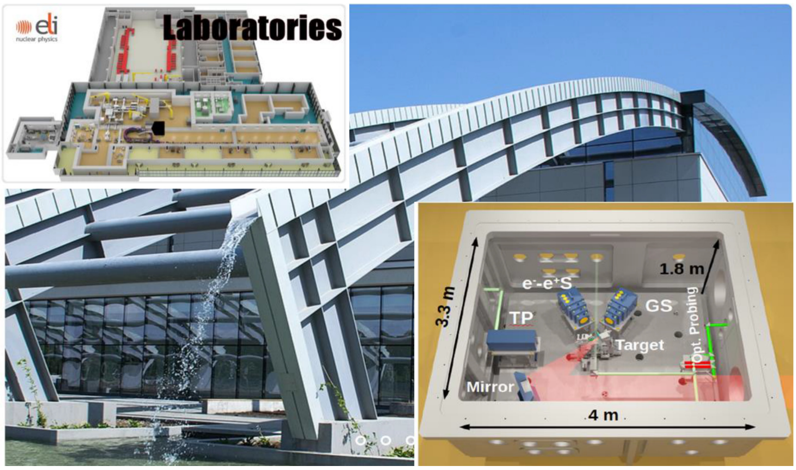


Figure 2.14. Schematic illustration of the ELI facility in Bucharest. Adapted from https://www.eli-np.ro courtesy of Sissy Körner.

charged particles, gamma rays, and neutrons, with a peak flux orders of magnitude higher than achievable with conventional accelerators.

In particular, at the ELI-NP pillar in Romania (see figure 2.14), which is more focused on nuclear physics, the short-duration pulses of its lasers will produce high fluxes of nuclear particles. These nuclear particles will be studied together with the new kinds of nuclear physics phenomena, such as:

- Exotic, heavy neutron-rich nuclei produced using new methods involving sequential reactions in plasmas
- The stopping power of charged particles bunches in dense plasmas
- Nuclear reactions in hot and dense plasmas simulating in the laboratory the astrophysical phenomena
- Nuclear excitations and de-excitations in plasmas leading to changes in (apparent) nuclear lifetimes

The investigations of exotic nuclear reactions aim primarily to explore the production of neutron-rich nuclei at around the $N = 126$ waiting point, which is relevant for the astrophysical r-process of nucleosynthesis. For this purpose, bunches of heavy ions generated and accelerated to around 10 MeV/nucleon by the laser through the radiation pressure acceleration mechanism will be employed. In this way a sequential fission–fusion will follow. This planned research will help to elucidate the mystery of high-Z element formation in the Universe. The proposed scheme is complementary to other approaches existing or planned for the production of radioactive nuclei and if successful, beyond demonstration, it will open interesting opportunities for the future.

2.4 Quantum gravity—an unfinished revolution

Claus Kiefer[1]
[1]Faculty of Mathematics and Natural Sciences, Institute for Theoretical Physics, University of Cologne, Cologne, Germany

It is generally assumed that the search for a consistent and testable theory of quantum gravity is among the most important open problems of fundamental physics. I review the motivations for this search, the main problems on the way, and the status of present approaches and their physical relevance. I speculate on what the situation could be in 2050.

2.4.1 Present understanding and applications

2.4.1.1 The mystery of gravity

Already one year after the completion of his theory of general relativity, Albert Einstein predicted the existence of gravitational waves from his new theory. At the end of his paper, he wrote[2]:

> …the atoms would have to emit, because of the inner atomic electronic motion, not only electromagnetic, but also gravitational energy, although in tiny amounts. Since this hardly holds true in nature, it seems that quantum theory will have to modify not only Maxwell's electrodynamics, but also the new theory of gravitation.

Thus already in 1916 Einstein envisaged that quantum theory, which at that time was still in its infancy, would have to modify his newly developed theory of relativity. More than hundred years later, we do not have a complete quantum theory of gravity. Why is that and what are the prospects for the future?

Gravitation (or simply gravity) is the oldest of the known interactions, but still the most mysterious one. It was Isaac Newton's great insight to recognize that gravity is responsible for the fall of an apple as well as for the motion of the Moon and the planets. In this way, he could unify astronomy (hitherto relevant for the region of the Moon and beyond) and physics (hitherto relevant for the sublunar region) into one framework. In the Newtonian picture as presented in his *Principia* from 1687, gravity is understood as action at a distance: any two bodies in the Universe attract each other by a force which is inversely proportional to the square of their distance (see appendix). For this, he had to introduce the so far unknown concepts of *absolute space* (which has three dimensions) and *absolute time* (which has one dimension). These entities exist independent of any matter, for which they act like a fixed arena that cannot be reacted upon by the dynamics of matter. Newton's discovery marked

[2] This is my translation from the German. The original reference can be found in Kiefer [65], p 26.

the beginning of modern celestial mechanics, which allowed the study of the motion of planets and other astronomical bodies with unprecedented accuracy.

The strength of the gravitational force between two bodies is proportional to their *masses*. Masses can only be positive, in constrast to electric charges, which can be both positive and negative. This difference is the reason why charges can attract each other (if they, unlike the masses in gravity, differ in sign) as well as repel each other (if they have the same sign). For elementary particles, mass, by which we mean rest mass, is an intrinsic property (the same holds for charge). There can also exist particles with zero mass, of which the only observed one is the photon; such particles must propagate with the speed of light c. Elementary particles are also distinguished by their intrinsic angular momentum (spin), by which they can be divided into bosons (having integer spin) and fermions (having half-integer spin).

Newton's theory of gravity was superseded only with the advent of general relativity in 1915. It was Einstein's great insight to recognize that gravity can be understood as representing the *geometry* of space and time as unified to a four-dimensional entity called spacetime. In this way, gravity acquires its own dynamical local degrees of freedom. Spacetime then no longer plays the role of a fixed background acting on matter, but takes itself part in the game and can be reacted upon—both by matter and by itself. Gravity itself creates a gravitational field as is reflected by the non-linear nature of Einstein's field equations (see appendix). That gravity possesses its own degrees of freedom can best be seen by the existence of gravitational waves, which propagate with the speed of light and which were detected directly for the first time by the laser interferometers of the Laser Interferometer Gravitational-wave Observatory (LIGO) collaboration in 2015. That gravity (and thus spacetime) is fully dynamical is also called *background independence*.

Gravity is very weak. The gravitational attraction between, say, electron and proton in a hydrogen atom is about 10^{40} times smaller than their electric attraction. A metallic body can be prevented from falling to the massive Earth by holding it with a small magnet. Still, because masses are only positive, it is the dominating force for the Universe at large scales, because positive and negative electric charges, being present in roughly equal amounts, average to zero at those scales.

Newton had carefully distinguished between gravity (interaction between bodies) and inertia (resistance of bodies to changes in their momenta). These two concepts are unified in Einstein's theory as expressed by the equivalence principle. The geometry of spacetime thus leads in appropriate limits to the traditional gravitational interaction as well as to inertial forces such as centrifugal or Coriolis forces.

Gravity is of a universal nature. Everything in the world is in spacetime and is thus subject to its geometry, that is, to gravity. So far, Einstein's theory successfully explains all observed gravitational effects from everyday life (e.g., the working of the GPS) to the Universe as a whole. Figure 2.15 presents a famous photograph showing galaxies at distances that cover cosmic scales in space and in time—because light propagates at the finite speed c we see these galaxies in a very early state of their evolution, billions of years ago. Astronomers measure cosmic scales in Megaparsec (Mpc) and Gigaparsec (Gpc). In conventional units, 1 Mpc $\approx 3.09 \times 10^{22}$ m, and 1

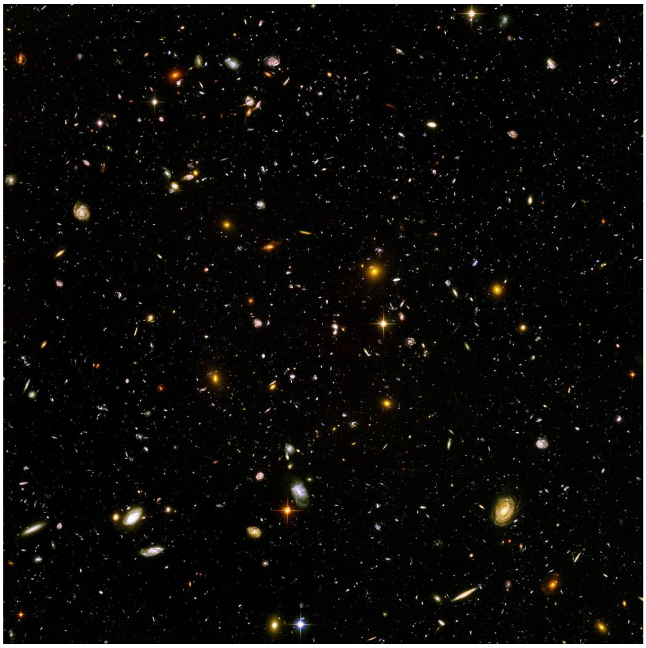


Figure 2.15. A glimpse into the macroscopic world: the *Hubble Ultra Deep Field*, a photograph taken from September 2003 to January 2004 in a small celestial region in the constellation Fornax. Figure credit: NASA and the European Space Agency.

Gpc is thousand Mpc. The size of the observable Universe is estimated to be about 14 Gpc[3].

Strictly speaking, there are two features for which it is presently open whether they can be fully accommodated into Einstein's theory or not: Dark Matter and Dark Energy. The two can only be observed by their gravitational influence; Dark Matter exhibits the same clumpiness as visible matter (and exhibits itself, for example, in the rotation curves of galaxies), but Dark Energy is of a homogeneous and repulsive nature and is responsible for the present accelerated expansion of our Universe (as measured by observing supernovae at increasing distances). Some scientists speculate that new physics is needed to account for Dark Matter and Dark Energy, but at present this is far from clear.

General relativity is what one calls a classical, that is, non-quantum, theory. Our current theories for the other interactions are all *quantum* theories or, more precisely, these interactions are described within a quantum framework, which uses concepts drastically different from classical physics. For example, whereas classical mechanics makes essential use of *trajectories* for bodies, the equations of which are determined by their initial positions and momenta, quantum mechanics no longer

[3] This is the so-called particle horizon: the distance in today's Universe up to which we can see objects, that is, the distance over which information (basically in the form of electrodynamic or gravitational waves) had enough time since the Big Bang to reach us. The age of our Universe is estimated to be about 13.8 billion years.

contains such trajectories in its mathematical description[4]. It instead features wave functions Ψ from which observable quantities such as energy values for spectra and interference patterns of particles can be obtained. The relation to positions, momenta (and other classical concepts) proceeds via the probability interpretation, and the limits can be expressed by the indeterminacy (or uncertainty) relations. The quantum-to-classical transition can be understood and experimentally studied using the concept of decoherence [66, 67]. The quantum framework and formalism seems to be of universal nature.

Quantum theory is usually applied in the realm of microphysics. This is the world of molecules, atoms, nuclei, and elementary particles. Quantum theory thus lies at the basis not only of physics, but also of chemistry and biology. The smallest scales investigated experimentally so far are the scales explored by particle accelerators such as the LHC at CERN. Figure 2.16 shows a glimpse into these smallest scales—the decay of the Higgs particle into other particles. Such 'microscopic' pictures are far more abstract than photos of the kind shown in figure 2.15; a great amount of theoretical insight is involved to construct them.

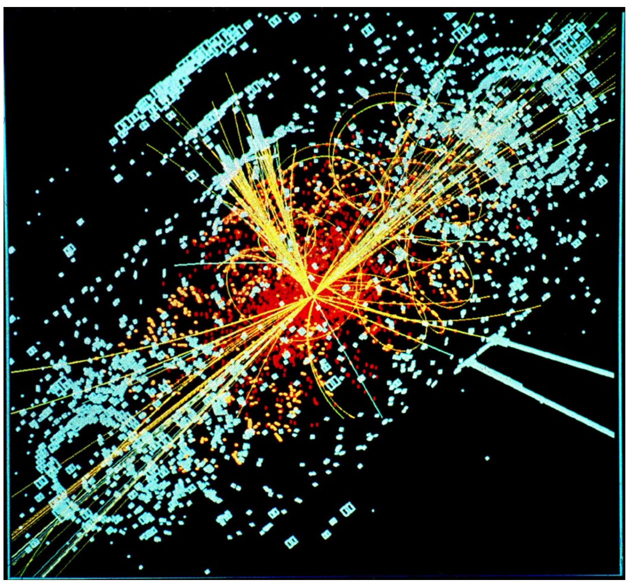


Figure 2.16. A glimpse into the microscopic world: simulation of the hypothetical decay of a Higgs particle into other particles at the detector CMS at CERN. Figure credit: Lucas Taylor/CERN—http://cdsweb.cern.ch/record/6284699 (Creative Commons License).

[4] The trajectories that appear in the so-called de Broglie–Bohm interpretation of quantum theory are of a non-classical nature.

These smallest explored scales are of the order of 10^{-18} m. Comparing it to the above cosmic scale of 14 Gpc, which is about 4×10^{26} m, we see that this corresponds to a difference of about 44 orders of magnitude[5].

The non-gravitational degrees of freedom are described by the SM of particle physics. It provides a partial unification (within the framework of gauge theories) of strong, weak, and electromagnetic interaction. The SM is a quantum *field* theory, that is, a quantum theory with infinitely many degrees of freedom. So far, there are no clear hints for physics beyond the SM. For theoretical reasons, one expects a unification of interactions at high energies. Some approaches to unification make use of supersymmetry (SUSY) in which fermions and bosons are fundamentally connected. Despite intensive search at the LHC, no evidence for SUSY was found.

Particle physicists measure energies in electron volts (eV). For high energies, one uses Megaelectronvolts (MeV), 1 MeV = 10^6 eV, Gigaelectronvolts (GeV), 1 GeV = 10^3 MeV, and Terraelectronvolts (TeV), 1 TeV = 10^3 GeV. The LHC reaches a collision energy of 13 TeV. Because of Einstein's famous relation $E = mc^2$, masses can be measured in eV over c^2. The proton mass is about 938 MeV c^{-2}, and the mass of the famous Higgs particle discovered at the LHC in 2012 is about 125 GeV c^{-2}.

The fields in the Standard Model all carry energies and thus generate a gravitational field. Because they are quantum fields, they cannot be inserted directly into the classical Einstein field equations. Only a consistent unification of gravity with quantum theory can describe the interaction of all fields at the fundamental level.

2.4.1.2 What are the main problems?

What do we mean when we talk about quantum gravity? Unfortunately, this term is not used in a consistent way. Here, we call quantum gravity any theory (or approach) in which the superposition principle is applied to the gravitational field.

The superposition principle is at the heart of quantum theory: for any physical states of a system (described, e.g., by wave functions Ψ and ϕ), any linear combination $\alpha\Psi + \beta\phi$, where α and β are complex numbers, is again a physical state. This principle is confirmed by an uncountable number of experiments. For more than one system it leads to entanglement between systems, which is relevant for atoms (e.g., the qubits used in quantum information), for particles (e.g., neutrino oscillations), and many other cases.

Now, because gravity couples universally to all degrees of freedom, this should entail also a superposition of different gravitational fields, for which a quantum theory of gravity is needed. At a famous conference at Chapel Hill (USA) in 1957, Richard Feynman explained this by a gedanken experiment (see figure 2.17).

In this, the superposition of microscopic states (e.g., electron spins) is transferred to the spatial superposition of a macroscopic ball, for which the gravitational field is measurable. But how do we describe the gravitational field of an object which is in a spatial superposition at different locations? Only a theory of quantum gravity can

[5] It is interesting to see that the geometric mean of the largest and the smallest explored distance corresponds to about 10 km, which is an everyday scale.

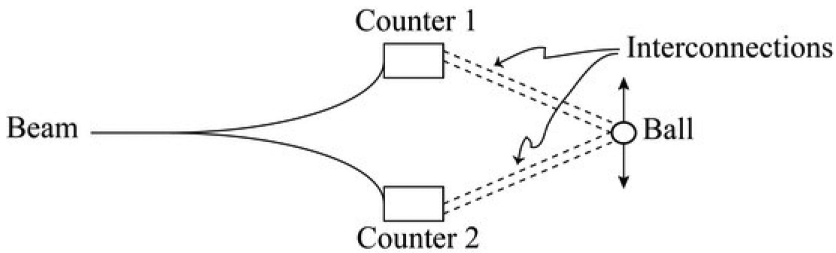


Figure 2.17. Stern–Gerlach type of gedanken experiment, in which the detectors for spin up respective spin down are coupled to a macroscopic ball. If the particle has spin right, which corresponds to a superposition of spin up and down, the coupling leads to a superposition of the ball being moved up and down, leading to a superposition of the corresponding gravitational fields. Figure adapted from DeWitt and Rickles, p 251, see DeWitt [68] CC BY 3.0.

achieve this. There exist attempts to realize superpositions à la Feynman in the laboratory; see, for example, Bose *et al* [69] and Marletto and Vedral [70]. Whether this is possible and whether one can draw conclusions on quantum gravity from this is currently subject of discussion.

There are other reasons in favour of the search for quantum gravity. As already mentioned above, if one aims at a unification of all interactions in Nature (a 'theory of everything' or TOE), one has to accommodate gravity into the quantum framework, since the quantum fields of the non-gravitational interactions act as sources for gravitational fields. One may, of course, envisage in principle a unified theory in which gravity stays classical. But there are at least two reasons that speak against this possibility. First, it is not very satisfactory to have such a fundamental hybrid theory. Second, there are various counter-arguments from the observational point of view against some hybrid theories (see, e.g., Kiefer [65] for details). But there exist no logical arguments that would force the quantization of gravity, and hybrid theories can indeed be constructed [71].

Einstein's theory, by itself, is incomplete. One can prove singularity theorems which state that, given some assumptions, there are regions in spacetime where the theory breaks down (Hawking and Penrose 1996). Concrete examples include the regions inside black holes and the origin of our Universe ('time zero'). Only a more general theory, such as a quantum theory of gravity, may be able to resolve these singularities and thereby allow a full description of black holes and the Universe.

There is also another kind of singularity. Quantum field theories are plagued by divergences which arise from probing space time at arbitrarily small scales, leading (by the indeterminacy relations) to momenta and energies of arbitrarily high values. On paper, these 'infinities' can be handled by regularization and renormalization. Regularization means that divergent expressions can be made finite by a mathematical procedure of 'isolating' the divergences (infinities). Renormalization means that the isolated divergences can be absorbed in physical parameters of the theory. These parameters cannot, of course, be calculated from the theory, but can only be determined empirically. The paradigmatic example is quantum electrodynamics (QED) and the parameters swallowing the infinities are the electric charge and the mass of the electron. Once this is done for finitely many parameters (typically a small number), the theory becomes predictive. While this procedure is consistent and can be successfully applied to the SM, the question arises whether a fundamental theory including gravity is finite by construction, that is, whether no divergences occur in the first place. Perhaps the root for both types of singularities (gravitational and quantum field theoretical) lies in the assumed continuum nature of space time.

Before we embark on a brief discussion of the main approaches, let us address the physical scales where we definitely expect quantum effects of gravity to become relevant (due to the universality of the superposition principle, such effects can, in principle, become relevant at any scale).

In the most recent version of the *Système International d'unités* (SI), which is valid since 2019, physical units are based as much as possible on fundamental constants[6]. In this, Planck's constant h, the speed of light (c), and the electric charge (e) are attributted fixed values. The units metre (m) and kilogram (kg) can then be inferred from h and c, while the second (s) is determined from atomic spectra. For us, h and c are relevant:

$$c = 299\,792\,458\ \mathrm{ms}^{-1}, \tag{2.1}$$

$$h = 6.626\ \ 070\ \ 040 \times 10^{-34}\ \mathrm{J} \cdot \mathrm{s}, \tag{2.2}$$

The gravitational constant G is known with much lower accuracy. On the *NIST Reference on Constants, Units, and Uncertainty*, one finds the following 2018 value for G:

$$G = 6.674\ \ 30(15) \times 10^{-11} \frac{\mathrm{m}^3}{\mathrm{kg} \cdot \mathrm{s}^2}. \tag{2.3}$$

It thus cannot serve the same purpose as h and c (otherwise, we could base our time unit on G). Einstein's theory also contains the cosmological constant Λ, which has dimension of an inverse squared length. From current observations one finds the value

$$\Lambda \approx 1.2 \times 10^{-52}\ \ \mathrm{m}^{-2} \approx (0.35\ \ \mathrm{Gpc})^{-2}, \tag{2.4}$$

which, however, is not precise enough for using Λ as a standard of units.

The three constants G, h (resp. $\hbar = h/2\pi$), and c provide the relevant scales for quantum gravity, because one can construct from them (apart from numerical factors) unique expressions for a fundamental length, time, and mass (or energy). Because Max Planck had formulated them already in 1899, they are called Planck units in his honour. The Planck length reads:

$$l_{\mathrm{P}} := \sqrt{\frac{\hbar G}{c^3}} \approx 1.616 \times 10^{-35}\ \ \mathrm{m}, \tag{2.5}$$

[6] See, e.g. Hehl and Lämmerzahl [72] for a thorough discussion.

the Planck time is:

$$t_{\rm P}:=\frac{l_{\rm P}}{c}=\sqrt{\frac{\hbar G}{c^5}}\approx 5.391\times 10^{-44}\ \text{s}, \tag{2.6}$$

and the Planck mass is:

$$m_{\rm P}:=\frac{\hbar}{l_{\rm P}c}=\sqrt{\frac{\hbar c}{G}}\approx 2.176\times 10^{-8}\ \text{kg}\approx 1.22\times 10^{19}\ \text{GeV}\,c^{-2}, \tag{2.7}$$

from which one can derive the Planck energy

$$E_{\rm P}:=m_{\rm P}c^2\approx 1.22\times 10^{19}\ \text{GeV}\approx 1.96\times 10^{9}\ \text{J}\approx 545\ \text{kWh}. \tag{2.8}$$

Whereas Planck length and Planck time are far remote from everyday (and experimentally accessible) scales, Planck mass (energy) seems to be of a more everyday nature. The point, however, is that the Planck mass is more than 10^{19} times the proton mass $m_{\rm pr}$ and more than 10^{15} times the maximal collision energy attainable at the LHC. This means that to generate particles with masses of order the Planck mass or higher, one needs to construct an accelerator with galactic dimensions. This is one of the most important problems in the search of quantum gravity: we cannot probe the Planck scale directly by experimental means.

The size of structures in the Universe is determined by the squared ratio of proton mass and Planck mass, sometimes called the 'finestructure constant of gravity':

$$\alpha_{\rm g}:=\frac{Gm_{\rm pr}^2}{\hbar c}=\left(\frac{m_{\rm pr}}{m_{\rm P}}\right)^2\approx 5.91\times 10^{-39}. \tag{2.9}$$

It is the smallness of this ratio that is responsible for the usual smallness of quantum-gravitational effects in astrophysics. It is an open question whether this number can be calculated from a fundamental theory or whether it remains unexplained as a phenomenological parameter that can only be determined from observations.

2.4.1.3 What are the main approaches and applications?

Before addressing the full quantization of gravity, it is appropriate to have a brief look at what is known about the relation between quantum theory and classical gravity[7].

The relation between quantum *mechanics* (quantum theory with finitely many degrees of freedom) and gravity is studied by using the Schrödinger (or Dirac) equation in a Newtonian gravitational field. This is the regime where experiments are available, such as by observing interference fringes of neutrons or atoms. The combination of quantum *field* theory (QFT) with general relativity ('QFT in curved spacetime') is much more subtle. The perhaps most famous prediction there is that black holes are, in fact, not black but radiate with a thermal spectrum. This effect

[7] References on this and the following sections can be found e.g. in Kiefer [65]. See also Carlip [73] and Woodard [74] for general accounts of quantum gravity.

was derived from Stephen Hawking in 1974 and is called Hawking radiation. The temperature of a black hole is given by

$$T_{\mathrm{BH}} = \frac{\hbar\kappa}{2\pi k_{\mathrm{B}} c}, \tag{2.10}$$

where κ is the surface gravity characterizing a stationary black hole. Within Einstein–Maxwell theory (coupled gravitational and electrodynamical fields), one can prove the *no-hair theorem* for stationary black holes: they are uniquely characterized by the three parameters mass (M), electric charge (Q), and angular momentum (J). Astrophysical black holes are described by the two parameters M and J (Kerr solution).

For a spherically symmetric (Schwarzschild) black hole with mass M, the Hawking temperature is

$$T_{\mathrm{BH}} = \frac{\hbar c^3}{8\pi k_{\mathrm{B}} GM} \approx 6.17 \times 10^{-8}(\frac{M_\odot}{M})\ \mathrm{K}. \tag{2.11}$$

The smallness of this value means that this effect cannot be observed for astrophysical black holes, which have a mass of at least three solar masses ($3M_\odot$).

Figure 2.18 shows an example of an observed black hole—a supermassive black hole with $M \approx 6.5 \times 10^9 M_\odot$ in the centre of the galaxy M87. For such black holes, the Hawking effect is utterly negligible.

There is, in fact, an analogue of the Hawking effect in flat (Minkowski) spacetime. An observer moving with constant acceleration a through the standard vacuum state of flat space time experiences a bath of thermally distributed particles with 'Unruh temperature':

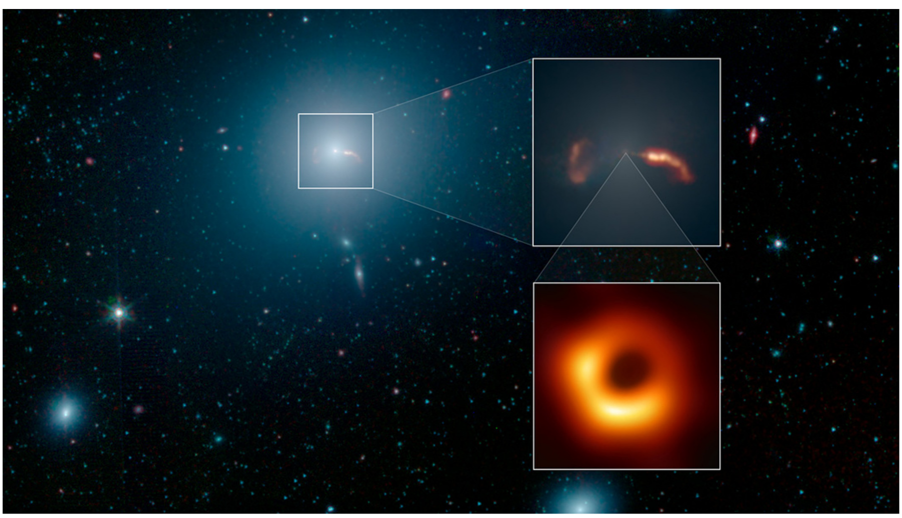


Figure 2.18. Shadow of the supermassive black hole in the centre of the bright elliptical galaxy M87. For this and all other black holes observed so far, only a consistent quantum theory can explain what happens in their inside regions. Image credit: NASA, JPL-Caltech, Event Horizon Telescope Collaboration.

$$T_{\mathrm{U}} = \frac{\hbar a}{2\pi k_{\mathrm{B}} c} \approx 4.05 \times 10^{-25} \ a\left[\frac{\mathrm{m}}{\mathrm{s}^2}\right] \ \mathrm{K}. \tag{2.12}$$

One immediately recognizes the similarity with (2.10), with a replaced by K. The reason for the appearance of this temperature is the fact that there is no unique vacuum (and thus no unique particle concept) for non-inertial observers in flat space time.

If black holes have a temperature, they also have an entropy, which is given by the 'Bekenstein–Hawking expression'

$$S_{\mathrm{BH}} = \frac{k_{\mathrm{B}} A c^3}{4G\hbar} \equiv \frac{k_{\mathrm{B}} A}{(2l_{\mathrm{P}})^2}, \tag{2.13}$$

where A denotes the area of the black hole's event horizon. In the Schwarzschild case, we can express the entropy as

$$S_{\mathrm{BH}} \approx 1.07 \times 10^{77} k_{\mathrm{B}} \left(\frac{M}{M_{\odot}}\right)^2. \tag{2.14}$$

S_{BH} is indeed much greater than the entropy of the star that collapsed to form the black hole. The entropy of the Sun, for example, is given approximately by $S_{\odot} \approx 10^{57} k_{\mathrm{B}}$, whereas the entropy of a solar-mass black hole is about $10^{77} k_{\mathrm{B}}$, which is 20 orders of magnitude larger. All the above expressions contain the fundamental units c, G, $\hbar$ and thus point towards the need for constructing a quantum theory of gravity. Such a theory should be able to provide a microscopic interpretation of the entropy formula (2.13).

Besides black holes, quantum effects are also important in cosmology. Assuming that the Universe underwent an (almost) exponential expansion at a very early state (a phase called *inflation*), density perturbations of matter and gravity (gravitons, see below) are generated out of quantum vacuum fluctuations. All the structure in the Universe (galaxies and clusters of galaxies) is believed to arise from these perturbations. The power spectrum of these density perturbations (also called 'scalar modes') can be derived to read

$$P_{\mathrm{S}} = \frac{1}{\pi}(t_{\mathrm{P}} H)^2 \varepsilon^{-1} \approx 2 \times 10^{-9}, \tag{2.15}$$

where ε is a 'slow-roll parameter' that is peculiar to the chosen model of inflation, and H is the Hubble parameter (expansion rate) of the Universe during inflation. One recognizes the explicit appearance of the Planck time t_{P}, equation (2.6), in this formula. The power spectrum of these density fluctuations is recognized in the anisotropies of the cosmic microwave background (CMB) radiation; see figure 2.19. The number 2×10^{-9} on the right-hand side of (2.15) comes from observations.

All of what has been said so far points towards the need for a quantum theory of gravity. But how can such a theory be constructed? The first attempts date back to work done in 1939 by Léon Rosenfeld, who was then an assistent to Wolfgang Pauli in Zürich. In two papers, he pioneered two approaches: the 'covariant approach' and

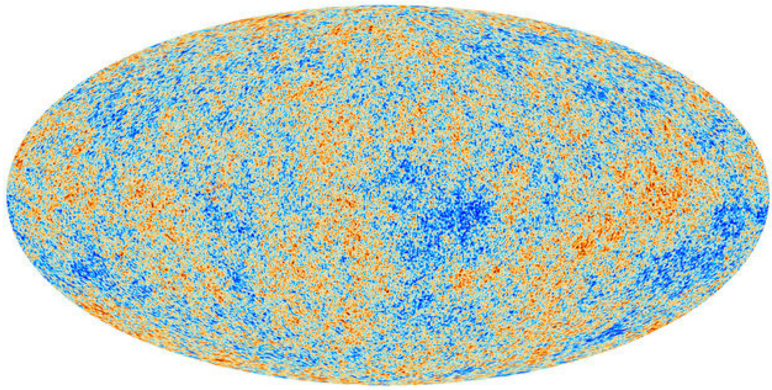


Figure 2.19. Anisotropy spectrum of the Cosmic Microwave Background (CMB). Image credit: ESA/Planck Collaboration.

the 'canonical approach'. Both approaches aim at the construction of a quantum version of general relativity. What is the status of these approaches?

The covariant approach has its name from the fact that a four-dimensional (covariant) formalism is employed throughout. In most cases, this formalism makes use of path integrals (in which, according to the superposition principle, four-dimensional spacetimes are summed over); see the appendix. Similar to the photon in quantum electrodynamics, a particle is identified as the mediator of the quantum-gravitational field, the *graviton*. It is massless, but has spin 2 (whereas the photon has spin 1). That it is indeed massless is indirectly confirmed by the detection of gravitational waves—they move with speed of light c. From this, the LIGO and Virgo collaborations report a limit of the graviton mass $m_g \lesssim 7.7 \times 10^{-23}$ eV. As remarked above, gravitons (also called 'tensor modes') are generated from the vacuum during an inflationary phase of the early Universe. Similar to the density spectrum (2.15), one can derive for them the power spectrum

$$P_T(k) = \frac{16}{\pi}(t_P H)^2. \tag{2.16}$$

A central quantity is the ratio between tensor and scalar modes,

$$r := \frac{P_T}{P_S} = 16\varepsilon. \tag{2.17}$$

So far, no observations have indicated a non-vanishing value for r. Observing such a value would constitute a direct test of quantum gravity at the linearized level.

As with all relevant quantum field theories, also the covariant quantization of general relativity exhibits divergences. But there is a major difference to the situation in the Standard Model. Whereas the perturbation theory for the SM is renormalizable, this does not apply for gravity. It is thus *not* possible to absorb divergent terms into a finite number of observable parameters; at each order of the perturbation theory, new types of divergences appear, and one would need infinitely many parameters to absorb them, rendering the theory useless. But the question arises whether higher terms in the perturbation expansion are indeed relevant. They come in powers of the parameter

$$\frac{GE^2}{\hbar c^5} \equiv \left(\frac{E}{m_{\rm P}}\right)^2 \sim 10^{-32}, \tag{2.18}$$

where E is the relevant observation energy, here taken to be 14 TeV, the energy of the planned LHC-upgrade. This is a very small parameter, so perturbation theory should in principle be extremely accurate. One could thus adopt the point of view that quantum general relativity is an *effective field theory* only, that is, a theory that is anyway valid only below a certain energy and must be replaced by a more fundamental, potentially renormalizable, or finite theory above that energy. An approach that makes use of standard quantum field theory up to the Planck scale is *asymptotic safety*. In this, G and Λ are not constants, but (as is typical for quantum field theory) variables that depend on energy. They may approach non-trivial fixed points in the limit of high energy and thus lead to a viable theory of quantum field theory at all scales. It is imaginable that the scale dependence of G could mimic Dark Matter; in this case, it would be hopeless to look for new particles as constitutes of Dark Matter.

To calculate quantum-gravitational path integrals is far from trivial and definitely not possible analytically. For this reason, computer methods are heavily used. One promising approach is *dynamical triangulation* which bears this name because the spacetimes to be summed over in the path integral are discretized into tetrahedra. This leads to interesting results about the possible microstructure of space time.

One candidate for a finite quantum field theory of gravity is supergravity, which combines SUSY with gravity more precisely; more precisely, a particular version called $N = 8$ supergravity. Heroic calculations over many years have shown that there are no divergences in the first orders of perturbation theory. Whether this continues to hold at higher orders and, moreover, whether this holds at all orders, is far from clear. Only new, so far unknown, principles can be responsible for this theory to be finite.

A candidate for a finite theory of quantum gravity of a very different nature is superstring theory (or M-theory). In the limit of small energy, the above covariant perturbation theory is recovered, but at higher energies, string theory is of a very different nature. Actually, its fundamental entities are not only one-dimensional entities as the name suggests, but higher-dimensional objects such as branes. Moreover, the theory makes essential use of a higher-dimensional space time (with 10 or 11 as the number of dimensions). The theory is not a direct quantization of gravity—quantum gravity appears only in certain limits as an emergent theory. In contrast to theories of quantum general relativity, string theory has the ambition to provide a unified quantum theory of all interactions (the TOE mentioned above). Such a theory should also allow to understand the origin of mass in Nature. One aspect of this is the *hierarchy problem*. We observe widely separated mass scales—neutrino masses (~0.01 eV), electron mass (~0.5 MeV), and top-quark mass (≈173 GeV), all of which are much smaller than the Planck mass (2.7). So far, the origin of this hierarchy is not understood. It is not clear whether there is new physics between the SM energy scale (as exemplified by the Higgs and the top-quark mass) and the Planck scale.

Out of string theory and the discussion of black holes grew insights about a possible relation between quantum-gravity theories and a class of field theories called conformal quantum field theories. The latter are defined on the boundary of the spacetime region in which the former are formulated. This is known as gauge/gravity duality, holographic principle, or AdS/CFT conjecture (see, e.g., [75]). Some claim that it will play a fundamental role in a full theory of quantum gravity.

The alternative to covariant quantization is the canonical (or Hamiltonian) approach. The procedure is here similar to the procedure in quantum mechanics where one construct quantum operators for positions, momenta, and other variables. This includes also the quantum version of the energy called Hamilton operator. In quantum mechanics, the Hamilton operator generates time evolution by the Schrödinger equation. In quantum gravity, the situation is different. Instead of the Schrödinger equation, one has *constraints*—the Hamiltonian (and other functions) are constrained to vanish. This is connected with the disappearance of spacetime at the fundamental level. Spacetime in general relativity is the analogue of a particle trajectory in mechanics; so after quantization space-time disappears in the same way as the trajectory disappears (recall the indeterminacy relations)—only space remains. This is sometimes referred to as the 'problem of time', although it is a direct consequence of the quantum formalism as applied to gravity. It is connected with the fact that already the classical theory has no fixed background, so there is no such background available to serve for the quantization of fields—different from the situation with the non-gravitational quantum fields of the SM. Background independence is one of the main obstacles on the route to quantum gravity.

If one uses the standard metric variables of Einstein's theory, one arrives at quantum geometrodynamics with the Wheeler–DeWitt equation as its central equation. Due to mathematical problems, the full equation remains poorly understood, but it can be applied to problems in cosmology and for black holes. An alternative formulation makes use of variables that show some resemblance with the gauge fields used in the SM. It is known under the name loop quantum gravity. At the kinematic level (before the constraints are imposed), it is well understood, but the exact construction of the Hamiltonian constraints and the recovery of quantum field theory in curved space time present problems. Applications of loop quantum gravity also include cosmology and black holes.

An important feature of the Wheeler–DeWitt approach is the possibility of building a bridge (at least at a formal level) from quantum gravity to quantum field theory in curved space time. In this way, spacetime (and, in particular, time) emerges as an approximate concept. This procedure is similar to the recovery of geometric optics ('light rays') from fundamental wave optics. In this, the separation of scales (the separation of Planck mass from masses of the Standard Model) is crucial[8]. This

[8] A review of this and other conceptual issues can be found in Kiefer [6, 7, 65] and the references therein.

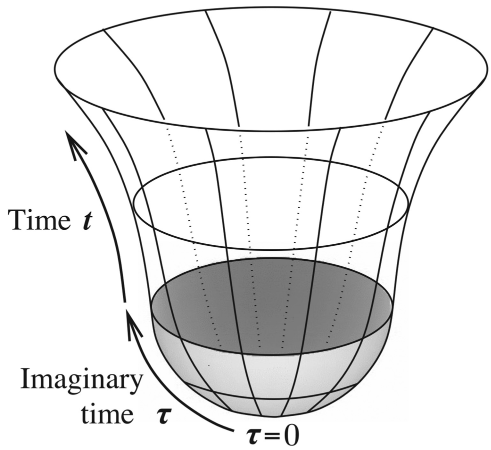


Figure 2.20. Hartle–Hawking instanton: the dominant contribution to the Euclidean path integral is assumed to be from half of a four-sphere attached to a part of de Sitter space. Reprinted with permission from [65], 2012, Oxford University Press.

emergence of time can also be described in the covariant approach. Using the method of path integrals, Hartle and Hawking constructed a certain four-dimensional geometry that elucidates the emergence of time by attaching a Euclidean ('timeless') geometry to a Lorentzian one. This 'Hartle–Hawking' instanton is shown in figure 2.20. It is frequently discussed in the application of quantum gravity to cosmology (quantum cosmology).

The Wheeler–DeWitt equation has a very peculiar structure. It is asymmetric with respect to the size of three-dimensional space and may thus allow to understand the origin of the arrow of time from fundamentally timeless quantum gravity [76]: there is an increase of entropy with increasing size of the Universe.

Besides the approaches already mentioned, there are a variety of others, and only space prevents me from discussing them in more detail. Many of these other theories make use of a discrete structure, either fundamentally imposed or derived from other principles. The reader may wish to consult Oriti [77] for more details.

2.4.2 Challenges and opportunities in the Horizon 2050

2.4.2.1 Theoretical challenges and opportunities

What can or should we expect in the coming decades? Physics is an experimental science. There can only be progress if we have testable predictions that can falsify a given approach and discriminate between different approaches. To derive such predictions is one of the main theoretical challenges.

It makes sense to distinguish between predictions at the linearized level and at the full level. The linearized level of quantum general relativity also follows from unified theories such as superstring theory, so tests at that level are very general. Looking at atomic physics, one can calculate the transition rate from an excited state to the ground state by emission of a graviton. In one example ([65], p 40) this gives a lifetime τ of the excited state as big as

$$\tau \approx 5.6 \times 10^{31} \text{ years.} \tag{2.19}$$

It thus seems forever impossible to observe such a transition. One should, however, not forget that the predicted lifetime of a proton in the simplest unified theory of particle physics (the minimal SU(5) theory) is about 10^{32} years, which one was able to falsify in the Super-Kamiokande experiment in Japan; it turned out that the proton has a lifetime of at least about 10^{34} years. The problem with (2.19) is that this decay is drowned in electromagnetic transitions, which are very fast. But if one could identify transitions in atomic or molecular physics that emit photons at no or low rate, there may be the option to observe gravitonic emissions in, for example, thin interstellar clouds. To the best of my knowledge, however, no one so far has attempted to identify and calculate such processes.

The power spectra (2.15) and (2.16) are, in a certain sense, already effects of linearized quantum gravity. The reason for this claim is that the calculation makes use of variables that combine gravitational (metric) and matter variables in a quantum sense. This is confirmed by the appearance of the Planck time t_P in these expressions. Calculations have also been performed to derive corrections to these expressions by going beyond the linear approximation. This has been achieved in particular for the canonical theory in both the geometrodynamic and the loop version. The corrections are proportional to the inverse Planck-mass squared and turn out to be too tiny to be observable at present. Similar correction terms should appear for the power spectra of galaxy distributions; so far, however, calculations of such terms do not seem to exist. Quite generally, one would expect that the first signatures of quantum gravity come from small effects. This was the case for quantum electrodynamics, where the theoretical understanding and the successful observation of the Lamb shift in atomic spectral lines led to the general acceptance of the theory.

A second major challenge is the construction of a viable full quantum theory of gravity, preferable one that gives a unified description of all interactions. On the one hand, it is not clear whether one can construct a separate quantum theory of gravity alone, without unification. Asymptotic safety may provide an example of a stand-alone theory, but most likely, such a separate theory would be an effective theory, one that is valid only below a certain energy scale. This would be sufficient for calculating small effects, but would lack an understanding of quantum gravity at the fundamental level. On the other hand, it is far from obvious that the programme of reductionism will continue to work and that a 'theory of everything' can be found. Superstring theory, the main candidate for such a theory so far, has not proven successful in the last 50 years.

The case of superstring theory also exhibits a deep general dilemma. One might expect that a really fundamental theory would enable one to predict most of the fundamental constants of Nature from a small number of parameters. One important example is (2.9), which sets the scale at which structures in the Universe appear, and which string theory cannot predict so far. But since one knows that only a very fine-tuned set of physical parameters (masses, coupling constants, etc) allow the existence of a Universe such as ours and the formation of life, this would leave the open question of why this is so. If, on the other hand, the fundamental theory does not lead to such a prediction and if, moreover, all possible parameter values are allowed in the world (which would then constitute a kind of 'multiverse'), it would leave us only with the *anthropic principle* as a way to understand the Universe (see, e.g., [78]). It may, of course, happen that we have a mixture of the two cases, so that most constants are determined by the fundamental theory and a few (such as the cosmological constant and the Higgs mass) can only be determined anthropically. A decision about this dilemma is one of the most important theoretical challenges, if not *the* most important one.

We have remarked above that general relativity is incomplete because it predicts the occurrence of space-time singularities. The general expectation is that a quantum theory of gravity will avoid singularities. The present state of quantum gravity approaches is not mature enough to enable the proof of theorems, but preliminary investigations in various approaches indicate that singularity-free quantum solutions can indeed be constructed. It is one of the main theoretical challenges of the next decades to clarify the situation and get a clear and mathematical precise picture of the conditions under which singularity avoidance follows. This would also throw light on one important open question in the classical theory—*cosmic censorship* (see, e.g., [79]). Black holes such as the one in figure 2.18 are characterized by the presence of a horizon from behind which no information can escape to external observers. The singularity predicted by general relativity is thus hidden. The hypothesis of cosmic censorship states that all singularities arising from a realistic gravitational collapse are hidden by a horizon, thus preventing the singularity from being 'naked'. Singularity avoidance from quantum gravity would immediately lead to the non-existence of hidden *and* naked singularities and would thus prove cosmic censorship to be true in a trivial sense.

2.4.2.2 Observational challenges and opportunities

Progress in quantum gravity can eventually only come from observations and experiments. As we have seen, quantum effects of gravity are usually small and become dominant only at the Planck scale. Laboratory experiments thus may look hopeless. One can try to generate superpositions of gravitational fields in the sense mentioned in connection with figure 2.17, but it is unlikely that this could enable one to discriminate between different approaches. One may also use laboratory experiments to decide whether the superposition principle is violated for gravitational fields as advocated, for example, in Penrose [79]. The main obstacle in this is to avoid standard decoherence effects from environmental degrees of freedom [66]. Laboratory experiments are also useful to test acoustic analogies to the Hawking and Unruhe effects, from which insight relevant for quantum gravity may be drawn.

The main observational input should thus come from astrophysics and cosmology, but also from particle physics. For this to be successful, large international collaborations are typically needed. We have already mentioned the anisotropy spectrum for the CMB, which was precisely measured by international projects such as PLANCK, WMAP, BOOMERANG, and others. Whether quantum gravity effects can be seen in future projects of this kind remains open. A major step would be the identification of a non-vanishing value for the r-parameter (2.17), from which the existence of gravitons could be inferred.

Another important class of experiments are gravitational-wave experiments. They are not designed primarily for quantum-gravity effects, but they may be helpful also in this respect by detecting, for example, a stochastic background of gravitons from the early Universe. One project is the Laser Interferometer Space Antenna (LISA) scheduled for launch in 2034[9]. A planned terrestrical project is the Einstein Telescope (ET) scheduled for starting observations in 2035[10].

Aside from cosmology, black holes are perhaps the most important objects for exploring quantum gravity experimentally. Due to Hawking evaporation, black holes have a finite lifetime. Taking into account the emission of photons and gravitons only, the lifetime of a (Schwarzschild) black hole under Hawking radiation is (see, e.g., [80])

$$\tau_{\mathrm{BH}} \approx 8895\left(\frac{M_0}{m_{\mathrm{P}}}\right)^3 t_{\mathrm{P}} \approx 1.159 \times 10^{67}\left(\frac{M_0}{M_\odot}\right)^3 \text{years}. \tag{2.20}$$

It is obvious that this lifetime is much too long to enable observations for astrophysical black holes. This would only be possible if small black holes exist, which most likely can only result from large density fluctuations in the early Universe—for this reason they are called *primordial black holes*. So far, observations gave only upper limits on their number and on the rate for their final evaporation. Since gamma rays are emitted in the final phase, gamma-ray telescopes are crucial for their detection (e.g., the Fermi Gamma-ray Space Telescope launched in 2008)[11]. There are also speculations about the presence of a primordial black hole with the size of a grapefruit in the Solar System ('Planet X'); whether this is really the case must be checked by future observations, such as by the upcoming Vera C. Rubin Observatory in Chile[12].

Hawking's calculations that led him to the black-hole temperature (2.10) break down when the mass of the black holes approaches the Planck mass (2.7). This means that the final phase can only be understood from a full theory of quantum gravity (beyond the approximation of small correction terms). Observations may then shed light on the 'information-loss problem', that is, whether the radiation remains thermal up to the very end (and may thus lead to loss of information about the initial state) or not.

Quantum-gravity effects may also be seen in particle accelerators. This may be due, for example, to the existence of higher dimensions or due to the presence of supersymmetry. So far, no hints for this or other quantum-gravity related effects were found at the LHC[13] or other machines. The upgrade High Luminosity Large Hadron Collider (HL-LHC) is planned to start operation in 2029. Plans for various other big machines scheduled for operation before 2040 exist.

9 https://www.lisamission.org
10 http://www.et-gw.eu
11 https://fermi.gsfc.nasa.gov
12 https://www.lsst.org
13 https://home.cern/science/accelerators/large-hadron-collider

2.4.2.3 A brief outlook on the year 2050

There is a quote attributed to Mark Twain—'Prediction is difficult—particularly when it involves the future'—which definitely also applies to predictions about the status of quantum gravity in 2050. Looking 30 years back (my postdoc years), most of the present quantum-gravity approaches did exist, some of them already for a while. Since then, there has been progress in both the mathematical formulation and the conceptual picture, but no final breakthrough was achieved. A hypothetical researcher time travelling from 1991 to 2021 would have no problems to follow the current literature on quantum gravity. But what about the next 30 years?

An optimistic picture would perhaps look as follows. We have a leading candidate for a quantum theory of gravity that provides an explanation of the cosmological constant (more generally, Dark Energy) and perhaps Dark Matter. It predicts testable effects for quantum-gravitational correction terms to power spectra of galaxies and the CMB and sheds light on the final phase of black-hole evaporation. Gravitons are observed as relics from the early Universe and in the form of tensor modes from the CMB. Primordial black holes are observed and their final phase can be studied in detail. Ideally, this theory should give a unified description of gravity and the other interactions.

A pessimistic version would look very differently. We still work on essentially the same approaches to quantum gravity as today and see no possibilities for testing them. The abovementioned projects for the 2030s and 2040s turn out to be very successful for astronomy and particle physics, but fail to shed light on quantum gravity. Already in 1964, Richard Feynman wrote (see [81]): 'The age in which we live is the age in which we are discovering the fundamental laws of nature, and that day will never come again. It is very exciting, it is marvellous, but this excitement will have to go.' What he means is that there are limits to performing experiments for fundamental physics coming from their sheer size and financial needs, and that these limits may appear rather soon. Still, I think, at least the next 30 years should remain exciting, and perhaps major progress, theoretically and empirically, will emerge from a totally unexpected side….

2.4.3 Appendix

In this appendix, I shall summarize some formulae which were omitted in the main text. For a clear and concise account of classical (Newtonian and Einsteinian) gravity I refer to Carlip [82].

The famous inverse-square law of Newtonian gravity reads

$$\mathbf{F} = -\frac{GM_1M_2}{r^2}\hat{\mathbf{r}}. \tag{2.21}$$

This force can be derived from a potential Φ, which obeys Poisson's equation

$$\nabla^2\Phi = 4\pi G\rho, \tag{2.22}$$

where ρ is the matter density.

In general relativity, Poisson's equation is replaced by the Einstein field equations

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}, \tag{2.23}$$

which are of a non-linear nature (a gravitational field generates again a gravitational field, and so on). A fundamental role is played by the metric $g_{\mu\nu}$, which instead of the one function Φ in the Newtonian case contains ten functions. The physical dimension of the energy–momentum tensor $T_{\mu\nu}$ is energy density (energy per volume), which is equal to force per area (stress). Einstein once spoke of the left-hand side as marble (because of its geometric nature) and the right-hand side as timber (because of the non-geometric nature of matter fields). In fact, $T_{\mu\nu}$ contains the fields of the SM. Because these fields are quantum operators, the Einstein equations cannot hold exactly but must be modified by an appropriate quantum equation.

Covariant quantum gravity can be defined by a path integral P, which contains a sum over all permissible metrics $g_{\mu\nu}$ and over all non-gravitational fields ϕ,

$$P = \int \mathcal{D}g_{\mu\nu}\mathcal{D}\phi \exp\left(\frac{\mathrm{i}}{\hbar}S\right), \tag{2.24}$$

where S denotes the total action of the system. In the canonical approach, one has constraints which are also fulfilled by the path integral, building in this way a bridge between the two approaches.

2.5 What is the Universe made of? Searching for dark energy/matter

2.5.1 Dark matter—general overview

Jochen Schieck[1]
[1]Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, Wien, Austria and Technische Universität Wien, Austria

2.5.1.1 Introduction

Understanding the nature of dark matter is one of the big unsolved questions of modern physics. The problem was first discussed in detail in the first half of the 20th century, and almost a century later the problem is still hotly debated. However, since then, significant progress has been made, mainly by confirming and constraining dark matter with different approaches at different scales [83]. We can now better classify and narrow down the dark matter problem. We consistently observe more gravitational pull than the pull we expect from visible matter. The various measurements all point to the same dark matter abundance, which is about five times more prominent than the visible matter, which we describe by the SM of particle physics [84]. However, the underlying nature of dark matter is still not understood.

The Swiss astronomer Fritz Zwicky [85] is frequently cited as a pioneer in dark matter. In the 1930s, he studied the dynamics of the Coma galaxy cluster, and he found a clear mismatch between the observed mass and the gravitational force generated by the mass required to prevent the Coma galaxy cluster from tearing apart. The gravitational force generated by the observed mass cannot compensate for the centrifugal forces that the galaxies experience due to their speed.

The conclusion that the lack of gravity stems from unobserved matter, particularly a new particle, is currently the best approach. This assumption is very well-founded and consistent with all our observations, but it stays an assumption. A different approach motivates a solution to the dark matter problem by changing the way we understand gravity [86]. An explanation of all observations with this approach, however, is far more challenging. In any case, gravity plays an essential role in the understanding of the origin of dark matter. We have an excellent theory that describes gravity at large scales called general relativity [87]. The Standard Model (SM) of particle physics, however, does not include gravity (see, e.g., [88]). The forces represented in the SM are much stronger compared to gravity, and a particle theory that formulates all forces, including gravity, in a standard description still not exists. A better description of dark matter, or even its discovery, could provide new clues to a combined standard definition of nature that contains all four fundamental forces.

The particle character of dark matter as a solution to the dark matter problem leads to the postulation of at least one new particle. The SM of particle physics provides no particle dark matter candidate. The various astrophysical observations of dark matter allow us to narrow down its properties. First of all, the particle has to be massive to generate gravity. Secondly, the particle cannot take part in the electromagnetic interaction; it has to be dark. We know that dark matter has to travel much slower than the speed of light, often referred to as dark matter being 'cold' [89]. The measurement of its content in the Universe is consistent during the different steps of the evolution of the Universe and dark matter has to be stable or have at least a lifetime comparable to the Universe's lifetime [90]. The fact that particle dark matter has not been observed yet means that interaction with ordinary matter is weak, but hopefully stronger than the gravitational interaction. A dark matter particle that interacts via gravitation only would be impossible to observe directly in any experimental setup.

The quest for dark matter is also a prime example of a topic bringing together astronomy, astroparticle physics, cosmology, and particle physics. While particle physics explores the fundamental particles and their interactions at the smallest scale, cosmology describes the Universe's evolution—the most extensive scale we are aware of. In astronomy, celestial bodies are measured and their properties are determined. Astrophysics uses physics, especially particle physics, to explain astronomical observations. Observations of dark matter at the largest cosmological scale point towards the existence of a new microscopic particle. Particle physics studies at the smallest scale allow us to discover and characterize the new particles, which gives feedback for understanding the Universe at the largest scales and astronomical observations—an excellent example of how the various disciplines are mutually beneficial.

The discovery of a dark matter particle that gives rise to five times more matter than the currently known matter would lead to a substantial change in the understanding of nature. Humanity will realize that the nature that we experience with our senses is only a fraction of the matter of the Universe. Dark matter observation would open an entirely new research field focusing on this new particle's characterization and the related interactions.

2.5.1.2 Observation of dark matter

Since the discovery of dark matter in the first half of the last century, several different observations based on different methods confirmed the existence of dark matter. In the following section, we will briefly present the various dark matter observations and discuss their impact for a deeper understanding of dark matter. We will start with the systematic study of galaxy rotation curves, an example that provides a very illustrative picture of the dark matter problem. The mismatch between the gravitational pull and the observed matter was already discussed in the literature before the study. However, only the measurement of the rotation curves of the velocity of stars at large distances from the galactic center convinced the scientific community of the presence of dark matter. In the 70s, a new tool became available, and the American Astronomer Vera Rubin and collaborators used it to measure the velocity distribution of stars within the Andromeda galaxy, the galaxy closest to our home galaxy [91]. The new spectrograph developed by her collaborator Kent Ford Jr allowed the measurement of the characteristic 21 cm spectral line from hydrogen, and therefore the velocity of the stars away from the center of the Galaxy. Like a stone circulating on a string and whose centrifugal force is compensated by the string, the force of gravity keeps the stars in their orbit around the galactic center. Contrary to expectations, the velocity distribution does not decrease with increasing distance from the Galaxy center but somewhat flattens to constant values. The determined velocity distribution away from the galactic center is higher than the one expected from the visible mass. Additional forces have to act on the stars to keep them in orbit. Later on, the flattened-out velocity distributions were observed in other galaxies besides the Andromeda galaxy as well, and the case for missing matter in galaxies became even stronger. Embedding the Galaxy in a halo of invisible matter that extends up to large radii was proposed as a possible solution to the problem of excessive velocity distributions—a halo made of dark matter. The Andromeda galaxy is about 2.5 million light years away, and these measurements originate from processes as old as 2.5 million years. Recent measurements of the Milky Way return a similar picture and point out that our local galaxy is also embedded into a halo of dark matter. Figure 2.21 shows the velocity distribution of the planets within our Solar System and the one for a galaxy embedded in a dark matter halo. While within our Solar System, the gravitation pull is mainly determined by the Sun's mass, the dominant part of the gravitational pull in the Galaxy comes from matter which is not visible, dark matter.

Today the most accurate measurement of the dark matter content in the Universe comes from a measurement of the photons from the cosmic microwave background (see [92, 93] for a review and references therein). With the Universe's expansion after the Big Bang, the Universe continues to cool down, changing the energy of the particles. The particles are in thermal equilibrium with the Universe, which means they have the same temperature and energy as the Universe's average temperature. About 380 000 years after the Big Bang, the thermal energy of the particles is low enough to capture electrons in the electromagnetic field of protons such that

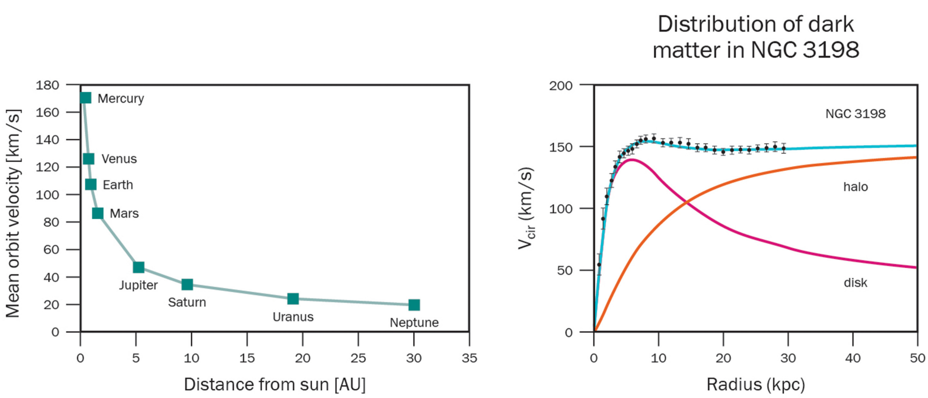


Figure 2.21. (Left) Velocity distribution of the planets orbiting around the Sun. The gravitational field is completely dominated by the Sun and the orbital velocity decreases with the distance from the Sun. (Right) The points represent the measurement of the velocity distribution of stars within the galaxy NGC 3198 [125]. For large distances the distribution flattens out and is not decreasing as expected from the visible mass indicated by the line named 'disk'. To explain the additional gravitational pull, to keep the stars on the orbit, the Galaxy has to be embedded in a halo of dark matter, indicated by the line name 'halo'. Only the combination of the gravitational pull originating from the disk and the halo can explain the observation.

hydrogen atoms can be formed, and the Universe undergoes a transition from a plasma with free charged particles to a gas of neutral hydrogen atoms. Suddenly the Universe becomes transparent to light since the charged particles are bound to neutral atoms. Photons no longer interact with scattering centers and suddenly propagate freely in space. The cosmic microwave photons represent an energy imprint at the the transition from an opaque to a transparent Universe. The measurement of these photons and a detailed subsequent analysis provides a precise determination of the Universe's matter content at the time of decoupling of the photons from the plasma. The average photon energy returns information about the average temperature when the plasma is transformed to a gas. In a gravitational field, the energy of a photon is slightly modified, as predicted by the general theory of relativity. Precision measurements of the photon energy therefore allow a mapping of the gravitational potential. Before the transition to the neutral hydrogen gas, the matter oscillates driven by the gravitational pull and radiation pressure from electromagnetic interactions. The analysis of these oscillations allows a precise measurement of the luminous matter and the non-luminous matter content of the Universe, dark matter. All matter feels the gravitational attraction, but dark matter only feels the attraction and no repulsion from the radiation pressure.

Currently, the most precise measurement of the energy and matter composition of the Universe comes from an analysis of cosmic microwave background data collected with the ESA Planck space observatory, with 26.8% [94] of the total energy density attributed to dark matter (figure 2.22).

At the transition time from the plasma to the neutral gas, the mass distribution in the Universe is almost uniform, and there are only minor density fluctuations of the gravitational field of the order of 0.001%. How did the nearly uniform Universe evolve in a Universe with matter accumulations and voids as we observe it today? The structure we observe today is the result of matter moving in the gravitational field of the Universe [95]. Sophisticated simulations can reproduce a matter structure

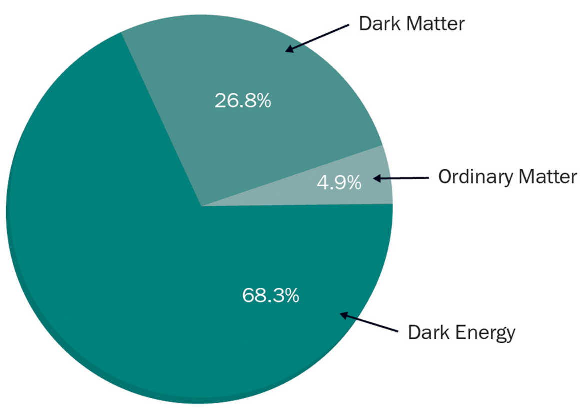


Figure 2.22. The pie chart shows the percentage of the various matter and energy contributions in the Universe as determined with data from the ESA Planck space observatory. Of the total energy and mass content, 26.8% is attributed to dark matter [94].

very similar to the one observed and, in addition, predict its evolution during the expansion of the Universe (see [96] for a review and references therein). However, the structure observed today can only be produced with dark matter included in the simulation. Visible matter only is not able to reproduce the Universe as we observe it today. In addition to the need for dark matter, we can also learn something about the dark matter properties. Dark matter must be cold, which means that it moves significantly slower than the speed of light. Otherwise, the smallest structures formed would be washed out in an early stage, and no galaxies and clusters of galaxies can form [89]. An example of the evolution of a dark matter distribution as expected from simulation based on cold dark matter is shown in figure 2.23. The similarity with observations is evident.

We know that other matter must exist from the observations discussed above, a matter that does not interact electromagnetically. However, how do we know that this matter is not made of a neutral accumulation of baryonic matter, which we know from the SM of particle physics?

A detailed study of the first nuclei generated about 3 min after the Big Bang, the so-called Big Bang nucleosynthesis, allows a precise determination of the abundance of light nuclei in the Universe [97, 98]. From this number, we can infer the Universe's total baryonic content of about 4, 9%, which is significantly lower

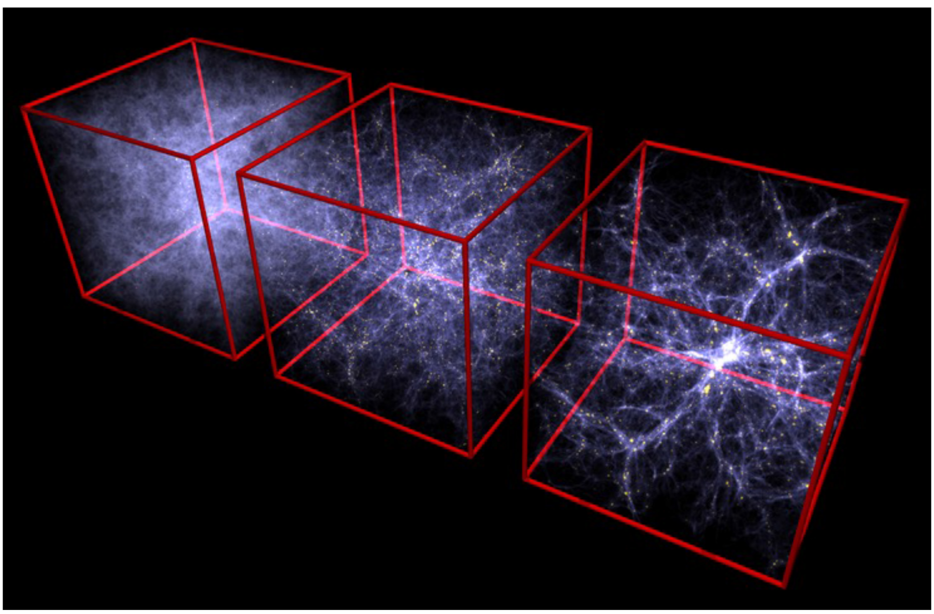


Figure 2.23. The picture shows a 3D view of the cosmic structure's evolution. The left box corresponds to the matter distribution of a part of the Universe with the age of one billion years (redshift $z = 6$), the middle one corresponds to about 3.3 billion years (redshift $z = 3$) and the right on as we would expect it today (redshift $z = 0$). The yellow areas correspond to stellar material [126] (Credit: Volker Springel, Max-Planck-Institute for Astrophysics), copyright (2008) reprinted by permission from Springer Nature.

compared to the total dark matter component [84]. Besides being invisible, dark matter is also not baryonic.

Dark matter has to exist to explain the observations presented above. All measurements agree that there is about five times more dark matter than the matter we know from the SM of particle physics. From the observations discussed above, we have some understanding of dark matter. The different measurements at different time scales during the evolution of the Universe are consistent. Dark matter consisting of one or multiple new, unobserved particles is currently the best working hypothesis for the search for dark matter.

We also have a good understanding of the dark matter in our own neighborhood, the Milky Way. Measurements show a dark matter density that corresponds roughly to the mass of a proton per cubic centimeter, which means that gravitational effects from visible matter dominate the Earth's atmosphere. The Solar System moves with a circular speed of about 230 km s^{-1} relative to the dark matter halo. Assuming a certain particle dark matter mass, we can then calculate the expected flux of dark matter particles here on Earth. The rotation of the Earth around the Sun of about 30 km s^{-1} leads to a seasonal modulation of the flux of about 5%–10%. An observation of an annual modulated signal originating from the movement of the Earth would be a clear dark matter signal [99].

2.5.1.3 Cold dark matter and open small-scale issues

On very large scales, the structure of the matter distribution can be very well described by simulations based on the Λ CDM model. The Λ CDM model is the cosmological model that provides the best description of our Universe's evolution. It offers an excellent illustration of all cosmological observations, and it is today's SM of cosmology. The Greek letter Λ refers to the cosmological constant, reflecting the so-called dark energy, the dominant part of the Universe's energy content, only discovered in 1998 (see figure 2.22). Another primary input to the Λ CDM model is cold dark matter, the central part of the understanding of several observations. For smaller scales, typically at the galactic and the sub-galactic scales, differences appear between the expectation from a collisionless Λ CDM dark matter simulation and the astronomical observations. These discrepancies are closely related to structure formation processes and to the corresponding matter being involved. A deeper understanding of astrophysical processes inside these regions is required to further improve modeling at sub-galactic scales. We already know that these simulations are incomplete, and contributions from baryonic processes, for example, are currently the subject of ongoing research. At large scales, the structure formation is entirely dominated by dark matter-caused dynamics; however, at smaller scales, baryonic processes play an increasing role and influence structure formation. Besides baryonic processes, specific properties of dark matter alter the evolution at small scales. Coherence effects of very light dark matter candidates could alter the structure at sub-galactic scales. Dark matter candidates could be warm and not cold, still in agreement with the expectations from structure formation and pointing towards a lighter dark matter particle, or dark matter might interact strongly with each other, leading to a different dynamical behavior in regions with sizable dark matter densities (see [100] for a review and references therein). Simulations using dark matter self-interactions on galactic and sub-galactic scales or with dark matter coherence effects might therefore guide the particle physics search for dark matter (see [96] for a review and references therein).

2.5.1.4 Approaches for solving the dark matter problem

Cosmology deals with our Universe's origin, evolution, and fate, and represents studies on the largest structure known to us. In addition to cosmology, which covers the largest scales known to us, the SM of particle physics plays a decisive role in unraveling the origin of dark matter. The SM of particle physics is one of the most telling physics theories ever developed and provides precision predictions with an agreement of several digits with experimental observations. It describes three of the known forces: the strong, the electromagnetic, and the weak force. Gravity, the by far weakest force, is not part of the SM of particle physics, which can explain all measurements on microscopic scales; however, it cannot be complete and is considered an effective theory of a more comprehensive theory. While explaining everything on a tiny scale, no particle in the SM could act as a dark matter candidate. In principle, the neutrino fulfills most requirements, but it nevertheless cannot act as a dark matter candidate. The SM neutrino is 'hot' and in contradiction to the condition of being a slow or 'cold' dark matter particle is necessary to explain the structure of the Universe we currently observe.

We are now in a situation where we have to have two excellent theories that describe our observations with very high accuracy. The Λ CDM model is based on general relativity, and relativistic quantum mechanics is the crucial ingredient to the SM of particle physics. We have a clear evidence for dark matter in the Λ CDM model, triggering an extended research program in particle physics. The potential discovery of a dark matter particle would have a significant impact on cosmology and therefore the Λ CDM model. Dark matter acts as a link between these two research areas dealing with the two most extreme length scales we are aware of—the largest and the smallest one.

Astrophysical observations demand the existence of dark matter; however, information on a concrete particle dark matter candidate is sparse. In particular, the mass and the size of the interaction of dark matter with SM particles are more or less unknown. It is impossible to cover the entire dark matter search region with a single experiment. In addition, any potential signal needs to be confirmed by another independent experiment and methodology before any final conclusion on particle dark matter can be drawn.

The search has to be guided by theory to set priorities for certain mass regions and interactions. New theoretical developments profit from previous work and guiding principles of physics, like symmetries and scales. These developments are complemented by new approaches, tackling the problem with alternative scenarios. A close collaboration between theory and experiment, particle physics, cosmology, and astronomy is essential to advance the field. Figure 2.24 shows the expected mass and the predicted interaction cross-section with SM particles for different dark matter models, guiding the different experimental approaches.

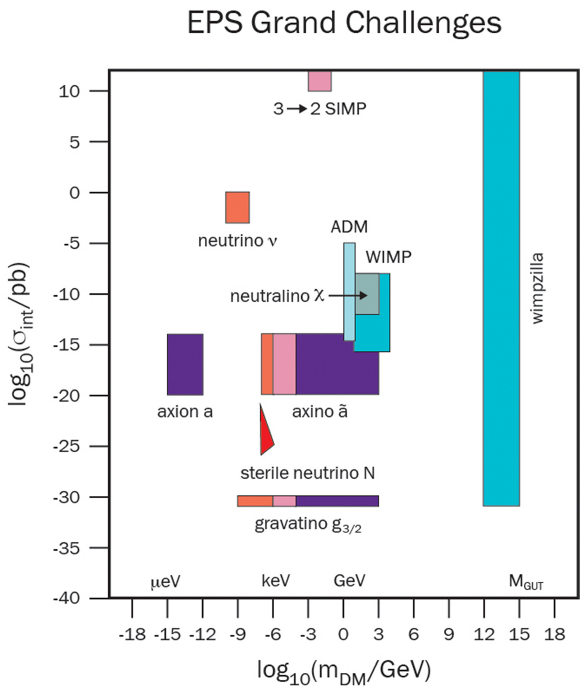


Figure 2.24. Possible dark matter models and its expectation for the corresponding mass and interaction cross-section. Adapted from [101], copyright (2015) with permission from Elsevier.

2.5.1.5 Particle dark matter candidates beyond the existing SM particle zoo

The mass range for dark matter candidates spans over several orders of magnitude. The lower mass limit is simply given by the space needed for each dark matter candidate, combined with the given dark matter relic density representing the amount of dark matter still remaining from the Big Bang. The fundamental properties of the particle, being a boson or a fermion, play a significant role. The upper limit is determined by the Universe's heaviest and most compact objects. Black holes produced during the earliest stage of the Universe ('primordial black holes') fulfill all requirements of being a dark matter candidate.

Particle dark matter candidates can be classified by different criteria. We already mentioned the classification via the temperature of the dark matter particle, like cold, warm, or hot dark matter, which reflects the particle's speed. Dark matter candidates can also be classified by the way how the dark matter relic density comes about. Thermally produced dark matter is such a possibility for the classification. All particles, including dark matter particles, are generated during the Big Bang. The particles are in thermal equilibrium with the temperature of the Universe. Dark matter particles can annihilate into SM particles, and vice versa, the temperature of the Universe determines the energy of the process. As the Universe expands, it cools down. Below a certain temperature, there is no longer enough energy available for the production of heavy dark matter particles, but dark matter can still annihilate into lighter SM particles. The Universe continues to expand and the dark matter density thins out. The annihilation process also becomes increasingly unlikely until it almost comes to a standstill and the amount of relic dark matter roughly stays constant. The expansion and cooling of the Universe leads to a freeze-out of dark matter particles from the thermal bath of the Universe. The remaining dark matter relic density is related to the properties of the interaction strength and the mass of the dark matter particle.

The most prominent thermally produced dark matter candidate is the so-called WIMP, a weakly interacting massive particle [102]. In this freeze-out scenario, the observed dark matter relic density can be obtained with a particle mass and weak interaction cross-section similar to the electroweak scale of the SM. The conjunction of these physical quantities is not regarded as a coincidence and motivates a class of models, leading to targeted searches for dark matter. Thermally produced WIMP-like dark matter particles are expected to have a mass between a few GeV up to several hundreds of TeV.

The supersymmetric extension of the Standard Model (SUSY) is based on the introduction of a new symmetry, which relates fermions and bosons—each particle of the SM gets assigned a supersymmetric partner. The supersymmetric partners of the neutral gauge bosons and the Higgs Boson of the SM can mix and form new particles, the so-called neutralinos. Under a certain symmetry, the lightest of these neutralinos has to be stable and has the expected characteristic of a WIMP particle precisely and is therefore a prime dark matter candidate. Despite an intensive search, SUSY and the neutralino have sill eluded discovery. However, the SUSY parameter space is large and this attractive theory is not ruled out yet.

Particles with an interaction strength to SM particles much smaller than the weak interaction will never reach thermal equilibrium and therefore cannot freeze-out as discussed above. These feebly interacting massive dark matter particles (FIMPs) are produced in a non-thermal process, like via decays of primordial heavier particles. This so-called freeze-in mechanism can also lead to the correct dark matter relic density with a similar mass range as WIMP particles. Only the interaction strength is much weaker, which leads to a more difficult experimental search.

Another non-thermally produced dark matter candidate is the so-called Axion, with a mass region between meV down to peV. A thermally produced dark matter candidate in that mass range would be 'hot', traveling with relativistic velocities, and in contradiction to the observed structure formation. Despite their lightness Axions are not 'hot'. Originally, they were introduced to solve another problem of the SM of particles physics, the so-called strong CP-Problem, and it turns out that Axions could also act as dark matter. Within the strong sector of the SM, a fundamental symmetry, the CP-symmetry, should be violated, but it is not. The non-existence of this broken symmetry can be explained by introducing a new particle, the Axion. Inspired by its properties, particles with similar properties can be introduced, so-called Axion-like particles (ALPs). While they are suited to explain the dark matter problem, they do not solve the strong CP-Problem [84].

There is also the idea to introduce a dark twin brother of the electromagnetic photon. This dark photon could act as a messenger between the visible and the invisible dark sector, or it could be dark matter itself. The SM neutrino has almost all the required dark matter particle properties, but as a thermally produced light particle it behaves like a hot dark matter candidate, which is in contradiction to structure formation. In the SM, each particle exists in two different orientations, so-called chirality states. The neutrino only interacts via the weak force and is therefore the only fundamental particle in which only one state of chirality can interact with other particles. The non-interacting neutrino chirality state, often referred as a 'sterile' neutrino, might be significantly heavier and a valid dark matter candidate. This scenario could also explain why normal neutrinos are so light in the first place and solve two open questions at once.

At even smaller dark matter mass scales, fuzzy dark matter models are introduced as a solution to the dark matter problem [103]. At these tiny mass scales, it is more appropriate to interpret particles as waves. The corresponding wavelengths are as large as known astronomical scales and coherence effects have to be taken into account. Fuzzy dark matter models offer a solution to the problems at sub-galactic scales discussed in the context of the Λ CDM dark matter model.

All models discussed above introduce a single new particle. However, with increasing knowledge about the particle dark matter problem, a solution with a single dark matter particle only is getting more and more challenging to defend. The answer might be rather a whole dark sector with several dark matter particles and fields (see [104] for a review and references therein). The interaction of the dark sector with the visible world might be feeble, even gravitational only. This could explain why the dark sector has so far eluded discovery and would make experimental discovery almost impossible.

2.5.1.6 Dark matter models beyond the particle hypotheses

In addition to the postulation of new dark matter particles, alternative explanations for the dark matter problem are discussed. Neutrinos are excluded as candidates for dark matter, but any neutral state built up from SM particles could be considered as well. The amount of SM matter is very well constrained by either the measurement of the total baryonic content of the Universe via the Big Bang nucleosynthesis or the analysis of the cosmic microwave background. Non-luminous astrophysical objects, MACHOs—'massive astrophysical compact halo objects', like neutron stars or dwarf stars—could also contribute to the dark matter density [105]. Primordial black holes are MACHOs and in particular the recent observation of mergers of medium-size black holes through gravitational waves intensified the discussion of primordial black holes, generated from gravitational collapses in the early Universe, as possible dark matter candidates and insensitive to the measurements from Big Bang nucleosynthesis. A MACHO or black hole only explanation for dark matter is strongly constrained, but not entirely excluded. A contribution to dark matter for a small MACHO mass range is still possible, and additional studies are ongoing (see [106, 107] for a review and references therein).

Modified Newtonian dynamics (MOND) tries to explain dark matter observations, in particular the mismatch between the observed matter and the corresponding gravitational pull, by modifying Newton's law of universal gravitation [86]. Several measurements of dark matter could be traced back to a deviation from the expected known $1/r$ gravitational behavior for large distances. An increased gravitational attraction for large distances would keep the systems from tearing apart.

While MOND modifies Newton's law of universal gravitation, extensions towards the inclusion of the underlying principle into general relativity exist as well [108]. While the solution of the dark matter problem with a modification of gravity or Newton's laws is elegant, a consistent explanation of all dark matter observations at different scales is much more challenging than introducing an additional particle dark matter candidate [109].

2.5.1.7 Search for dark matter

The observation of dark matter is undisputed; the range of proposed solutions is extensive. The existence of a new or several unobserved particles to explain the additional gravitational pull is currently a major direction of research. The difference in interaction strength with ordinary matter and in mass scales leads naturally to many different dark matter experiments focusing on different search regions. The interaction rate together with the involved energy allow us to draw conclusions about the abundance and the mass of dark matter particle candidates. To gather more information several possible approaches are used. Dark matter particles can annihilate to ordinary matter particles (so-called indirect detection; see [110] for a review and references therein), or vice versa, standard matter particles can annihilate to dark matter particles (so-called production; see [111, 115] for a review and references therein). Dark matter particles can also scatter from ordinary matter particles, including transfer of energy to the SM particle (so-called direct detection; see [112] for a review and references therein). The different interactions are sketched in figure 2.25. In principle, one can also gain new insight from the interaction between dark matter particles only, but a complete characterization or discovery of a dark matter particle with this process only is impossible.

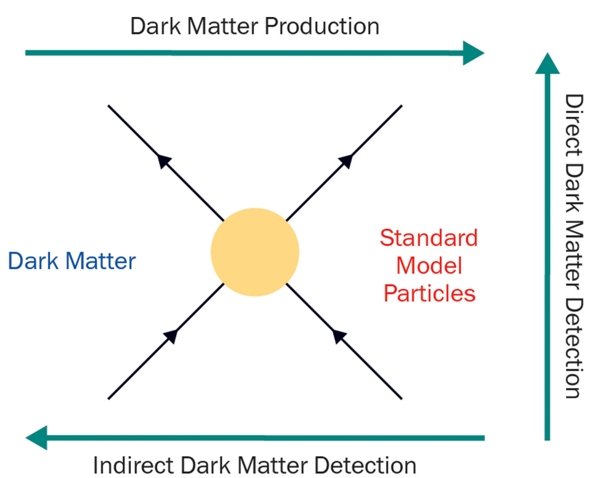


Figure 2.25. Sketch of possible interaction scenarios between dark matter particles and ordinary matter. The arrow indicates the direction of time, leading to three different interaction scenarios [84].

The annihilation of dark matter particles into SM particles is enhanced in areas with large dark matter density. This process is identical to the mechanism that would lead to a freeze-out of dark matter particles in the early Universe and to the relic dark matter content we observe today. Possible regions with enhanced dark matter density are the center of the Sun, the center of our Galaxy, or dwarf galaxies, which are expected to have a particular high dark matter content. The search is based on observing the annihilation products, leading to certain signatures, which can be well distinguished from background processes. This method includes the search for dark matter decay products, like antiparticles, which rarely exist otherwise. For a review and the references for the experiments mentioned below see [110]. Space-based experiments look for an excess of antimatter with energy distributions that indicate the annihilation of dark matter. The AMS-02 experiment at the International Space Station or the PAMELA experiment reconstruct the charge, identity, and the momentum of cosmic particles and determines the fraction of anti-particles over particles. The annihilation of a pair of dark matter particles would leave a clear signature in the anti-particle over particle spectrum. Another dark matter annihilation signature would be an increased number of dark matter annihilation products that originate from a region with increased dark matter density. These decay products must reach the observer as undisturbed as possible. Charged particles, however, are deflected by the galactic magnetic fields, and tracing back to the source is impossible. Therefore, neutral particles, such as photons or neutrinos, are an ideal way to look for dark matter decay products and their directional origin. The interpretation of indirect detection crucially depends on the understanding of the underlying astrophysical processes, like the propagation of matter and anti-matter in space or the dark matter density of potential dark matter sources. Photon detectors searching for dark matter signals are operating either as satellites in space, like the Fermi Gamma-ray Space Telescope or as Large Air Cherenkov telescopes, located at high altitudes, like the H.E.S.S. telescope in Namibia or the MAGIC telescope on the Canary islands. These telescopes detect Cherenkov light being emitted from extensive electromagnetic showers, initiated by high-energetic photons hitting the atmosphere of the Earth. Even larger detectors are necessary to detect weakly interacting neutrinos (e.g., Ice Cube at the South pole), searching for an excess of neutrinos originating from potential dark matter decays.

If dark matter annihilates into ordinary matter, the latter can also be converted into dark matter by the same underlying process—equivalent to turning around the time arrow of the process as indicated in figure 2.25. Additional kinetic energy is necessary to generate the mass of heavy dark matter particles. Since the main detection principle of particle detectors relies on electromagnetic processes, dark matter particles do not leave any signature in any particle physics experiment. However, energy and momentum conservation help to overcome this problem. Accelerator-based experiments have a well-defined initial state, and momentum or energy carried away by dark matter particles leads to energetically unbalanced final states with an unmistakable signature. Any observation of unbalanced energy or momentum proves the involvement of new particles which do not interact electromagnetically. These studies do not allow us to make a statement about the relic dark matter density and give only limited information of the lifetime of the particle. The

LHC at the European Center for Particle Physics (CERN) is a prime accelerator to search for dark matter particles. Protons with the highest energies collide at the two multi-purpose experiments, ATLAS and CMS, to produce dark matter particles in a controlled environment. Besides the LHC experiments, dark matter is also searched at other accelerators. Experiments with electron–positron collisions, like Belle II, or experiments at beam-dump facilities are also set up to search for missing energy and momentum signatures, pointing towards the production of dark matter particles. For a review and the references see [111, 115].

The working principle of experiments for direct dark matter detection is based on the fact that the Earth moves through the dark matter halo of our Galaxy. Dark matter particles from the halo can scatter with SM particles, and the measurement of the recoil energy of this scattering process leads to a featureless, exponential energy distribution. Due to the constant dark matter density on Earth, the expected particle flux for light dark matter particles is more prominent than for heavier particles, leading to a lower signal rate for heavy candidates. Taking into account the movement of the Earth around the Sun leads to well-defined annual modulation of the flux, indicating a clear dark matter signal [113]. The expected interaction strength between dark matter particles and ordinary matter is at most weak and despite the high flux of dark matter particles only few scattering events are expected [114]. Dark matter particles can scatter with the nucleus or the electrons of the target, and from the detection rate the interaction cross-section between dark matter and ordinary matter can be determined. Due to the low interaction rate the experiment has to be shielded from any potential background sources, like interactions induced by cosmic ray events. For this reason, these direct detection experiments are operated in underground laboratories, and for the construction of the experiment, material with very low natural radioactivity is used [112]. The energy transfer of the scattering event is deposited in the detector and allows to infer the particle dark matter mass. The capability to measure smallest energies is crucial in order to obtain sensitivities down to the smallest dark matter masses. Solid-state experiments, like Edelweiss, SuperCDMS, or CRESST, currently provide the best sensitivity towards smallest energy depositions and give the best limit for sub-GeV dark matter particles. For dark matter particles in the GeV mass range, experiments based on noble gases such as XENON1T, PandaX, or LUX provide the best sensitivity. A summary of the limits from ongoing direct dark matter searches is shown in figure 2.26 [84]. For a review and references see [112, 115].

Axions and ALPs interact weakly with matter and radiation. However, due to the lightness of these particles any energy deposition from scattering events is below any direct detection threshold and different experimental approaches have to be used. The main search strategy follows the two-photon interaction properties. In a strong magnetic field, Axions or ALPs can transform into a photon, and the produced photon can then be detected. The characteristic two-photon vertex is also used to generate Axions or ALPs artificially by sending a light source through a strong magnetic field. These artificially generated Axions or ALPs are sent through a light tight obstacle and another strong magnetic field is used to re-transform them back to photons, which can then be detected [116].

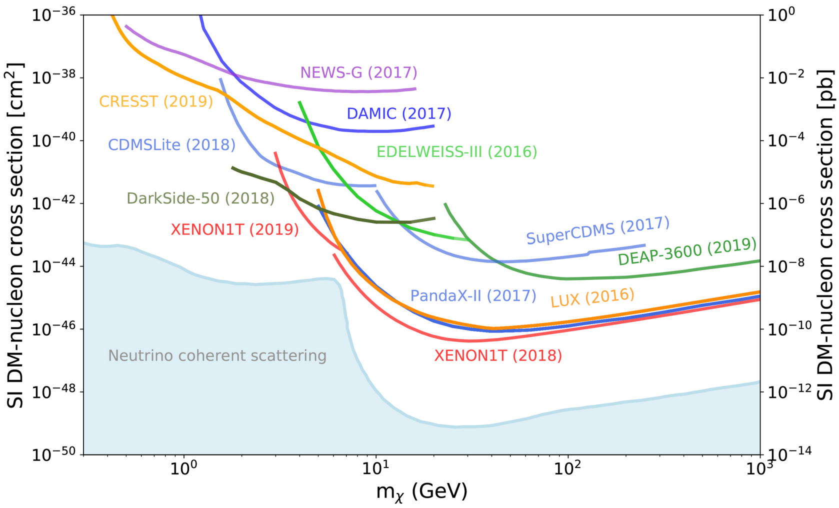


Figure 2.26. Overview of limits for direct dark matter searches performed by various experiments. The interaction strength of the dark matter particle with ordinary matter as a function of the mass of the dark matter particle is shown. The regions above the colored lines are excluded by the experiment. The blue shaded area at the bottom is the parameter range where neutrinos from Supernovae, atmospheric neutrinos, or neutrinos from the Sun leave a similar signature in the experiment. Reproduced from [84] CC BY 4.0.

The discovery of gravitational waves [117] opens a complete new possibility to study dark matter phenomena. While all measurements described above use non-gravitational interactions between dark matter and ordinary matter, gravitational waves will allow dark matter to be studied directly via gravity—the fundamental interaction all large-scale dark matter observations are based on. Besides large objects, like primordial black holes as potential dark matter candidates, gravitational waves can also make statements about possible microscopic dark matter particles. Scientific observations about the connection between dark matter and gravitational waves just started and still leave a lot of room for future theoretical and experimental studies [118].

2.5.1.8 Challenges and opportunities

Despite intense searches during the last decades, the origin of the additional gravitational pull caused by dark matter is still unknown. Intensive experimental investigations are carried out, but none has provided evidence for an undisputed particle dark matter discovery. Several theoretical approaches exist to solve the dark matter problem based on experience, scientific intuition, and extrapolation from existing knowledge. These theories are well thought out and justified to be scrutinized by different experiments. In the end, only the experimental results can give the direction towards the solution of the dark matter problem. A challenge for the scientific community is to propose suitable experiments to draw the correct conclusions from the results.

The search for a new particle as the solution to the dark matter solution provides a unique opportunity to bridge the largest and the smallest scales of fundamental physics. The proof of a dark matter particle would provide leverage to tackle significant conceptual challenges of physics. While gravity and the resulting dynamics at the largest scales are theoretically well described by general relativity, the smallest scales we can experimentally access are described by quantum mechanics and special relativity. Cosmology is based on general relativity and the SM of particle physics on relativistic quantum mechanics. Dark matter would open an opportunity to link these two fundamental theories. We know that at tremendous energies, at the so-called Planck Scale, these two theories can no longer be treated separately, and gravitational interactions have to influence quantum mechanics and vice versa. However, the required energies cannot be reached experimentally, now and not soon, and we need different input to relate these two fundamental concepts. The question for dark matter, driven by gravitational observation and the observation of a new dark matter particle, might open the possibility to give first insights into this problem.

The amount of theoretical models explaining dark matter is large and an extensive discussion of the scientific community's final confirmation and broad acceptance is expected. The fact that dark matter is reliably observed by various astronomical observations, combined with very little information on its origin, poses quite strong expectations on its scientific proof. Any potential dark matter explanation, being of particle character or any other solution, must consistently explain all existing dark matter observations at the same time. A great advantage arises from the fact that dark matter is searched by using different orthogonal experimental methods based on different assumptions. Some experimental approaches are based on the fact that we are embedded in a dark matter halo and any positive signal in direct detection experiments allows us to conclude on the relic dark matter density and provides a direct connection to astrophysical measurements. The production and observation of dark matter particles in an accelerator-based experiment is entirely independent of astrophysical conditions and provides additional input and consistency checks. A single approach can hardly solve degeneracies for specific fundamental parameters. Observations with different experimental methods are required to establish a consistent picture of dark matter and to dissolve degeneracies [119]. The establishment of dark matter measurements by different scientific approaches will provide a strict path towards the claim of discovery. The Initiative for Dark Matter in Europe and beyond (IDMeu)[14] is such a joint venture of the astroparticle physics community (represented by APPEC), the nuclear physics community (represented by NUPPEC), and the particle physics community (represented by ECFA) to tackle together the question of dark matter. This initiative includes joint scientific events, collaborative software tools as well as outreach activities.

Open science will play an increasingly important role in the future. This refers not only to open access of published results but even more to the open access to the underlying data used to produce the individual scientific results. Publicly available

[14] http://www.idmeu.org

data from different experimental sources will encourage collaboration between the respective disciplines. Scientists not being directly involved in generating these data will be able to provide input to their interpretation. Studies with cross-experiment data are becoming increasingly crucial for developing new knowledge.

The discovery will not only solve a long-standing problem of modern physics, it will rather establish a completely new research field, working on the detailed characterization of this new phenomena, including new insights in the underlying theories of quantum mechanics, general and special relativity.

The quest for dark matter has been ongoing for almost a hundred years with significantly increased interest during the last decades. The theoretical solutions to the dark matter problem push the experimental requirements more and more to the limits. The searches are carried out using plausible and target-oriented assumptions. It must be always kept in mind at all times what the evidence-based requirements and what reasonable assumptions are; the latter has to be questioned occasionally. Technical challenges are tempting, but the experimental search must always be driven by physics and not by technical challenges and possibilities.

A very close interaction between theory and experiment at all times is crucial for progress. Theory input is necessary to define new dark matter search strategies and for the interpretation of data from existing experiments in light of recent findings. In particular, theoretical studies are of utmost importance for bringing together experimental results from the different approaches. These studies include particle physics calculations for predicting and interpreting interactions with matter as well as theoretical astrophysical processes for the distribution and interaction of dark matter in the Universe.

2.5.1.9 Technical challenges, technology development, and knowledge transfer

Like all other fundamental sciences, technology drives the experimental progress in the field, and physics requirements from the research will drive technology. The experimental searches will advance into a new parameter space that can only be investigated with the help of new or improved experimental methods. The expected signatures for the different experimental approaches are known. There is also a good understanding of potential background sources, complicating the measurement of a dark matter signal. In addition, there will be 'unknown unknowns,' unexpected signals, which might lead to a misinterpretation of the result.

For direct dark matter detection, the current technologies will soon reach the so-called neutrino floor. An irreducible background of neutrinos from the Sun, hadronic interactions in the atmosphere, and supernovae will leave the same signature in the experiment as the one expected from dark matter particles. A discrimination between dark matter- and neutrino-induced events is no longer possible. To continue the search for dark matter with the highest sensitivity, the technology of the experiments must be adapted and further improved. Like the direction of the dark matter particle flux, additional information needs to be reconstructed to enhance the sensitivity for dark matter particles further.

Experimental searches for dark matter require cutting-edge technology, and these technological challenges can only be tackled together in extensive international collaborations. These experimental challenges for the various approaches differ significantly, from space missions for indirect dark matter searches, the construction of large-scale accelerators for the production of dark matter particles to the detection of smallest energy depositions from dark matter in our Milky Way, requiring an experiment operating in an extremely low background environment. The scale of the experiments is diverse, from tabletop experiments to massive experiments operated by large international collaborations. The challenge is to find the right balance between the approaches at different scales to ensure that all areas are covered and equipped with the necessary resources. Sometimes, these experiments are so extensive that only a few, or even only one experiment, can be implemented worldwide. The development, construction, and operation of such experiments pose new challenges for science. Financing across national borders is increasingly becoming the norm in large-scale equipment research. The scientific exchange across borders promotes scientific progress and encourages the interaction of young scientists and thus opens up the possibility that the best scientists in their field can work together on the most pressing questions. For some dark matter-related approaches, the work in large collaborations is unavoidable. The scientific work in large collaboration poses significant challenges for the scientific community, like evaluating the scientific performance of individual scientists in a large group. While scientific excellence is a crucial ingredient for recognizing personal achievements, successful work in large collaborations requires additional skills beyond the traditional assessment criteria, like coordination skills or empathy.

In addition to increasing the detector volume, improving the sensitivity to detect the smallest energy deposits plays an important role to cover more dark matter models. The conventional detection methods soon reach their limits, and new technologies and detection principles must be developed to overcome them. One possible step will be the development and application of quantum sensors that use quantitative phenomena, such as interferometry or entanglement. With this technology, the future search for dark matter will extend beyond particle and astroparticle physics towards new areas like quantum technology, further expanding the interdisciplinary nature of this topic [120].

Experiments for the direct search for dark matter must be shielded from cosmic radiation and have to be carried out in underground laboratories. With a growing number and ever-more extensive experiments, the need for laboratory space increases. The construction and operation of underground laboratories are complex, and it will be more challenging to meet all requests for laboratory space in the future. Worldwide there are only few suitable underground laboratories. Italy hosts the largest general-purpose underground laboratory in operation, the Laboratorio Nazionale del Gran Sasso (LNGS) [121]. More underground laboratories are currently under construction worldwide, and these laboratories will continue to be critical infrastructure for conducting dark matter experiments. Soon the largest and deepest underground laboratory will be operated by China.

The experiences gained during the research and development in the laboratory have to be transferred to mini-series to build these large-scale experiments. These quantities can no longer be produced by research laboratories alone, and often the production has to be carried out together with industry. Some technological challenges can only be solved together with industry and need close collaboration right from the beginning. The technical difficulties and the knowledge developed in developing new cutting-edge technologies for dark matter experiments are passed on to the industry. For example, the Astroparticle Physics European Consortium (APPEC) regularly organizes joint workshops between academia and industry[15]. In the past, developments driven by dark matter research led to successful technology transfers to society and the establishment of spin-off companies. Many developments in the field of particle detection find their way into medical technology.

Another example is the development and construction of radiation detectors based on noble gases, developed in technologies to search for dark matter. This technology development led to the spin-off now producing and selling radiation detectors[16]. The astroparticle community organizes regular joint meetings between scientists and industry on dedicated experimental topics to enhance scientific exchange further and increase knowledge transfer. The technology developed for dark matter experiments also inspires other related scientific disciplines. Scientists working on sub-GeV dark matter searches used the technology to investigate coherent neutrino scattering [122]. This technology can potentially enable tabletop experiments for neutrino physics, allowing studies of nuclear reactors without intervening directly in the reactor [123].

However, one key aspect of knowledge transfer is the education of young scientists. Young researchers, in particular students at various levels, from bachelor to PhD students, work on different dark matter-related projects as part of their training. The search for dark matter is undoubtedly one of the most exciting questions of modern physics and thus attracts many young people. The academic job market is too limited to offer all students a future in science, and most are leaving to work in the industry. The dark matter thus fulfills two critical aspects—the attraction and inspiration of young people for science and technology and an ideal, international environment to acquire all the skills necessary for a successful career in industry. Besides the science and technology aspects, the global environment of the dark matter research field, similar to other sciences, and the necessity to operate in large collaborations are beneficial for students' education.

The discovery of dark matter will open a new research field working on the detailed characterization of this recent phenomenon. Dark matter is not a minor add-on to ordinary matter; it is five times more abundant and is expected to play an essential role in the Universe's evolution. Unraveling the origin of dark matter will provide a new insight to our fundamental understanding of nature.

[15] https://www.appec.org/implementation/technology

[16] https://www.arktis-detectors.com

2.5.2 Dark energy

Emmanuel N Saridakis[1]
[1]National Observatory of Athens Lofos Nymfon, 11852 Athens, Greece

We provide a review on the dark energy, namely the unknown component that triggers the accelerated expansion of the Universe and comprises 70% of its energy content.

2.5.2.1 Introduction

In the history of science in general, and in the history of physics and astronomy in particular, the role of observations in changing our view has been crucial. In the Aristotelian-Ptolemaic cosmological and physical paradigm the Earth is spherical, motionless, and exists in the center of the Universe, around which revolve the spheres of the planets, Sun, and fixed stars (the Earth is composed of four elements and their combinations: Earth, Water, Fire, and Air, while the spheres are made of a perfectly transparent substance known as 'quintessence'—the 'fifth' element). It remained the absolute physical model for more than 1500 years, and actually it was the most long-lived scientific system in history.

It was only after the 11th century AD, where various Arab and Persian scholars incorporating observations performed in Maragha observatory, started putting into doubt the details of the paradigm such as the epicyles and the Earth's non-rotation. And despite the theoretical considerations of Copernicus, who was based on Aristarchos of Samos, it was only after the detailed observations made by Tycho Brahe, Johannes Kepler, and Galileo Galilei that the paradigm shift was established. In a similar way, despite the successes of Newtonian-Keplerian astronomical and physical model, a new paradigm shift was made necessary after the observations of the precession of Mercury's perihelion, which could not be explained in the previous framework.

The following decades were characterized by a significant advance in the quality and quantity of observations, leading to corrections and improvements in the new cosmological paradigm. Astronomers discovered extra galaxies beyond the Milky Way that are moreover moving away from each other (this was deduced through the Doppler effect, namely the shift of their emitted radiation wavelength towards larger values). Since every galaxy moves away from every galaxy, one can establish the framework of an expanding and cooling Universe originating from a primordial super-dense and super-hot state. This 'Big Bang' theory offered verified quantitative predictions (e.g., the abundance of primordial elements and the cosmic microwave background radiation) and was able to describe all observations.

However, theoretical investigation revealed that it might have some theoretical 'problems' (or at least issues whose explanation was not 'natural'), such as the horizon, the flatness, and the magnetic monopole ones. Thus, after 1980 the phase of inflation was established as a necessary ingredient of the cosmological paradigm at its early stages. Finally, in the last decades the cosmological paradigm underwent through another modification in order to incorporate the 'indirect observation' of the dark matter sector (see section 2.5.1). Hence, in the mid-1980s a concordance cosmological paradigm had been well established, namely an expanding Universe governed by General Relativity, in which one has all particles of the SM of particle physics (proton, neutrons, electrons, photons, and neutrinos) plus the unknown sector of Cold Dark Matter (CDM).

2.5.2.2 Accelerated expansion

The only open question was to determine the rate in which the Universe expansion is slowing down. If it was too small then the Universe would expand forever with reducing speed, while if it was sufficiently large then the Universe expansion would stop at a finite time, after which it should re-collapse resulting eventually in a Big Crunch. Finally, if the Universe expansion had a very specific value in-between, the Universe would expand with a speed that would asymptotically become zero at infinite time. The fact that the expansion should be decelerating was considered undoubtful, since gravity is an attractive force and hence the clusters of any form of matter (baryonic or dark) would tend to either re-collapse or move mutually away with lower and lower speed.

In the mid-1990s two Collaborations, namely the Supernova Cosmology Project and the High-z Supernova Search Team, started to measure in detail the 'deceleration parameter' of the Universe in order to provide a definitive answer to the above question. The way to do that was the following: if we find objects in the sky for which we know their moving-away speed, as well as their distance (and hence how long ago their image corresponds to, since their light needs some time to reach us) we can determine the expansion rate of the Universe at various times and hence deduce the 'deceleration parameter'.

Measuring speeds is quite easy in cosmology due to the Doppler effect and the redshift of the observed spectrum mentioned above. Measuring small distances is also quite easy due to the parallax phenomenon (perceived change in position of a relatively nearby object seen from two different places of the Earth's orbit) and simple geometrical calculations, but measuring large distances can be challenging. Nevertheless, one can still measure large distances if he can find objects emitting light with known flux. In particular, knowing how much radiation the object emits, observing how much radiation reaches us, and considering the inverse-square law for the reduction of light intensity with distance, we can calculate the object's distance. Fortunately, such 'standard candles' exist, and they are a particular class of supernovae explosions called Type Ia (these are believed to be caused by the core collapse of white dwarf stars when they accrete material to take them over a particular limit). Such objects have the same luminosity, although they are at different distances, and hence even if that luminosity is not known it will appear merely as an overall scaling factor.

Observing 50 such Type Ia supernovae in 1998 the Project teams resulted in a conclusion that was completely unexpected. The supernovae were less bright than expected, which implies that light had traveled more time in order to reach us, and hence the Universe expansion is not slowing down but on the contrary it is accelerating! This result was indeed verified later on by observations of completely different origin, namely the structure of the spectrum of the cosmic microwave background (CMB) radiation (there is an almost isotropic black-body microwave photon radiation coming to us from all points of the sky, with a temperature of around 2.7 K, which nevertheless has very small anisotropies that provide valuable information), the baryonic acoustic oscillations (BAO) (fluctuations in the density of the baryonic matter caused by acoustic density waves in the primordial plasma of the early Universe), etc, and this cross-verification gave to the leaders of the supernovae Collaborations Saul Perlmutter, Adam Riess, Brian P Schmidt the Nobel Price in Physics in 2011 [124].

The results delivered a major surprise to cosmologists. None of the usual cosmological models was able to explain the observed luminosity distance curve, called the apparent magnitude-redshift diagram, presented in figure 2.27. Speaking in quantitative terms, introducing the density parameters for the various components of the Universe (i.e., Ω_i for the 'ith' component), observations indicate that all known SM sectors (heavy elements, stars, free hydrogen, and helium, neutrinos,

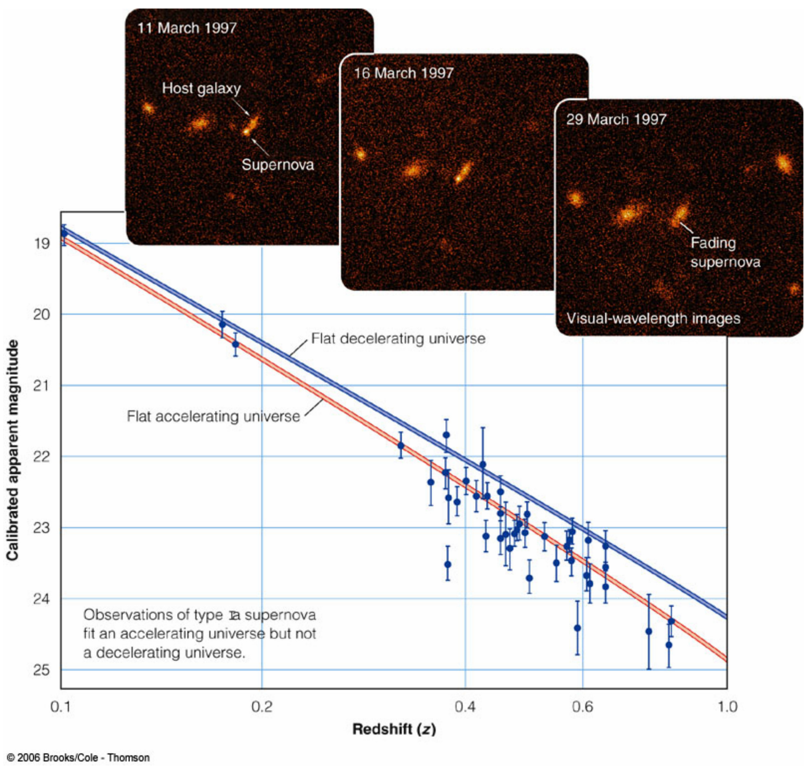


Figure 2.27. The apparent magnitude-redshift diagram for 50 Type Ia supernovae elaborated by the Supernova Cosmology Project and the High-z Supernova Search Team, i.e. the observed luminosity distance curve as a function of the redshift (larger redshifts correspond to larger velocities, thus larger distances, thus earlier times). The data show that the expansion rate of the Universe is currently larger than it used to be in the recent cosmological past, and thus the expansion is accelerating.

photons, etc) constitute only 5% of the total Universe content, cold dark matter constitutes approximately 25%, and the remaining approximately 70% corresponds to this unknown cause of acceleration. The Universe is not only accelerating, but the source of acceleration is by far its dominant sector. So what can it be?

2.5.2.3 ΛCDM concordance model

When Einstein was formulating the theory of general relativity, astronomers believed that the Universe was static. General relativity could not permit this, simply because it gives rise to attractive gravitational interactions and hence a static Universe should collapse under its own gravitational attraction. In order to allow for a static Universe, in 1917 he proposed a change to the equations by introducing a 'cosmological constant', called Λ, which at cosmological scales could compensate the gravitational attraction. However, when the Universe was found to be expanding, and thus general relativity could describe it without the need of such a term, Einstein removed it famously calling it the 'greatest blunder' of his life.

Interestingly enough, the cosmological constant Λ term is exactly what is needed in order to describe the accelerating expansion of the Universe. Since it compensates the attractive gravitational interaction, it can have exactly this effect. Equivalently, viewing it from the fluid perspective, the cosmological constant is completely different than any other form of matter, since it corresponds to negative pressure, which is what would be needed to obtain accelerated expansion. Hence, after the 1998 discovery, the cosmological constant was proudly re-introduced to cosmology as the main component of the Universe, giving rise to the current concordance model, the so-called ΛCDM paradigm. Once again, observations had played a crucial role in leading to a paradigm shift.

Now, in order to describe an expanding Universe in the framework of general relativity we introduce the concept of the four-dimensional space time. This is imposed to be homogeneous and isotropic in order to comply with the 'Cosmological Principle' (which states that at large scales the Universe is the 'same' everywhere). Hence, we use the Friedmann-Robertson-Walker (FRW) metric,

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta \, d\phi^2)\right], \tag{2.25}$$

in which t is time, r, θ, ϕ are the spatial dimensions in spherical coordinates, c is the speed of light, and $a(t)$ is the 'scale factor', the single quantity that quantifies the 'size' of the Universe (in an expanding Universe $a(t)$ is an increasing function of time). Finally, the parameter k accounts for the geometrical features of the three-dimensional space, and it is 0 for 'flat' space, +1 for 'closed' space, and −1 for 'open' space.

The essence of general relativity is that 'matter tells space time how to curve and the curved space time tells matter how to move'. More accurately, the distribution of the material that constitutes the Universe determines its evolution. Hence, according to the Standard Model of Cosmology, namely the ΛCDM paradigm, the Universe evolution is determined by the two Friedmann equations, which provide the differential equation for the scale factor [124]:

$$\left[\frac{\dot{a}(t)}{a(t)}\right]^2 = \frac{8\pi G}{3}\sum_i \rho_i(t) - \frac{kc^2}{a(t)^2} + \frac{\Lambda c^2}{3}, \tag{2.26}$$

$$\frac{\ddot{a}(t)}{a(t)} = -\frac{4\pi G}{3}\left[\sum_i \rho_i(t) + \frac{\sum_i p_i(t)}{3c^2}\right] + \frac{\Lambda c^2}{3}. \tag{2.27}$$

In these equations ρ_i and p_i are respectively the energy density and pressure of the various sectors that constitute the Universe material (baryonic matter, dark matter, photon radiation, neutrino radiation), which at large scales in a homogeneous and isotropic Universe depend only on time (which is consistent with the metric choice (2.25)). Moreover, G is the usual Newton's constant, which determines the strength of gravitational interaction. Finally, the equations close by considering the 'equation of state' of the various sectors, namely the expression $p \equiv p(\rho)$ that relates the pressure with the energy density.

The cosmological constant can perfectly describe quantitatively the observed acceleration of the Universe expansion. But what is its physical nature and more importantly why does it have the value it has? Although in mathematical terms Λ is just a constant that is introduced in the equations of general relativity, from the physical point of view it corresponds to the energy of space, or equivalently the vacuum energy. In any field theory one can estimate the vacuum energy following basic theoretical steps. The problem is that these basic field-theoretical estimations of the value of the zero-point energy (i.e., of the cosmological constant) lead to a number that (depending on the cutoff and other factors) is around 120 orders of magnitude larger than its observed value, which is found to be $\sim 7 \times 10^{-30}$ g cm^{-3}. This 'largest discrepancy between theory and experiment in all of science' is the renowned 'cosmological constant problem'.

As we can see, although Λ CDM paradigm can describe the Universe evolution and features, it conceptually suffers from the aforementioned problem on the nature and value of the cosmological constant. Seeing from the fluid point of view, according to Friedmann equations (2.26) and (2.27) the cosmological constant corresponds to a 'fluid' with energy density $\rho_\Lambda = \frac{c^2}{8\pi G}\Lambda$ and negative pressure $p_\Lambda = -\frac{c^2}{8\pi G}\Lambda$, and thus with equation-of-state parameter $w_\Lambda \equiv p_\Lambda/\rho_\Lambda = -1$.

2.5.2.4 Dark energy models

Having the above in mind, and trying to come closer to a microphysical description of the cosmological constant, or more fundamentally to the source that drives the Universe acceleration, one can substitute the cosmological constant by the more general concept of 'dark energy'. This umbrella term accounts for all possible explanations of the Universe late-time acceleration, either the cosmological constant itself or any other alternative. The term is adequate to describe the main features of this unknown source of the Universe acceleration, namely that it is a kind of 'energy' and that it is 'dark', in the sense that it does not interact with electromagnetism (in a similar way that dark matter is a form of matter which is 'dark').

In order for a mechanism, field, fluid etc, to be the dark energy, it should satisfy some necessary requirements. First of all it should not interact with Standard Model particles, including photons, or if it does the interaction cross-section should be extremely small; otherwise, we should have observed it. Additionally, seeing from the fluid point of view, dark energy should correspond to negative pressure, and in particular its effective equation-of-state parameter today should be close to -1 (i.e., to the cosmological constant value), since this is what observations show. In particular, in figure 2.28 we present the observational constraints on the value of the dark-energy equation-of-state parameter at the present Universe w_0, versus the matter density parameter Ω_m (thus the dark-energy density parameter is approximately $1 - \Omega_m$). Any dark energy model should satisfy these constraints. Hence, one could follow two main ways in order to obtain the above basic features: Construct models with known microphysical behavior, or construct phenomenological, effective models, with unknown microphysics but efficient to play the role of dark energy.

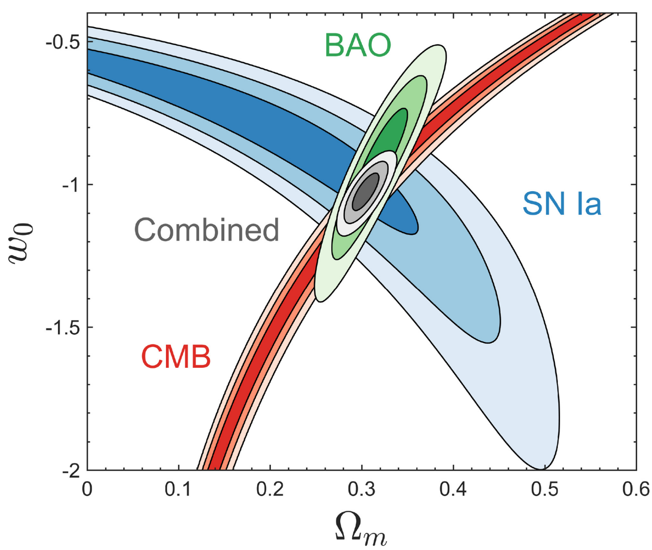


Figure 2.28. Constraints on the dark-energy equation-of-state parameter at the present Universe w_0, versus the matter density parameter Ω_m, arising from various observational datasets, namely from supernovae Type Ia (SNIa), cosmic microwave background (CMB) radiation, and baryonic acoustic oscillations (BAO) (see text) and their combination. The contours correspond to 68.3%, 95.4%, and 99.7% (1σ, 2σ and 3σ) likelihood, respectively, and one assumes a spatially-flat universe. Reproduced from [147], copyright IOP Publishing Ltd. All rights reserved.

The most well-studied model that can be a candidate for dark energy is a simple scalar field (similarly to the case of the inflaton scalar field, which is the most well-studied mechanism for the inflation realization in the early Universe). In particular, the whole Universe is filled with a scalar field ϕ whose dynamics drives the Universe acceleration. This scalar field is not a part of the SM of particle physics, and it is called 'quintessence' (Aristotle's revenge one could say!). Hence, one introduces the Lagrangian

$$\mathcal{L}_{\text{quint}} \equiv -\frac{1}{2}\partial^{\mu}\phi\partial_{\mu}\phi - V(\phi), \tag{2.28}$$

with μ taking the four coordinate values 0, 1, 2, 3, and where $V(\phi)$ is the potential of the quintessence field. Note that no interaction term between ϕ and SM particles is introduced, as required. In the cosmological FRW metric (2.25), in which the quintessence field depends only on time, the above Lagrangian leads to the equation of motion:

$$\ddot{\phi}(t) + 3\frac{\dot{a}(t)}{a(t)}\dot{\phi}(t) + \frac{dV(\phi)}{d\phi} = 0. \tag{2.29}$$

Additionally, the energy density and pressure of the scalar field, respectively, are

$$\rho_{\text{quint}} = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{2.30}$$

$$p_{\text{quint}} = \frac{\dot{\phi}^2}{2} - V(\phi). \tag{2.31}$$

Hence, the Friedmann equations in a quintessence Universe will be (2.26) and (2.27), but instead of the cosmological constant terms ρ_Λ and p_Λ one should use the quantities above. Finally, the equation-of-state parameter of quintessence dark energy is simply

$$w_{\text{quint}} = \frac{\frac{\dot{\phi}^2}{2} - V(\phi)}{\frac{\dot{\phi}^2}{2} + V(\phi)}. \tag{2.32}$$

Interestingly enough, by choosing suitably the quintessence potential (e.g., cases that lead to $V(\phi) \gg \dot{\phi}^2$), we can obtain $w_{\text{quint}} \approx -1$ and in this case the quintessence field plays the role of dark energy (i.e., it drives the Universe acceleration). We mention that although the scenario at hand is quantitatively similar to the cosmological constant, physically it is radically different since now the dark energy density is not the vacuum energy but the energy density of a scalar field, and thus there is no cosmological constant problem, and no 120-orders-of-magnitude error. Nevertheless, we should say here that if one wants to describe quantitatively the Universe acceleration then he should use potentials that in particle terms correspond to quintessence mass of the order of the cosmological constant, namely $\sim 10^{-33}$eV. Hence, we do not face the cosmological constant problem but we do face a huge hierarchy problem, namely we need to explain why in the Universe we have a massive particle that is tens orders of magnitude lighter than all the others (some people call this 're-parametrization of our ignorance').

Inspired by the above simple dark energy scenario, in the literature there appeared hundreds of dark energy models based on field Lagrangians. For instance, one can consider a large number of possible potentials, he can assume non-minimal interactions of the field to gravity, consider generalized kinetic terms (the so-called K-essence models), consider the case where the scalar field is tachyonic, phantom, or dilaton, use more than one scalar fields, use vector fields, etc. Recently, the most general scalar-field dark-energy models were investigated, the so-called Horndeski theories, or Generalized Galileon theories, in which one has many possibilities and many parameters and free functions of the field in order to successfully describe the cosmological data.

One different approach to the dark energy is the classic historical approach of introducing an unknown 'fluid'. In particular, since we know the basic requirements that dark energy sector needs to fulfill at macroscopic-cosmological level, we can introduce by hand a fluid with the specific phenomenological macroscopic properties, without caring about its microphysical description. Since a fluid in a homogeneous and isotropic Universe is determined just by its equation-of-state parameter (at least for the simple 'barotropic' fluids used in cosmology), one just introduces a desired form for the equation-of-state parameter of the dark-energy fluid, namely $w_{\rm DE}$, in order to phenomenologically fit the data. In figure 2.29 we depict the

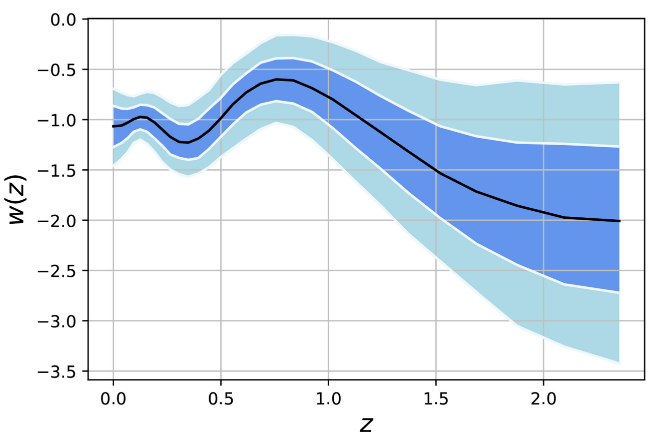


Figure 2.29. The data-driven reconstructed dark-energy equation-of-state parameter $w_{\rm DE}$ as a function of redshift z. Today $z = 0$, while $z = 1$ corresponds to approximately 8 Gyrs ago, and $z = 2.5$ to approximately 11 Gyrs ago. The contours correspond to 68.3% and 95.4% (1σ and 2σ) likelihood respectively, while the black curve denotes the best-fit value. Reprinted by permission from Springer Nature [148], copyright (2017).

data-driven reconstructed w_{DE} as a function of the redshift. Hence, any dark energy phenomenological model should satisfy these constraints.

Definitely, as we discussed above, the cosmological constant itself lies within this effective fluid description, with $w_{DE} = \text{const} = -1$. The first extension would be to consider dark-energy fluids with $w_{DE} = \text{const} = w_0$, the so-called wCDM dark energy models. One could then proceed to a large variety of w_{DE} parametrizations, like the Chevallier–Polarski–Linder (CPL), Linder, Linear, Logarithmic, oscillating, etc, ones, in which w_{DE} is considered to be time-varying (equivalently scale-factor-varying or redshift-varying) with a particular form. For instance, in the CPL parametrization one assumes that

$$w_{DE} = w_0 + w_a(1 - a), \tag{2.33}$$

with a the Universe scale factor and w_0, w_a two parameters. Alternatively, one could assume that the dark energy fluid has richer properties, satisfying, for instance, the equation of state class of the Chaplygin gas, whose pressure depends on energy density as

$$p_{Cg} = -A\rho_{Cg}^{-\alpha}, \tag{2.34}$$

with A, α two parameters. All these effective dark energy models remain at the phenomenological level, without offering an explanation for the microphysical nature of dark energy. However, they are very efficient in describing the dark energy features, as well as the cosmological data, and thus they are extensively used.

Let us now refer to a completely different approach to dark energy and Universe acceleration that has attracted a huge amount of research effort the last two decades. The essence of general relativity, which lies in the foundation of cosmology, is that the Universe evolution (i.e., the dynamics of space time) is determined by the distribution of the Universe content. In the dark energy approaches described above the strategy was to introduce an extra component in the Universe content, with suitable properties that can change the space-time dynamics in the desired way. However, one could follow a different strategy: instead of assuming that there are extra components in the Universe, to consider that the gravitational interaction is not general relativity, but a modified/extended one, which will lead to different space-time dynamics and thus to different Universe evolution.

Hence, in the modified gravity approach to dark energy, the latter arises in an effective way due to the extended gravitational interaction. In particular, one constructs new gravitational theories, which of course accept general relativity as a particular limit (in a similar way that general relativity tends to Newtonian gravity at a particular limit), but which in general have extra degree(s) of freedom and richer dynamics that can trigger the acceleration of the Universe without the need of extra fields, fluids, particles, etc. Additionally, contrary to the phenomenological/fluid approaches to dark energy, in the modified gravity approach one does obtain a microphysical description and a full explanation, which is exactly the modified gravity.

Having in mind the structure of general relativity, by changing it in various ways one can obtain numerous classes of modified/extended gravity. For instance, one can generalize the Einstein–Hilbert action through the addition of extra geometrical terms and their possible couplings to scalars and vectors such as in tensor-vector-scalar (TeVeS) theories. Alternatively, one may consider extra vector or tensor degrees of freedom, such as in massive or bimetric gravity. Moreover, one can use richer geometries, e.g. incorporating torsion, non-metricity, and non-commutativity, or consider extra dimensions, such as in Braneworld models or in string theory, etc. In all these theories and models, in the end of the day one chooses the details, as well as the model parameters, in order to be able to obtain effective dark energy features in agreement with observations. Actually, since cosmology (amongst others) is our laboratory for gravity, confrontation with detailed observational data is a necessary and important test for every gravitational theory. Even if a gravitational theory is constructed with different goals and motivations than to describe dark energy, at the end of the day it should be confronted with Universe acceleration data just to ensure that it does not lead to wrong effective dark energy features.

2.5.2.5 Conclusions

Let us close this section with a historical apposition. In the second half of the 19th century, astronomers started to observe differences in the precession of Mercury's orbit comparing to the predicted behavior. This led Urbain Le Verrier in 1859 to propose that this was a result of another, extra, planet between Mercury and Sun, namely the 'Vulcan' (note that Verrier in 1846 was able to describe the discrepancies with Uranus's orbit by predicting through purely theoretical calculations the existence and current position of a new planet, namely Neptune, which was indeed discovered by Johann Gottfried Galle at exactly the predicted coordinates). However, the answer to the problem of Mercury's precession was not an extra planet, a planet that we had not seen, and thus in a sense a 'dark' planet, the answer was modified gravity, namely general relativity. This new gravitational theory had different foundations, physical interpretation, and mathematical structure from Newtonian gravity; nevertheless at the level of predictions it was served as its modification, exhibiting the latter as a particular limit and offering corrections that were larger at scales that had just then started to become accessible by technological advance.

Interestingly enough, from the beginning of the 21st century and until now, we are in a similar situation that physicists were in the the beginning of the 20th century. Technological advance made it possible to access scales and perform accurate observations that suggest modification of the concordance cosmological paradigm. In general, we have two main ways to explain it: either we will consider new forms of material that in the framework of general relativity will do the job, or we will modify the theory of gravity itself. In figure 2.30 we present a schematic classification of all such dark energy approaches. Will the final outcome be similar to the one of the previous century, namely modified gravity, or it will it be an unknown, for the moment, 'dark' material? The combined theoretical and observational research will provide the answer. Stay tuned!

Figure 2.30. A schematic, not unique, classification of the dark energy approaches, representing the current status of cosmological research. These models arise from various extensions and modifications of the concordance paradigm of ΛCDM cosmology, which is based on general relativity (GR), cosmological constant (Λ), cold dark matter (CDM), and all particles of the Standard Model of particle physics. Details can be found in [149].

Acknowledgment

This section is partially supported by the National Natural Science Foundation of China and by the University of Science and Technology of China Fellowship for International Cooperation.

2.6 A gravitational universe: black holes and gravitation waves

Nelson Christensen[1]
[1]Artemis, Université Côte d'Azur, Observatoire Côte d'Azur, CNRS, CS 34229, 06304 Nice Cedex 4, France

A century after their prediction by Albert Einstein, numerous gravitational-wave events have been directly detected by the two Michelson interferometers that form the LIGO in the United States, and the European interferometric detector Virgo. This is the start of a global network of gravitational-wave detectors that begins the era of gravitational-wave astronomy. The majority of the events observed have been from binary black hole mergers. Two binary neutron star mergers have also been observed, one of which was in coincidence with a gamma-ray burst, and the remnant *kilonova* was found. Gravitational waves are the best way to get information on black holes, and test general relativity. The current gravitational-wave detector network will grow to include the KAGRA detector in Japan, and a third LIGO detector to be located in India. The third generation of ground-based gravitational-wave detectors are presently being designed, with the hope to be operational in the mid-2030s. The European Space Agency is leading the development of the LISA, which will operate in space, and observe a lower frequency band. LISA is planned for launch also in the mid-2030s. Pulsar timing arrays will likely soon be observing gravitational waves at even lower frequencies. In the decades to come there will be observations of gravitational waves over a broad spectrum of frequencies, thereby giving information on stellar-mass black holes, intermediate-mass black holes, and supermassive black holes. Presented here is a description of the plans for gravitational-wave detection in the coming decades, and what can be learned from these observations, especially pertaining to black holes and what they tell us about the Universe.

2.6.1 Introduction

A century after their prediction by Albert Einstein [127–129], gravitational waves, produced from a binary black hole merger, were directly detected for the first time by the two Michelson interferometers that form the LIGO [130]. LIGO, along with the European detector Virgo, are the start of a global network of gravitational-wave detectors that begin the era of gravitational-wave astronomy. Together LIGO and Virgo have announced the detection of 50 gravitational-wave events from their first three observational runs in the advanced detector era (O1, O2, and the first half of O3, O3a), namely 48 binary black hole mergers and two binary neutron star mergers [131]. These second-generation ground-based interferometric gravitational-wave detectors are referred to as Advanced LIGO [132] and Advanced Virgo [133]. With arm lengths of 3 km for Virgo and 4 km for LIGO, these detectors can measure a relative displacement of the interferometers' arm lengths of 10^{-18} m.

Presently numerous collaborations are building and operating second-generation interferometers in order to detect gravitational waves. Advanced LIGO in the

United States consists of two interferometers located in Livingston, Louisiana, and Hanford, Washington. Advanced LIGO started observations in 2015, and will be working over the coming years to achieve its design sensitivity, with the goal to reach it by 2025 [134]. The European Advanced Virgo interferometer is near Pisa, Italy [133]. Virgo started acquiring data in 2017, and will also be aiming for its target sensitivity in the coming years. GEO 600, a German–British collaboration, is a 600 m detector near Hanover, Germany [135], and is currently operational. KAGRA is the Japanese 129 km interferometer that is presently under commissioning, and will commence observations in 2022 [136]. There will be a third 4 km LIGO interferometer, LIGO-India [137], located in India, with the goal to be operational in the coming years. All of the km-length detectors will be attempting to detect gravitational waves with frequencies from 20 Hz up to a few kHz. The higher the mass of a binary system, the lower the frequency of the gravitational-wave signal. The second-generation gravitational-wave detectors will observe lighter binaries, like binary neutron star systems with individual component masses around 1.4 $M_{\odot}$ ($M_{\odot}$ represents the mass of our Sun), up to binary black hole systems with total masses up to hundreds of $M_{\odot}$.

In Europe there are plans to build a third-generation gravitational-wave detector, the Einstein Telescope [138], while in the United States there are plans for the Cosmic Explorer [139]. These detectors would have a better sensitivity over the present second-generation detectors by a factor of 10, thereby seeing 1000 more of the Universe. The low-frequency sensitivity will be pushed down to 5 Hz, and maybe even 1 Hz, which will allow for the observation of more massive binary black hole systems. Intense research and development is presently ongoing so as to allow for these detectors to observe in the mid-2030s. The Einstein Telescope will be in a triangular configuration, 10 km arm-length, with three low-frequency detectors and three high-frequency detectors. In order to minimize seismic and anthropogenic noise, the Einstein Telescope is planned to be 100–300 m below ground. The Cosmic Explorer will have an L-shape (like the present LIGO and Virgo), but gain in sensitivity with a possible arm-length of 40 km; this detector will be on the surface of the Earth. The third-generation detectors will measure binary black hole systems up to a few thousand $M_{\odot}$.

The LISA will be a gravitational-wave detector in space [140]. It will consist of three satellites, forming an equilateral triangle, with arm lengths of 2.5 million km. Laser beams propagating between each satellite create three gravitational-wave detectors. LISA will orbit the Sun, but 20° behind the Earth's orbit. LISA will observe gravitational waves in the 10^{-5} to 1 Hz band. LISA will measure binary black hole systems up to a few million $M_{\odot}$.

Pulsars, or spinning neutron stars, emit radio signal in a very stable fashion, like clock. The received timing of the pulses can be perturbed due to gravitational waves. As such, Pulsar Timing Arrays are being used to search for gravitational waves at very low frequencies, 10^{-9}–10^{-6} Hz. Binary systems containing supermasssive black holes would emit in this frequency band. The NANOGrav collaboration could already be observing a first hint of a gravitational-wave signal [141]. There is a growing effort around the world to monitor pulsars for pulsar timing gravitational-wave detection. In addition to existing collaborations in the USA, Europe, and Australia, the upcoming radio telescope network, the Square Kilometer Array, will also contribute to pulsar timing observations in the decades to come [142–144]. Pulsar timing will measure binary black hole systems up to a few billion $M_\odot$.

2.6.2 Gravitational waves

An accelerating electric charge produces electromagnetic radiation—light. It should come as no surprise that an accelerating mass produces gravitational light, namely gravitational radiation (or gravitational waves). In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational waves because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational waves in 1916, and Joe Taylor and Joel Weisberg [145] indirectly confirmed their existence through observations of the orbital decay of the binary pulsar 1913+16 system. The direct detection of gravitational waves has been difficult, and has literally taken decades of tedious experimental work to accomplish. The only possibility for producing detectable gravitational waves comes from extremely massive objects accelerating up to relativistic velocities. The gravitational waves that have been detected so far have come from the coalescence of binary neutron star or binary black hole systems. For example, the first direct detection of a gravitational-wave event, GW150914 [130], was produced by the merger of a 29 $M_\odot$ black hole and a 36 $M_\odot$ black hole some 1.3×10^9 light years away. The total energy radiated in gravitational waves was equivalent to 3 $M_\odot c^2$, with a peak luminosity of 3.6×10^{56} ergs s^{-1}, or about 10^{23} times the luminosity of the Sun.

Other possibly detectable gravitational-wave sources are also astrophysical: core-collapse supernovae, pulsars, neutron star—black hole binary systems, newly formed black holes, or even early Universe inflation. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravitational waves from the Big Bang would provide unique information of the Universe at its earliest moments. Observations of core-collapse supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies up to hundreds of Hertz, and the signals from these objects will help to decipher their characteristics. Gravitational waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter; they would also help to explain the mechanism that produces short gamma-ray bursts. The observation of black hole formation from these binary systems, and the ringdown of the newly formed black hole as it approaches a perfectly spherical shape, would be the *coup de grâce* for the debate on black hole existence, and the ultimate triumph for general relativity.

Advanced LIGO [132] and Advanced Virgo [133] are second-generation interferometric gravitational-wave detectors. Initial LIGO and Virgo conducted observations from 2002 through 2010. Advanced LIGO and Advanced Virgo will ultimately have better sensitivities, by a factor of 10, over their initial designs. They search for gravitational waves from 20 Hz up to a few kHz. Their target sensitivities will allow them to observe signals from the coalescence of binary neutron star systems (1.4 $M_\odot$–1.4 $M_\odot$) out to distances past 300 Mpc for Advanced LIGO and past 200 Mpc for Advanced Virgo. The mergers of more massive binary black holes systems will extend much farther. Already from their first three observational runs Advanced LIGO and Advanced LIGO have detected gravitational waves from 48 binary black hole coalescences, and two binary neutron star inspirals [131].

2.6.2.1 Some general relativity

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational waves will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational waves will deform the fabric of space time. Just like electromagnetic radiation there are two polarizations for gravitational waves. Let us imagine a linearly polarized gravitational wave propagating in the z-direction, $h(z,\ t) = h_{0+}e^{i(kz-\omega t)}$. The fabric of space is stretched due to the strain created by the gravitational wave. Consider a length L_0 of space along the x-axis. In the presence of the gravitational wave the length oscillates like

$$L(t) = L_0 + \frac{h_{0+}L_0}{2}\cos(\omega t)$$

hence there is a change in its length of

$$\Delta L_x = \frac{h_{0+}L_0}{2}\cos(\omega t).$$

A similar length L_0 of the y-axis oscillates, like

$$\Delta L_y = -\frac{h_{0+}L_0}{2}\cos(\omega t).$$

One axis stretches while the perpendicular one contracts, and then vice versa, as the wave propagates through. Consider the relative change of the lengths of the two axes (at $t = 0$),

$$\Delta L = \Delta L_x - \Delta L_y = h_{0+}L_0,$$

or

$$h_{0+} = \frac{\Delta L}{L_0}.$$

So the amplitude of a gravitational wave is the amount of strain that it produces on space time. The other gravitational wave polarization ($h_{0\times}$) produces a strain on axes 45° from $(\boldsymbol{x},\boldsymbol{y})$. Imagine some astrophysical event produces a gravitational wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L_0. The first gravitational wave observed by LIGO had an amplitude of $h \sim 10^{-21}$ with a frequency at peak gravitational-wave strain of 150 Hz [130]. The magnitude of a gravitational wave falls off as $1/r$, so it will be impossible to observe events that are too far away. However, when the detectors' sensitivity is improved by a factor of n, the rate of signals should grow as n^3 (the increase of the observable volume of the Universe). This is because the gravitational-wave detectors measure signals from all directions; they cannot be pointed, but reside in a fixed position on the surface of the Earth.

A Michelson interferometer can measure small phase differences between the light in the two arms. Therefore, this type of interferometer can turn the length variations of the arms produced by a gravitational wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational-wave detectors have evolved. Imagine a gravitational wave of amplitude h is incident on an interferometer. The change in the arm length will be $\Delta L \sim h\,L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. The Advanced LIGO and Advanced Virgo detectors will measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, much smaller than an atomic nucleus [146]. The recent observation of gravitational waves has been one of the most spectacular accomplishments in experimental physics, and has been greeted with much excitement across the globe.

The optical systems for a laser interferometric gravitatational-wave detector are quite complex. Figure 2.31 displays the optical set-up for the Advanced LIGO detector. Every photon that enters the detector can be thought of as a *meter stick* that is used to measure the length difference between the two arms. The statitics of repeated measurements then implies that the more photons used, the more measurements that are made, and ultimately the better the detection statistic. Hence, high-power lasers are used. Also, various schemes are employed to *recycle* the light to build up power and the signal. Because these interferometers are so large, 4 km for LIGO and 3 km for Virgo, the required lasers beams are large in size. Figures 2.32 and 2.33 show photographs of actual optical components. Even though they are large in size, they still must be prestine in their optical qualities. This is a major technological challenge for gravitational-wave detectors.

Figure 2.34, top, presents an aerial view of the LIGO site at Hanford, Washington State in the USA. The magnitude of the 4 km system is apparent. Figure 2.34, bottom, displays the Virgo detector with its 3 km, located near Pisa, Italy.

2.6.3 Black holes

Black holes are predicted by general relativity. They are so dense that nothing can escape from them, including light. Far from being a purely theoretical concept, their presence has been observed in our Galaxy and Universe via numerous means. For example, the presence of black holes can be inferred from x-ray observations [150]. Advanced LIGO [132] and Advanced Virgo [133] have now directly observed gravitational waves from stellar-mass binary-black-hole systems [151]. With the observation of GW190521, the birth of a $142 M_{\odot}$ intermediate mass black hole has

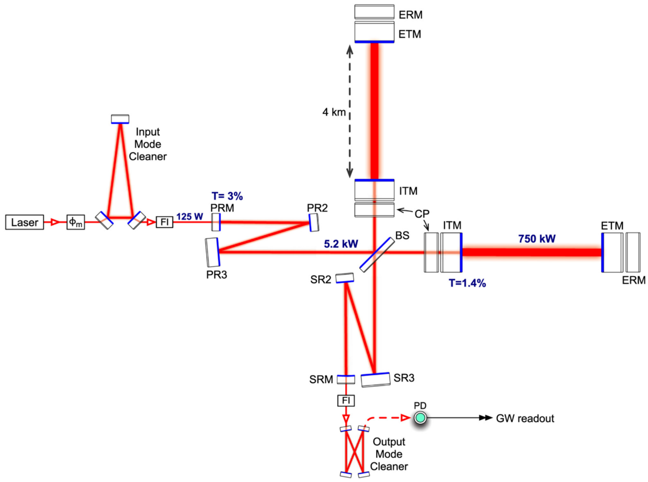


Figure 2.31. The advanced LIGO optical system. The laser (~200 W) light propagates from the stabilized laser through a phase modulator (ϕ_m) to the input mode cleaner, then through a Faraday Isolator (FI) to the power recycling mirror (PRM). The folding mirrors, PR2 and PR3, direct the light to the beamsplittler (BS) input of the interferometer. Note that approximately 125 W of light impinges upon the power recycling mirror, resulting in 5.2 kW at the input port to the beam splitter. The Fabry–Perot cavities are formed with the input test mass (ITM), which is coupled to a compensation plate (CP), and the end test mass (ETM) which is coupled to a end reaction mass (ERM). The Fabry–Perot cavities will contain 750 kW of light power. Note too that the output signal from the interferometer can itself be recycled and amplified at specific frequencies, dependent on the reflectivity of the signal recycling mirror (SRM); SR2 and SR3 are folding mirrors. The output beam also has its spatial features cleaned with the output mode cleaner before the light falls upon a photodetector (PD). Reproduced from [132], copyright IOP Publishing Ltd. All rights reserved.

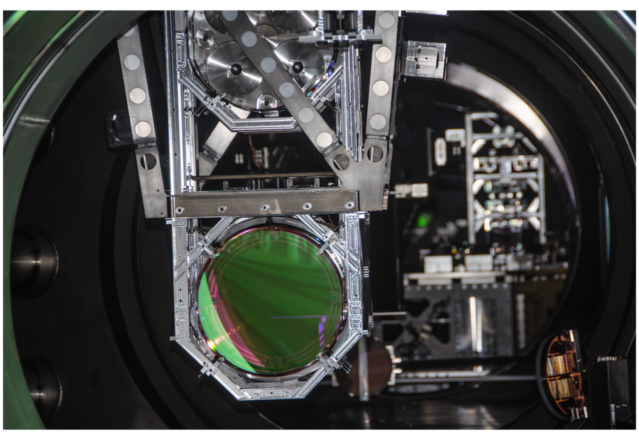


Figure 2.32. A picture of the input test mass (mirror R_1) for Advanced LIGO within its vibration isolation suspension system. The fused silica component is 40 kg, 34 cm in diameter and 20 cm thick. Photograph courtesy of LIGO/Caltech/MIT.

Figure 2.33. Advanced Virgo optics. Left: The power recycling substrate (35 cm in diameter, left) beside the beam splitter of 55 cm in diameter. Right: One of the Fabry–Perot cavity mirrors, which also serves as a test mass for the gravitational wave detector. Reproduced from [133], copyright IOP Publishing Ltd. All rights reserved.

Figure 2.34. Top: Aerial view of the LIGO Hanford, Washington site. The vacuum enclosure at Hanford contains the 4 km interferometer. Photograph courtesy of LIGO/Caltech/MIT. Bottom: Aerial view of the Virgo detector, with 3 km arms, located near Pisa, Italy. Photograph courtesy of the European Gravitational Observatory.

been observed [152, 153]. Observations of the orbital dynamics of stars in the center of the Milky Way indicate that there is a supermassive black hole of mass $4 \times 10^6\ M_\odot$ [154, 155]. The Event Horizon Telescope has recently imaged the shadow of a $6.5 \times 10^9\ M_\odot$ supermassive black hole in the center of the galaxy M87 [156]; see figure 2.36.

While black holes are a consequence of general relativity, they were not initially recognized as we consider them today. Soon after Einstein published his description of general relativity in 1915 [127], Schwarzschild quickly followed up with the solution for the metric about a spherical body [157]. The concept of a black hole was not immediately apparent from the solution, and it took many decades for the modern interpretation to emerge [158]. Other interesting features of black holes were subsequently deciphered. A proposition known as *cosmic censorship* states that there should be no naked singularities; specifically, we should not be able to view from afar the point where the curvature of space time for the black hole goes to infinity. As such, the magnitude of the dimensionless spin vector, $\overrightarrow{\chi} = \overrightarrow{S} c/(GM^2)$, where M is the mass of the black hole and $\overrightarrow{S}$ is its spin angular momentum, must not exceed 1 [159, 160]. The *no-hair theorem* says that black holes are described by only three parameters; their mass, spin angular momentum, and charge [161, 162]. Finally, black holes obey laws of thermodynamics, and radiate as a black-body with temperature $T = \frac{\hbar c^3}{8\pi k_B GM}$, where $\hbar$ is the reduced Planck's constant, c is the speed of light, k_B is Boltzmann's constant, G is Newton's constant, and M is the black hole mass [163, 164].

2.6.4 Gravitational-wave detections and what they tell us about black holes

On September 14, 2015, at 09:50:45 UTC a gravitational wave was detected directly for the first time. The gravitational wave was first observed at the LIGO Livingston Observatory (Louisiana), and then 7 ms later at the LIGO Hanford Observations (Washington). An online transient search algorithm identified the signal in 3 min. An offline examination of the data using a template-based search for compact binary coalescence signals identified the gravitational wave with a signal-to-noise ration of 24 [130]. Parameter estimation routines were used to determine that the gravitational-wave signal was emitted from the merger of two black holes with masses of 36 $M_\odot$ and 29 $M_\odot$. The newly created black hole had a mass of 62 $M_\odot$, meaning that the total energy of gravitational wave emitted was equivalent to 3 $M_\odot c^2$. The system was 1.3 billions light years away from us when it merged.

The measured gravitational-wave signal, GW150914, from the two LIGO detectors is displayed in figure 2.35 [130]. The peak amplitude of GW150914 is $h \sim 10^{-21}$ which corresponds to a displacement of the interferometers' arms of $\Delta L \sim 2 \times 10^{-18}$ m. The exquisite sensitivity of these interferometers can be seen from these numbers. In addition to GW150914, during Advanced LIGO's first observing run two other gravitational-wave events were observed (also stellar-mass binary black hole mergers) [165, 166].

In their first three observing runs, O1, O2, and O3 (first half O3a, with results from the second half O3b still to be announced), Advanced LIGO, and Advanced Virgo reported the observations of 50 gravitational-wave events; 48 from binary black hole mergers and two from binary neutron star mergers [131]. The first reported three-detector (the two LIGO detectors plus Virgo) observation of gravitational waves was GW170814, a binary black hole coalescence [167]. The binary

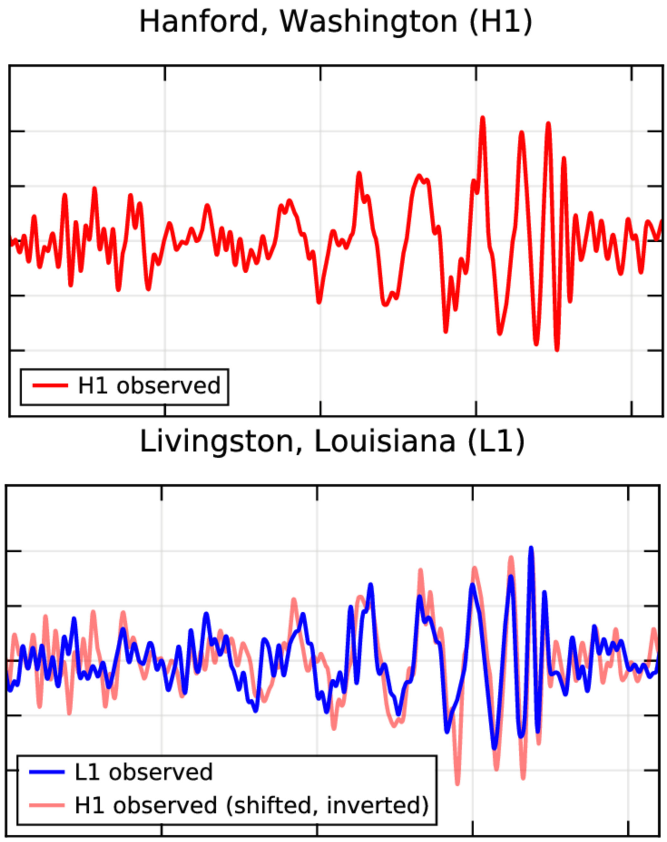


Figure 2.35. The measured gravitational-wave signal GW150914 as observed at the two LIGO interferometric detectors. The data has been bandpass filtered (35–350 Hz), and the gravitational-wave signal is clearly observable by eye. Top: The signal as observed from the LIGO Hanford detector. Bottom: The signal as observed from the LIGO Livingston detector (blue). In addition, the Hanford signal (red) is superimposed after it has been displaced by 7 ms and inverted (due to the relative orientation of the two detectors). The similarity of the two measured signals is clearly visible. Figures [130] courtesy of the LIGO Open Science Center (losc.ligo.org).

neutron star event, GW170817 [168], was also observed by the three LIGO and Virgo detectors, and as such, the position of the event could be estimated to a spot size of 28 deg^2 on the sky and a distance of 40^{+8}_{-14} Mpc. There was also a gamma-ray signal observed 1.7 s after the merger time of the binary neutron star system [169, 170]. From the gravitational-wave and gamma-ray position estimates it was possible to find the location of the source of these signals [171]. GW170817 signaled the birth of gravitational-wave multimessenger astronomy.

2.6.4.1 Binary black holes

With so many observed binary black hole mergers it is possible to study the statistical distributions of the physical parameters of the black holes. This includes the masses, the spins, and their distance. These parameters are estimated via Bayesian methods [172, 173]. The frequency with which the observations are made, and the distances to the sources allow for an estimation of the merger rates. For example, the LIGO-Virgo observations to date predict a binary black hole merger rate of 23.9 $\text{Gpc}^{-3}\ \text{yr}^{-1}$, where a Gpc is 3.3 billion light years. This means that the binary black hole mergers that LIGO-Virgo are observing (stellar mass) are happening about once every 4min in the observable Universe. For comparison, the rate of binary neutron star mergers is estimated to be 320 $\text{Gpc}^{-3}\ \text{yr}^{-1}$. Further observations will help to reveal how the merger rate of binary black holes varies with redshift (or their occurrence as a function of the age of the Universe). The present LIGO-Virgo observations indicate the the binary black hole merger rate does increase as one goes back in time, but not as fast as the rate at which stars form looking back in the history of the Universe. Knowing better the binary black hole merger rate over the history of the Universe will provide important information as to how these systems were formed, and even information on the evolution of massive systems, such as galaxies [174].

2.6.4.1.1 Formation in the field

The observation of the spins of the initial component black holes, and the resulting statistical distribution, can provide information on the possible formation mechanism. The field formation consists of two initial massive stars in a binary system. One star's life ends, and it forms a black hole. Material is then pulled off of the second star, into the black hole. A common envelope of material then forms about the pair, and a close orbit is produced. Eventually the core of the second star collapses to a black hole [175, 176]. The binary black hole's orbit decays via subsequent gravitational-wave emission, and ultimately there is the merger into a new and larger black hole.

For the binary black holes that have formed by field formation the spins of the initial black holes are more likely to be aligned with the orbital angular momentum. This scenario is consistent with many, but not all, of the LIGO-Virgo observed binary black hole mergers.

However, the reality of stellar physics can make this formation scenario difficult for more massive black holes. The formation by stellar processes of black holes in the $\sim 64 - 135 M_\odot$ range is prohibited by a process known as the (pulsational) pair-instability supernova [177, 178]. For these masses, the star's core becomes unstable, and the star disintegrates, leaving no remnant behind. The recent detection by LIGO-Virgo of GW190521 had the initial black hole masses at $85 M_\odot$ and $66 M_\odot$ [152, 153]. These masses are in tension with stellar formation, especially for the $85 M_\odot$ black hole. Multiple formation processes are probably necessary to describe all of the events that LIGO-Virgo have observed.

It is also important to consider the lower mass limit for black holes, especially for those formed via stellar processes. LIGO and Virgo have observed a unique system GW190814 Abbott:2020khf, with initial component masses of $23.2 M_\odot$ and $2.59 M_\odot$. This small secondary mass is difficult to explain, and an important question is whether this is a neutron star or a black hole. Neutron stars with masses above $2 M_\odot$

are also difficult to construe. Possibly this small mass seen in the GW190814 signal could be a black hole formed by the merger of two neutron stars, and would be at the lower limit for black hole masses formed by stellar processes.

2.6.4.1.2 Dynamic formation

Dense regions, like the center of galaxies or in globular clusters, could be environments where black holes have multiple interactions with one another, mergers happen, and more massive black holes are formed. This is the other major formation scenario to explain the binary black holes that LIGO-Virgo are observing. With such dynamic formations the spins of the initial black holes can be expected to be randomly distributed, independent from one another, and independent of the orbital angular momentum [179, 180]. LIGO and Virgo see evidence for some binary black hole systems where the initial component masses have spins parallel to the orbital plane (or perpendicular to the orbital angular momentum).

When the spins are perpendicular to the orbital angular momentum, there is spin–orbit coupling (similar to atomic systems), and this will cause a precession of the orbital plane. This was the case with the very massive GW190521 system, where there were indications of such a precession [152, 153]. The statistical studies of all the LIGO-Virgo binary black hole observations to date show the presence of orbital precession. This, plus the character of the mass distribution at high masses, supports the assumption that some of the observed binary black hole systems have been formed by dynamical processes [174].

2.6.4.1.3 Primordial black holes

The observations of numerous binary black hole mergers has generated much discussion as to whether black holes might be an explanation to the missing mass, or dark matter, problem. About 5% of the mass–energy of the Universe is baryonic matter (protons and neutrons are baryons), namely the material that makes up the elements (H, He, C, O, N, etc) that we are familiar with. However, there is much evidence that there is missing mass, and missing energy. It is estimated that 27% of the mass energy in the Universe is dark (unobserved) and non-baryonic mass.

The numerous binary black hole mergers observed by LIGO and Virgo has renewed interest in the possibility that black holes could be formed in the early Universe, namely primordial black holes [181, 182], and these could contribute to the missing mass [183, 184]. It has been theorized that cosmic strings could also be responsible for primoridal black hole formation [185]. The possibility that black holes could be dark matter has been constrained by microlensing observations, and the structure of the cosmic microwave background. However, a window in allowable masses exists, from $30M_{\odot}$ to $100M_{\odot}$, consistent with the LIGO-Virgo observations. Some studies claim that the black holes observed by LIGO-Virgo could account for all of the dark matter [186]. For the LIGO-Virgo observations, primordial black holes in binary systems would have their spins randomly distributed, independent of the orbital angular momentum of the binary system [187].

If black holes were formed in the early Universe it is expected that they would be created with a distribution of masses. LIGO and Virgo are conducting searches especially targeting compact binary systems where the component masses are under $1 M_{\odot}$ [188]. If such systems were to be found, it would be difficult to explain the presence of a compact object less than $1 M_{\odot}$. This would be too small to be a neutron star, so presumably it would be a black hole. In such a case the best assumption would be that the black hole was formed in the early Universe, namely a primordial black hole. Continuing searches for such sub-solar mass sources by LIGO-Virgo, and future third-generation ground-based detectors, will be critical in determining the existence of primordial black holes.

The work of Jacob Bekenstein showed that black holes should have entropy [189], which was followed by Stephen Hawking who showed that such an object with entropy would have a temperature, and would therefore radiate like a black body [164]. According to the derivation by Hawking, a black hole formed in the early Universe would have evaporated by now (universe age of about 14 billion years [189]) if the initial mass was 1.7×10^{11} kg or less. Such a decay would release many photons and neutrinos, especially at the final moments; such an explosion has not been observed. The cosmic microwave background has a temperature of 2.7°, and if one considers a thermal equilibrium between Hawking radiation and the acquisition of energy from the cosmic microwave background, then the mass limit for primordial black holes would be approximately 5×10^{22} kg, very similar to the mass of the moon.

2.6.5 The path to supermassive black holes

The centers of most galaxies seem to have a supermassive black hole, of millions of solar mass, or even much larger. Our Milky Way Galaxy contains a black hole of $4 \times 10^{6} M_{\odot}$ [154, 155]. The Event Horizon Telescope has recently imaged the shadow of a $6.5 \times 10^{9} M_{\odot}$ supermassive black hole in the center of the galaxy M87; see figure 2.36 [156]. An important question is how these supermassive black holes were formed, and what role they play in the formation of galaxies. One assumption is that there were very massive stars in the early Universe that quickly became black holes with masses of tens to hundreds of solar masses. These black holes then consumed the material in its environment (gas, stars, other black holes) and grew [190]. Another scenario has dense regions of gas and dark matter collapsing directly to a black hole of hundreds of thousands of solar masses, without even forming stars. These would then be the seeds for the supermassive black holes in the center of galaxies [191].

However, it is difficult to explain very large black holes that exist in a short time after the Big Bang. For example, the quasar ULAS J1120+0641 was estimated to have a central black hole mass of $8 \times 10^{8} M_{\odot}$ at just 690 million years after the Big Bang [192]. Another quasar, J0313-1806, is estimated to have a black hole with mass $1.6 \times 10^{9} M_{\odot}$ at just 670 million years after the Big Bang [193]. Such large masses early in the Universe are difficult to explain. Gravitational-wave observations of supermassive binary black hole systems with pulsar timing and LISA will hopefully provide more information about these systems.

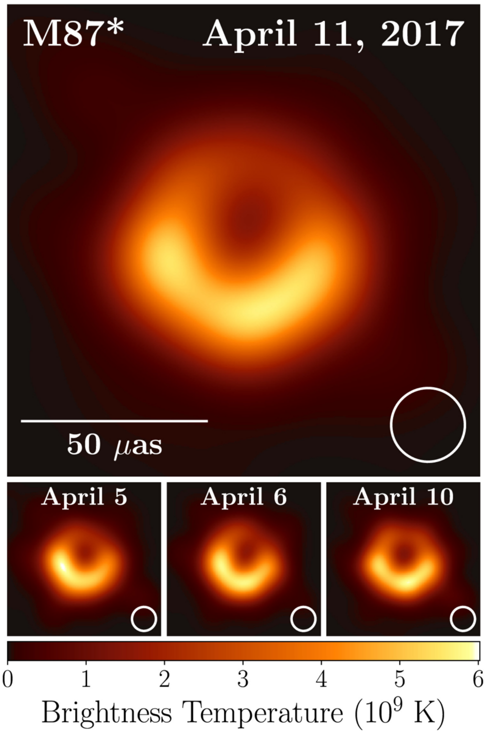


Figure 2.36. The Event Horizon Telescope image of the black hole in the center of the galaxy M87. Radio telescope observations were combined to display the shadow of the supermassive black hole at the center of the Galaxy. Reroduced from [132],

2.6.5.1 Intermediate mass black holes

There is much observational evidence for stellar-mass black holes and supermassive black holes. However, there are relatively few observations of intermediate-mass black holes, $100M_{\odot}$ to $100\,000M_{\odot}$, the bridge from stellar mass to supermassive black holes. It will be important for our understanding of the formation of supermassive black holes to accurately measure the presence of intermediate mass black holes in the Universe. Gravitational-wave observations are likely to be the best way to do this.

The direct observation by LIGO-Virgo of the birth of a $142M_{\odot}$ black hole displayed how intermediate mass black holes can be observed and measured [152, 153]. The other observations of intermediate mass black holes are sparse, and indirect. For example, evidence comes from the dynamics of the central regions of galaxies and star clusters [194]. LIGO and Virgo gravitational-wave observations should provide more observations up to a few hundred solar masses, third-generation detectors like the Einstein Telescope and Cosmic Explorer could observe systems up to a few thousands of solar mass, while the space-based LISA could extend the observations through hundreds of thousands to a few millions of solar mass. Gravitational-wave observations truly give the best means to measure the distribution of intermediate mass black holes, and build the bridge between stellar mass and supermassive black holes.

2.6.6 Testing general relativity

The observations of gravitational waves offer an important means to test general relativity, especially from regimes where gravity is extremely strong. Important tests have already been made with the LIGO–Virgo observations for binary black holes and binary neutron stars. That general relativity is correct is the basic assumption in these tests. The data can be analyzed to see if the total signal is consistent with general relativity. Comparisons can also be made to alternative models of gravity. At some level it is expected that quantum mechanics must play a part in the correct description of gravity, and gravitational-wave observations are being used to look for these quantum effects.

2.6.6.1 Signal residual test

One test that has been done for binary black hole merger observations is the signal residual test. Bayesian parameter estimation routines [172, 173] use general relativity to model the signal. The estimate of the signal is then subtracted from the data, and an examination is done to see if the residual is consistent with LIGO-Virgo detector noise [195]. None of the LIGO-Virgo observations have failed this test. See figure 2.37 for an example of this test for binary black hole merger GW170104 [196].

2.6.6.2 Gravitational-wave polarization

Gravitational waves have two polarizations. This is similar to light, which also has two polarizations. The two LIGO detectors are essentially aligned with respect to each other, although separated by 3000 km. As such, the two detectors tend to respond to the same polarization. The first three-detector observations were made when the European detector, Virgo, joined the network. Due to its large distance displacement from the American detectors, and its orientation, it is able to measure polarization content from gravitational waves that complements that observed by the LIGO detectors. This was first demonstrated with the three-detector observations of binary black hole-produced signals GW170814 [167] and GW170818 [151]. General relativity predicts a type of polarization, tensor, that is different from electromagnetic polarization, vector. Alternative theories of gravity predict vector and scalar polarizations, in addition to the tensor polarization. The polarization content of the binary black hole-produced gravitational-wave signals observed by the LIGO-Virgo three-detector network are consistent with general relativity. This type of polarization test will become more stringent once the worldwide gravitational-wave network expands to include KAGRA [197] in Japan and LIGO-India [137].

2.6.6.3 Speed of gravity

It is predicted by general relativity that gravitational waves travel at the speed of light. In addition, it is thought that short gamma-ray bursts are created by the

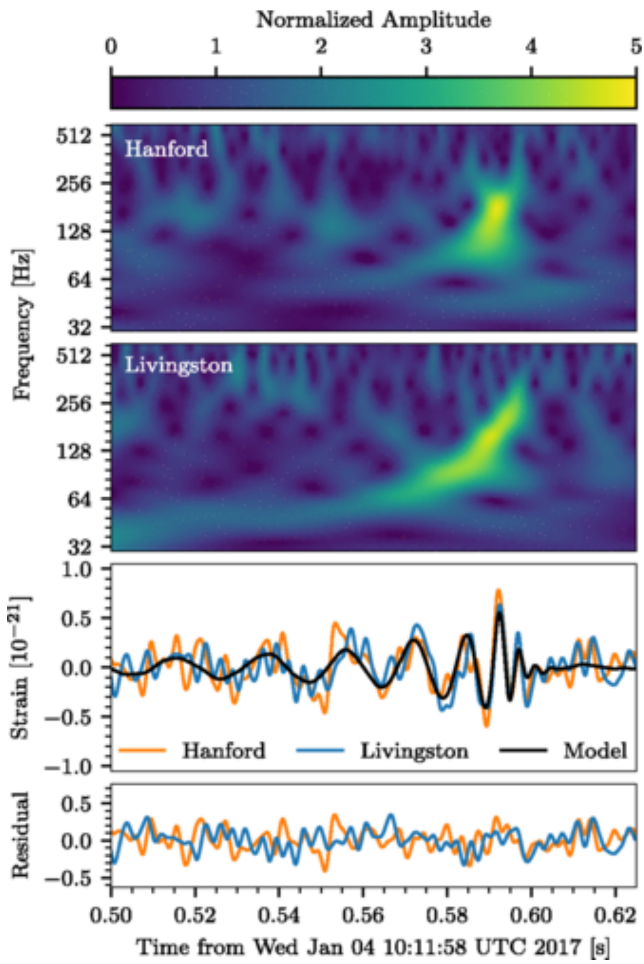


Figure 2.37. Gravitational-wave signal GW170104 observed by the two LIGO detectors. The time-frequency representations of the signal as observed in the LIGO Hanford (top panel) and the LIGO Livingston (second panel) detectors. The third panel shows the data time series from Hanford (orange) and Livingston (blue), along with the waveform best describing the signal (black). The bottom panel is the result when the waveform model is subtracted from the two data streams, leaving Gaussian noise and no indication of a remaining signal. See [196] for more details. Reproduced from [196] CC BY 4.0.

coalescence of binary neutron stars. These predictions were verified when gravitational waves from a binary neutron star merger, GW170817 [198], were seen 1.74 s before the gamma-ray burst, GRB 170817A [199]. This was detected by two satellite gamma-ray detectors, the Anti-Coincidence Shield for the Spectrometer for the International Gamma-Ray Astrophysics Laboratory (INTEGRAL) [200], and the Fermi Gamma-ray Burst Monitor (Fermi/GBM) [201]. The gravitational-wave signal from LIGO and the gamma-ray signals from INTEGRAL and Fermi/GBM are shown in figure 2.38. The difference between the gravitational-wave merger time and the gamma-ray arrival time of 1.7 s can be seen.

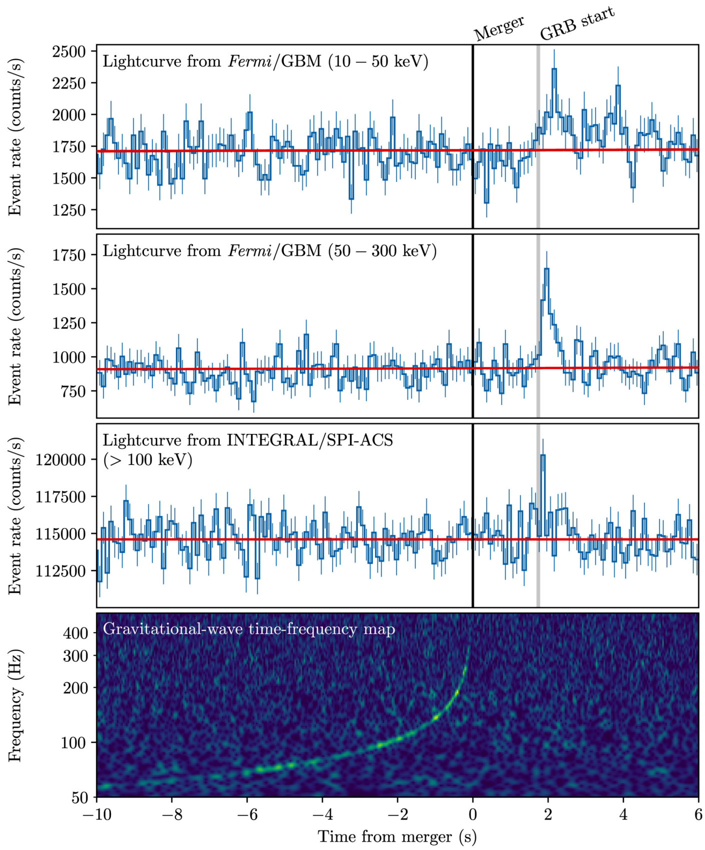


Figure 2.38. The LIGO observed gravitational-wave signal GW170817 and the Fermi/GBM [201] and INTEGRAL [200] observations of gamma-ray burst GRB 170817A. See [199] for more details. Reproduced from [199] CC BY 4.0.

From the gravitational-wave signal alone it is possible to estimate the distance (technically the luminosity distance) to the location of the binary merger. For GW170817, the estimate is 40^{+8}_{-14} Mpc [198], or about 130 million light years. The subsequent searches for an optical counterpart to this event succeeded, with the kilonova found in the galaxy NGC 4993. This observation also showed that the formation of the heaviest elements truly does come from neutron star mergers.

The difference between the arrival times of the gamma rays and the gravitational waves, along with the measurement of the distance, gives the information needed to calculate the difference in the speeds of propagation.

For making the comparison between the speed of gravity v_g, and the speed of light c, LIGO-Virgo used the most conservative distance estimate, namely the lower limit of $D = 26$ Mpc. One does not know, however, the time of the emission for the gamma rays relative to the time of the merger of the neutron stars. While it could very likely correspond to the 1.74 s observation, one can assume that the gravitational waves and gamma rays were emitted simultaneously. As such, the speed of gravity would be greater than the speed of light, and $\Delta v = c - v_g$, $\Delta v/c \approx -3 \times 10^{-15}$. On the other hand, a conservative assumption would be that the gamma emission occurred 10 s after the merger. For this scenario $\Delta v/c \approx 7 \times 10^{-16}$. The gravitational-wave and gamma-ray data set a limit for the difference between v_g and c of [199]

$$-3 \times 10^{-15} \leqslant \frac{\Delta v}{c} \leqslant +7 \times 10^{-16}. \tag{2.35}$$

2.6.6.4 Binary black hole inspiral—merger—ringdown

The gravitational-wave signal from a binary black hole merger has different regimes that correspond to the different physical effects from the source. The assumption is that general relativity provides the description of all these processes. When the two black holes are orbiting one another, energy is lost by gravitational-wave emission, and the orbit decays. There is an increase in the orbital- and gravitational-wave frequencies, and the amplitude of the gravitational wave grows. This is the so-called inspiral, or *chirp* signal. Eventually the system arrives at the innermost stable circular orbit; general relativity does not allow for stable circular orbits for objects too close to each other. At this point the masses fall toward one another, the merger part of the signal. Finally the two black holes meet. By the no-hair theorem, the final black hole can only described by its mass, spin, and electric charge. When the two black holes meet it will resemble something like a peanut shell, but in the end the final remnant black hole must approach a stable configuration. It does this by oscillating, ringing like a bell, losing energy by gravitational-wave emission.

LIGO and Virgo have tested general relativity by comparing the physical parameters estimated from the inspiral part of the signal, with those from the merger-ringdown part. This test has been done with a number of binary black hole merger signals. To date the parameter estimation for the different parts of the signal give consistent results, and hence no deviations from general relativity were observed [202].

2.6.6.5 Echoes and quantum gravity

There has been much research and discussion pertaining to applying quantum mechanics to gravity. For example, it has been theorized that the area of a black hole event horizon is quantized [203]. It is further speculated that this could create a scenario where gravitational waves produced during a binary black hole merger could reflect off of the event horizon of the newly formed black hole [203]. This would produce gravitational-wave *echoes*, and multiple signals would be subsequently observed in gravitational-wave detectors [204]. It has been claimed that the gravitational-wave signal GW150914 [205] had observable echoes [206]. This claim has been challenged by others, however [207]. As LIGO and Virgo (and eventually KAGRA) improve on their sensitivities, it will be important to look for such echo signals in the binary black hole-produced gravitational-wave signals.

2.6.7 Hubble constant

As already noted, one can estimate the luminosity distance to a coalescing compact binary gravitational-wave source. If the source can be found and the redshift measured (giving the velocity) it is then possible to estimate the Hubble constant, the measure of the present expansion rate of the Universe [208, 209]. The definition of the Hubble constant, H_0 comes from

$$v = H_0 d, \tag{2.36}$$

where d is the distance to the source and v is the velocity for the source moving away from us because of the Universe expanding. This gravitational-wave approach would complement the existing methods of measuring the Hubble constant using the cosmic microwave background data [210, 213] or supernovae observations [211]. There is presently a tension between the cosmic microwave background and supernova estimates of the Hubble constant, and so using gravitational waves could help to clarify the description of the expansion of the Universe.

For the binary neutron star-produced gravitational-wave signal GW170817 it was possible to identify the source in the Galaxy NGC 4993, and then make a redshift measurement and obtain the velocity.

The measurement of the Hubble constant was $H_0 = 70^{+12}_{-8}$ km s^{-1} Mpc^{-1} [212]. This observation is consistent with the results from the other means used to estimate the Hubble constant. Cosmic microwave background observations give $H_0 = 67.74 \pm 0.46$ km s^{-1} Mpc^{-1} [213], and supernovae measurements have $H_0 = 73.24 \pm 1.74$ km s^{-1} Mpc^{-1} [211]. Note that the uncertainty for the gravitational-wave Hubble constant measurement is quite large compared to the other two methods. However, with order of 50–100 such binary neutron star gravitational-wave observations the error should be reduced to the level of the other methods, and could help to explain the tension between the cosmic microwave background and supernovae estimations.

2.6.8 LIGO, Virgo, and KAGRA observation schedules

The LIGO and Virgo detectors sensitivity improvements are described in a document that is regularly updated [214]. The arrival of the KAGRA gravitational-wave detector to the worldwide network is expected in 2022 [136], and is the next important step for worldwide gravitational-wave observations. KAGRA, located in Japan, will observe with LIGO and Virgo at for the fourth observing run O4 in 2022. Note too that a third LIGO detector, located in India, will join the worldwide gravitational-wave network in 2027 [197]. The LIGO-Virgo-KAGRA observing scenario calls for a succession of data taking periods separated with detector upgrades and commissioning periods to reach the detectors' target sensitivities. Advanced LIGO and Advanced Virgo recently completed their third observational run O3 at the end of March, 2020. lIGO and Virgo release their data to the public after a proprietary period; the data can be downloaded from the Gravitational Wave Open Science Center [215].

LIGO, Virgo, GEO, and KAGRA have created a new type of telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. This has started with the unexpected observation of gravitational waves produced by binary black hole systems with tens of solar masses. Physicists do know that there will be other signals that they can predict. Binary systems containing neutron stars, for example. It is suspected that short gamma-ray bursts come from the coalescence of binary neutron stars, or neutron star—black hole binary systems. The binary neutron star merger producing GW170817 proved this assumption [199]. A core collapse supernova will produce a burst of gravitational waves that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravitational waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort. The exciting initial observations of gravitational waves have been made, but it is just the beginning of a new astronomy. The second-generation gravitational-wave detectors started detection in 2015 and will continue through the 2020s. The third-generation detectors should arrive in the 2030s, alongside the space-based LISA detector. By 2050 we should have a tremendously more complete view of neutron stars, black holes, and our Universe through gravitational-wave observations from these detectors. Plans will certainly be developing for future improvements.

2.6.9 Challenges and opportunities

To date ground-based gravitational-wave detection has already achieved remarkable success. The second-generation detectors, Advanced LIGO [132] and Advanced Virgo [133], have already announced 50 gravitational-wave observations [131]. More observations will soon be announced as the results from the second half of the third observing run, O3b, are presented. The arrival of the Japanese detector KAGRA [136] for the fourth observing run O4 in 2022 will be a tremendous improvement for the worldwide detector network. With the LIGO-India detector [137] commencing observations around 2027, the 2020s looks very bright for continued gravitational-wave observations. The gravitational-wave observations have given us new insight into the physics associated with neutron stars, black holes, short gamma-ray bursts, kilonovae, tests of general relativity, and cosmology.

2.6.9.1 Future gravitational-wave detectors

These second-generation gravitational-wave detectors observe gravitational waves from about 20 Hz up to the kHz regime. The low-frequency cutoff determines the possible observable mass range for binary black hole systems. The second-generation detector are able to observe binary black hole systems with masses up to a few hundred solar masses.

The third-generation detectors will have sensitivities that are about 10 times better than the second generation. This will allow them to observe 1000 times more of the Universe. Importantly, a goal for the third-generation detectors is to have a lower sensitivity of about 5 Hz (although if the technology advances, this might be pushed to 1 Hz). This will allow for observing binary black hole systems with masses of a few thousands of solar masses. This is a mass range for intermediate mass black holes that is relatively unknown. The ability to measure the population of such binary black hole systems will be extremely important for our understanding of the black hole mass spectrum in the Universe.

In Europe the Einstein Telescope [138] is the proposed third-generation gravitational-wave detector. Whereas the second-generation detectors are Michelson interferometers in a 90° L-shape configuration, the Einstein Telescope will have the configuration of an equilateral triangle with 10 km long arms. There will actually be six interferometers, each with arms at 60°. There will be three interferometers optimized for low-frequency observations, from 1 to 250 Hz. The mirrors for the low-frequency interferometers will be cooled to 10–20 K in order to reduce thermal noise. The Einstein Telescope will then have three high-frequency interferometers, operating from 10 Hz to 10 kHz. For the high-frequency interferometers the mirrors will be at room temperature. To reduce seismic and anthropogenic noise, the Einstein Telescope will be 100–300 m underground. In order to achieve its sensitivity goals much technology needs to be further developed. Researchers in Europe, and around the world, are actively conducting research in a number of areas in order to meet the Einstein Telescope requirements. These include high-power and stabilized lasers (>500 W), mirror coatings with low thermal noise but able to tolerate the 3 MW of light stored in the arms of the high-frequency detectors, cryogenic mirrors for the low-frequency detectors, sophisticated seismic isolation systems for the interferometer mirrors, vacuum systems to accommodate the six interferometers with arm lengths of 10 km, and an extensive underground tunnel network to accommodate the detectors. Bringing the Einstein Telescope to its proposed sensitivity will require tremendous scientific and engineering developments, but the prospects are realistic. The scientific payoffs from the observations will be tremendous, which is the motivation for developing the third-generation detectors. The goal is to have the Einstein Telescope operating by the mid-2030s. Figure 2.39 shows a simplified picture of the Einstein Telescope.

The American led third-generation gravitational-wave detector project is known as Cosmic Explorer [139]. One of the ways that Cosmic Explorer hopes to achieve an improvement in sensitivity by a factor of 10 is to make an L-shaped Michelson interferometer with 40 km arms. This is to compared with the 4 km arms for the LIGO detectors. There will be technological advancements over the current LIGO detectors, but the bulk of the sensitivity gain comes from the arm-length increase. Cosmic Explorer will be on the surface of the Earth, but a difficulty concerns the curvature of the Earth over this distance. In addition, a vacuum system with two 40 km arms will be a challenge. A simple artistic representation of Cosmic Explored is given in figure 2.40.

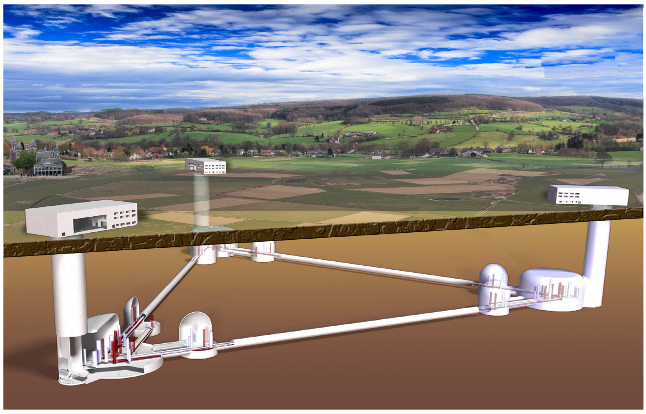


Figure 2.39. The Einstein Telescope will be located in Europe, have six actual interferometric gravitational-wave detectors with arm lengths of 10 km, and located 100–300 m below ground. Figure from the Einstein Telescope collaboration.

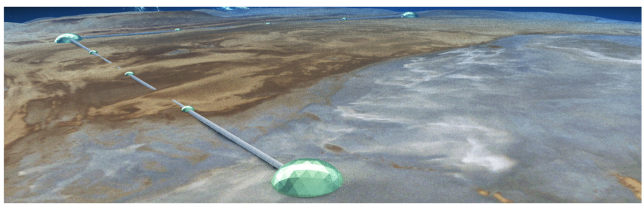


Figure 2.40. The Cosmic Explorer will be located in the United States. It will be an L-shaped interferometric gravitational-wave detector with arm lengths of 40 km, and located the surface of the Earth. Figure from the Cosmic Explorer collaboration.

The third-generation detectors will be expensive, with costs easily exceeding a billion Euros. Tremendous scientific and technical planning will be necessary. The importance of the scientific goals also will need to have universal recognition from the broader physics community. While initial observing is planned for the mid-2030s, the infrastructures created are expected to last for many decades. It is expected that upgrades will be made to specific elements in the detectors, but the vacuum systems and tunnels need to survive for many decades. As LIGO and Virgo have already seen, maintaining such large vacuum systems is not trivial, and problems do arise.

The LISA mission is led by the European Space Agency, but also supported by NASA [140]. The launch is currently planned for 2034. The nominal duration for the mission is 4 years of data acquisition, but this could be extended to 10 years. LISA will be searching for gravitational waves in the 10^{-4}–0.1 Hz band, but it may be possible to extract useful information out to a band of 10^{-5}–1 Hz. The three LISA satellites will form an equilateral triangle with arm lengths of 2.5 million km; see figure 2.41. Each satellite will have two test masses in free fall. The optical system will measure the distance between the masses in the different satellites to an accuracy

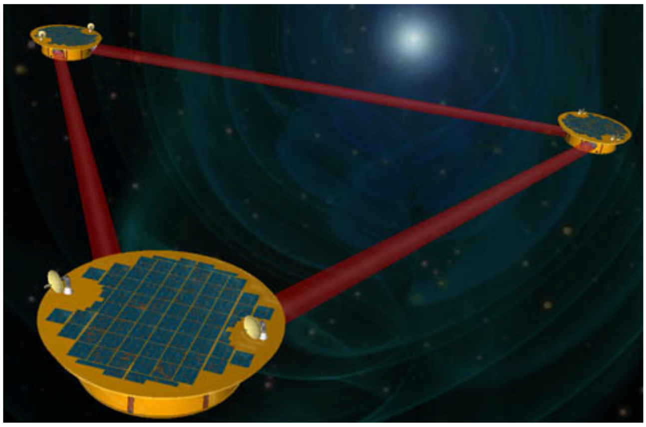


Figure 2.41. A simple figure displaying the three LISA satellites, separated from one another by 2.5 million km. The LISA constellation will orbit the Sun, 20° behind the orbit of the Earth. Figure from NASA.

of pico-meters (10^{-12} m) over the 2.5 million km total distance. There will be of order 1 Watt of laser light exiting the satellite, but only a few hundred pico-Watts of light will be detected after this long distance. The beam will have diverged significantly. As such, it will be impossible to make a true Michelson interferometer, like LIGO or Virgo. LISA will use other optical techniques to measure the small distance displacements. But like LIGO and Virgo, the gravitational wave will make a change to the detector arm lengths, which will induce an optical phase shift in the laser light, which will be detected. Unlike LIGO, Virgo, and the other ground-based detectors, LISA must work on the first attempt. There will be no means for human intervention after the launch. The physics necessary for gravitational-wave detection by LISA are already established; it is a question as to whether such a detector can be properly engineered to work as expected.

Much of the technology for making accurate distance measurements between test masses in space has already been demonstrated by the LISA Pathfinder Mission [216, 217]. This mission was launched on December 3, 2015, conducted experiments for 576 days, and was deactivated on June 30, 2017. The experiment consisted of two test masses in the satellite. There were two cubes (46 mm per side), 1.928 kg, made of a gold–platinum alloy. One mass was in a true free fall, with the satellite fixing its position about the mass. The second mass was electrostatically controlled to remain at a fixed distance separation of 376 mm. The differential acceleration between two freely falling test masses in space was measured to sufficient accuracy to demonstrate that the LISA mission could succeed in its quest to measure gravitational waves in space. This measurement was made in the observational band of LISA. Figure 2.42 displays the LISA Pathfinder and its optical system.

In the coming years gravitational-wave searches will continue to be conducted by pulsar timing. Radio telescope networks around the world will be observing pulsars. The Square Kilometer Array is currently being constructed in South Africa and Australia, with the final network assembly coming online in 2027 or possibly later.

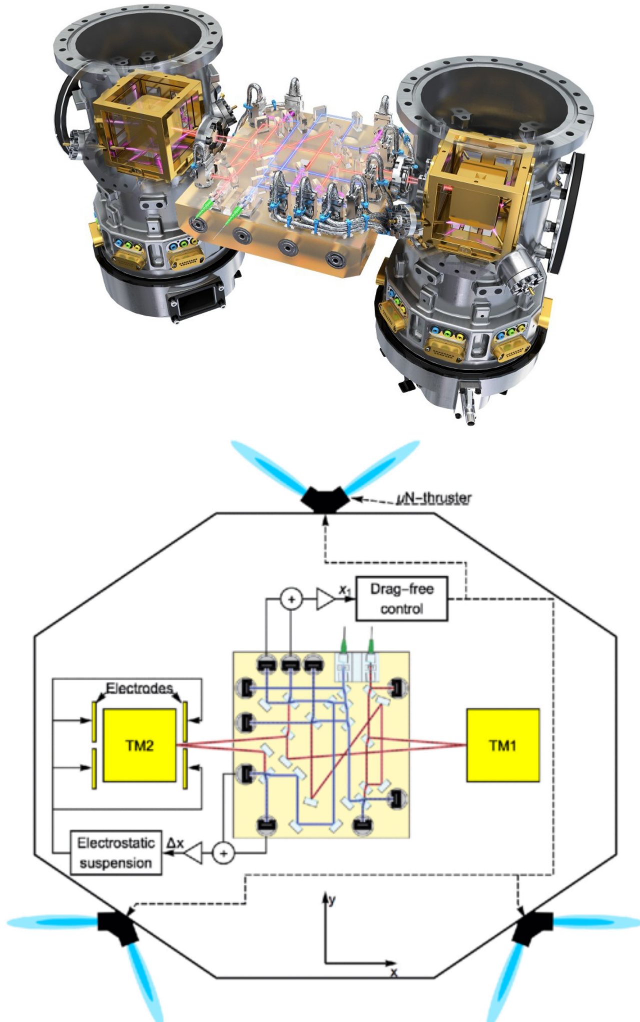


Figure 2.42. Top: Representation of the LISA Pathfinder experiment. Note the two masses and the laser interferometric system to measure the distance between them. Bottom: Schematic of the spacecraft, test masses, and optical system for measuring the distance between the masses. Top figure from the European Space Agency, bottom figure from [216] CC BY 4.0.

The Square Kilometer Array will augment other radio observations and search for gravitational waves in the 10^{-9}–10^{-6} Hz frequency band [142–144].

The 2.7° cosmic microwave background is the remnant of the thermal radiation present when free protons and electrons formed neutral atoms 370 000 years after the Big Bang. Gravitational waves produced in the Big Bang, during inflation, should have left and imprint in the distribution of the polarization of the cosmic microwave background. Numerous observations are being conducted, and will continue to be conducted, to observe this signature of gravitational waves and hopefully gain knowledge of the parameters describing inflation [218].

The coming decades will be seeing extensive observations of gravitational waves over a broad frequency range. This should propel our knowledge of astrophysics and cosmology.

2.6.9.2 Future science results from gravitational-wave observations

The 2030s will see the arrival of the third-generation ground-based gravitational-wave detectors, the Einstein Telescope, and Cosmic Explorer. Similar to LIGO, Virgo, and KAGRA, these detectors will continue to search for gravitational-wave signals from binary neutron star mergers, binary black hole mergers, supernovae, pulsars, a stochastic gravitational wave background, and exotic objects such as cosmic strings. With an improvement in sensitivity by a factor of 10, the ability of third-generation detectors to observe these signals is greatly improved.

Binary neutron stars are relatively light objects with a total mass of around $3M_{\odot}$. Still, the third-generation detectors will be able to detect the gravitational waves from a binary neutron star merger out to a redshift of $z \approx 2$ (when the Universe had an age of 3.3 billion years), and maybe even $z \approx 3$ (when the Universe had an age of 2.2 billion years). For comparison, the age of the Universe is approximately 13.8 billion years [179, 213]. It is estimated that the third-generation detectors could detect 70 000 binary neutron star merger events per year [219]. The large number of observations, including many with large signal-to-noise ratio, should tell us much about the nuclear physics associated with neutron stars, including their equation of state. This will happen when tidal effects change the shape of the neutron stars before merger, thereby changing the gravitational-wave signal.

Some of the observed binary neutron star mergers will probably also be seen in electromagnetic observations, be they short gamma-ray bursts or kilonova identification. Multimessenger astronomy will greatly benefit from the large number of these observations. This depends, of course, on there being sufficient electromagnetic observatories, especially gamma-ray detection on satellites. Coincident observations of high-energy neutrinos will also be very informative. These observations will tell us more about the physics of short gamma-ray bursts, the formation of gamma-ray jets, what part neutrinos play in these mergers, and the formation of heavy elements in the Universe.

Joint gravitational-wave and electromagnetic observations will also help to reduce the uncertainty in the estimation of the Hubble constant. The gravitational-wave based Hubble constant measurement will be as statistically accurate as other methods. This will certainly be important in explaining the tension between the Hubble constant measurement from observations of the cosmic microwave background [213] or supernovae [211]. The accurate determination of the Hubble constant is critical for our ability to describe the Universe.

Being more massive, binary black hole events will be easier for the third-generation detectors to observe, and so they will be seen even further back in the Universe. In fact, almost every binary black hole merger in the observable Universe with masses up to a few thousand solar masses will be detected by the third-generation detectors. This will amount to around a million observations per year. This will provide the means to do detailed statistical studies of the observations. This will include measurements of the merger rate as a function of total mass, component spins, and redshift (distance). This should provide the means to understand the formation channels for black holes, from stellar based, to dynamical black hole mergers in dense environments. Any evidence of primordial black hole formation would be revolutionary for cosmology. The important information about intermediate mass black holes, from 100 solar mass up to thousands of solar mass, will be extremely important for understanding black holes in the Universe, and making the bridge between stellar-mass black holes and supermassive black holes.

The observation of gravitational waves produced during the Big Bang would provide cosmological information that is impossible to obtain by any other means [220]. This cosmologically produced stochastic gravitational-wave background could provide data about inflation, phase transitions in the early Universe, or the presence of exotic objects like cosmic strings. However, the merger of binary neutron stars and binary black holes over the history of the Universe will create another stochastic gravitational-wave background that will obscure the observation of the cosmological background. Techniques have been proposed to separate these different backgrounds using data from third-generation detectors [221, 222].

However, the third-generation detectors will be so sensitive that they will directly detect virtually all binary black mergers in the Universe, and most binary black hole mergers. As such, these events can be removed from the data, and a search for the cosmologically produced stochastic gravitational-wave background can be done without this astrophysically produced foreground. This could lead to an important discovery, or setting stringent limits on the cosmological gravitational-wave background [223, 224].

While the third-generation gravitational-wave detectors will begin observations in the mid 2030s, they will continue observing well past 2050. For example, the proposal for the Einstein Telescope is to have an infrastructure that will last for at least 50 years. We can expect that the technology for gravitational-wave detection will progress, and important scientific observations will happen of the decades.

The LISA Mission will be launched in the mid 2030s, with a plan for 4 years of observations, with a possible extension to 10 years. Within the observation band of 10^{-4}–0.1 Hz LISA will be expected to observe thousands of signals. The science case for LISA is extremely strong.

Starting closest to us, LISA will tell us much about compact binary stars in the Milky Way. This will include observations of binaries comprised of white dwarfs, neutron stars, and stellar-mass black holes. The observation of how the tidal disruption of the objects affects the orbital evolution will be important. Some galactic binary systems will be observable with gravitational waves and electromagnetic observations. The comparison of these observations will be important for understanding tidal interactions and other features than can perturb the orbit.

LISA will be sensitive to binary black merger signals for systems with a total mass up to a few million solar mass. The merger of such systems will be observable out to the earliest moments of the Universe. Specifically, these mergers will be observed out to redshifts of $z \approx 10$ (when the Universe had an age of 500 million years) to $z \approx 20$ (when the Universe had an age of 200 million years). This will be extremely important as it can help us understand the starting process for the the formation of the supermassive black holes that we now observe in the centers of galaxies. Mapping out the black hole masses as a function of redshift,as well as the merger rates as function of redshift, will also be critical for our understanding of structure and its formation in the Universe. The observation of an electromagnetic signal in coincidence with an observed binary black hole merger would also provide information about the environment occupied by these black holes.

LISA should be able to observe gravitational-wave signals from extreme mass ratio inspirals. This would be when a stellar-mass black hole ($\sim 50 M_{\odot}$) orbits around and then plunges into a $10^5 M_{\odot}$ to $10^6 M_{\odot}$ black hole. Such a signal would be given information about the environment about large black holes, and the rate that they consume smaller black holes. The redshift dependence of such mergers would also help to understand how black holes grow in size throughout the history of the Universe.

The stellar-mass black holes mergers that have been observed by LIGO and Virgo could also be seen by LISA. In some cases, it might be possible for LISA to observe the inspiral signal from a binary black hole months to years before the merger would be observed by ground-based gravitational-wave detectors [225]. Such an observation across numerous frequency decades would provide for interesting tests of general relativity. While the merger signal is outside of the LISA observing band, the observation of stellar-mass black hole binaries would contribute to our knowledge of such systems, and possibly also contribute to deciphering how such systems are formed.

LISA's observations of binary black hole mergers will allow for testing general relativity and alternative theories of gravity in multiple ways. The observation of a post-merger signal from binary black hole merger, namely the ringdown signal from the newly formed black hole, will help to confirm whether general relativity truly is the theory that describes black holes. It would also help to confirm that these objects are, in fact, black holes. Gravitational waves from extreme mass ratio inspirals also have a more complicated structure than when the two masses are approximately equal. Comparing such signals to models will be another test of general relativity. These signals can also be used to confirm that gravitational-waves travel at the speed of light. The presence of dark matter (such as axions) in the environment of an extreme mass ratio inspiral could also affect the gravitational-wave signal.

Just as LIGO and Virgo have already shown that the use of gravitational waves to measure the Hubble constant is possible, LISA will also contribute to this important measurement of the expansion of the Universe. This will be especially true when the source can be found, for example, with an electromagnetic counterpart. Any electromagnetic signal from a binary black hole merger would provide critical information on the environment about the black hole. Such observations at very high redshift would also give information on the dark energy content of the Universe, as the Hubble constant does change of the history of the Universe (so it is not really a constant, as per its name).

LISA will also attempt to measure the stochastic gravitational-wave background [220]. It will certainly observe a stochastic background made by binary black hole mergers throughout the Universe. There will also be a foreground signal from the binary systems galaxy. In order to try to measure a cosmologically produced stochastic background methods must be developed to simultaneously characterize the galactic foreground [226], and the binary black hole produced background [227]. In the LISA observational band a cosmologically produced stochastic background could be produced by phase transitions in the early Universe or cosmic strings. The gravitational waves produced from inflation are likely too small for LISA to observe.

LISA will be observing thousands of signals and have the opportunity to make significant discoveries. See [140] for a more complete description of the scientific goals for LISA.

Pulsar timing observations in the decades to come will also be contributing important science. At the very low frequencies (nanoHertz) there is the opportunity to observe gravitational waves from supermassive black hole binaries, with masses of billions of solar mass [228]. The addition of the Square Kilometer Array for the radio observation of pulsars will be an important contribution for the future [142–144].

Gravitational-wave astronomy started in earnest in 2015 with the first direct detection of gravitational waves by LIGO. In the five years since LIGO and Virgo have announced 50 events, with many more arriving in the years to come. The future looks bright for gravitational-wave science in the decades to come. With detectors with even better sensitivities, and operating in many observation bands, the scientific output should accelerate. However, pushing the detection technology is not easy, and will require substantial work and resources.

Acknowledgments

The production of this article was supported by funds from the Observatoire Côte d'Azur, and U.S. National Science Foundation grant PHY-1806990.

2.7 Stars, the Sun, and planetary systems as physics laboratories

Patrick Eggenberger[1]
[1]University of Geneva, Geneva, Switzerland

2.7.1 General overview

As a very simple and general definition, stars can be seen as *spheres of hot gas*. While such a definition may seem at first sight to be too simplistic to be meaningful, it already contains some key aspects of the physical nature of stars. First, the spherical shape of stars directly reveals the dominant role played by gravity for these objects. This is of course directly related to the important mass of these objects that enables gravity to dominate over electric forces, which are responsible for the shape of the objects that we encounter in our daily life. This evolution of the shape of the objects according to their sizes as a result of the competition between gravity and electric forces is visible when observing the different celestial bodies of our Solar System. One indeed notes that the small objects of the Solar System like asteroids exhibit various shapes, which differ from the spherical symmetry that characterizes the more massive objects like the Moon, the planets, or the Sun.

2.7.2 The mechanical equilibrium of a star

While the first part of the simple definition of stars as spheres of hot gas underlines the dominant role played by gravity in these objects, the second part underlines the need to have another force in order to compensate for the contraction that would be induced by gravity alone. A key element of stellar evolution is indeed related to the fact that during most of their lifetime, stars are in hydrostatic equilibrium. In order to reach this equilibrium state, the gravity force acting on a fluid element inside the star must be compensated by a force of the same amplitude but acting outwards. This is where the second part of the definition related to hot gas comes into play. Indeed, this force is directly related to the properties of the matter in stellar interiors: this is the pressure of the gas, and even more precisely the variation of this pressure with the radial distance inside a star that is able to sustain a star against gravity. This constitutes the first basic equation of stellar structure, the equation of hydrostatic equilibrium:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\rho\frac{GM_r}{r^2}, \tag{2.37}$$

where the radius r is defined as the distance to the stellar center, P is the pressure, ρ the density, G the constant of gravitation, and M_r the total mass inside a sphere of radius r. This equation is directly obtained by expressing the difference of pressure dP in a thin shell inside a spherically symmetric star between a radius r and $r + dr$. Recalling that the local gravity $g = \frac{GM_r}{r^2}$, one sees that this equation simply expresses that at any point inside a star, the pressure gradient must sustain the stellar matter against gravity in order to reach an equilibrium state.

While the first basic equation of stellar structure equation (2.37) constitutes a very simplified version of the conservation of angular momentum, the second basic equation of stellar structure is simply obtained by expressing the conservation of the mass:

$$\frac{\mathrm{d}M_r}{\mathrm{d}r} = 4\pi r^2 \rho. \tag{2.38}$$

It is interesting to notice that in theses two equations (2.37) and (2.38), only the pressure P and the density ρ appear explicitly and not the temperature T. To solve the equations of stellar structure, one needs to know the microphysics and in particular the equation of state of the stellar matter that expresses the relation between the three physical parameters P, T, and ρ:

$$\frac{\mathrm{d}\rho}{\rho} = \alpha \frac{\mathrm{d}P}{P} - \delta \frac{\mathrm{d}T}{T}, \tag{2.39}$$

with $\alpha = (\frac{\partial \ln \rho}{\partial \ln P})_{T,\,\mu}$, $\delta = -(\frac{\partial \ln \rho}{\partial \ln T})_{P,\,\mu}$ and μ the mean molecular weight (i.e., the average number of atomic mass units per particle). The physics describing the state of the stellar matter is included in the different values assigned to α and δ. For instance, in the case of a perfect gas, α and δ are equal to one, while in the case of a degenerate non-relativistic gas, $\alpha = 3/5$ and $\delta = 0$. In the latter case, the density depends then only on the pressure and not on the temperature. As mentioned above, equations (2.37) and (2.38) form then a complete set of equations describing the internal structure of the star without any need to consider the thermal part of the problem. However, the pressure generally also depends on the temperature in stellar interiors as illustrated by the perfect gas case that is more representative of the state of the stellar matter during the main duration of the evolution of a star (the degenerate case describing the evolved phases of evolution of a star). In this case, there is an important coupling between the mechanical and the thermal state in stellar interiors.

2.7.3 The energy of a star

We are now interested in discussing in more detail the second part of our simple definition of stars as spheres of *hot gas*. As discussed above, a gradient of pressure is necessary in stellar interiors to ensure hydrostatic equilibrium. In the general case of an equation of state where the pressure depends both on the density and temperature (as for instance in the perfect gas case), this results in a gradient of temperature with a higher temperature in the central part of the star than at its surface. For instance, in the solar case, a central temperature of about 15 million of Kelvin (K) is reached, while the surface is characterized by a much lower temperature of about 5800 K. These differences of temperature lead to a transport of energy from the central layers to the surface: a star shines as a direct consequence of hydrostatic equilibrium and its lifetime then directly depends on the reservoir of energy available to maintain its luminosity.

This leads us to consider the different sources of energy available for a star. A first source of energy for a star consists in simply extracting this energy from its gravitational potential through a global contraction. Based on the simple equations derived in the preceding section, we can discuss the impact of a uniform contraction of a star. As a result of such a contraction, there is a change of pressure dP and of radius dr. From equation (2.37) expressing hydrostatic equilibrium, these two variations are then related by:

$$\frac{\mathrm{d}P}{P} = -4\frac{\mathrm{d}r}{r}. \tag{2.40}$$

Similarly, using equation (2.38) that expresses the conservation of the mass, the variation of density $d\rho$ is then related to the one of the radius by:

$$\frac{\mathrm{d}\rho}{\rho} = -3\frac{\mathrm{d}r}{r}. \tag{2.41}$$

Combining these two equations and using equation (2.39) for the general expression of the equation of state leads to the following relation between the variations of temperature and density during a slow uniform contraction of a star:

$$\mathrm{d}\ln T = \frac{4\alpha - 3}{3\delta}\mathrm{d}\ln\rho. \tag{2.42}$$

In the perfect gas case representative of stellar matter in non-degenerate conditions, the coefficients α and δ are equal to one so that a slope of 1/3 is obtained from equation (2.42) between the variations in temperature and density. This situation corresponds to the case of a star that extracts its energy from its gravitational potential only. In order to estimate the lifetime of a star based on this sole source of energy, one can simply divide its total gravitational energy by its luminosity L to obtain:

$$t_{\mathrm{KH}} \cong \frac{GM^2}{RL}, \tag{2.43}$$

where M, R, and L are the total mass, radius, and luminosity of the star, respectively. The timescale t_{KH} is referred to as the Kelvin–Helmholtz timescale and corresponds to the lifetime of a star that would only be able to produce its luminosity at the sole expense of its gravitational energy. By introducing numerical values in equation (2.43) a typical duration of about 30 million of years is obtained for the Sun. In this context of the study of the lifetime of a star, the interest of a multi-disciplinary approach can be nicely illustrated. Indeed, in the case of the Sun, the lifetime predicted by the Kelvin–Helmholtz timescale can be directly compared to other constraints available for the age of the Solar System. In particular, a comparison with geophysical constraints coming from the dating of the oldest rocks analyzed on Earth reveals that this gravitational timescale is much too short to be compatible with the geophysical understanding of the evolution of our planet. This disagreement reveals that another important source of energy is at work in stellar interiors. This additional source of energy is not needed to explain the luminosities of the stars (since as seen above they can be easily reproduced from the sole gravitational energy) but to explain how these luminosities can be maintained over long durations and in particular over billion of years instead of a few million of years as predicted by the Kelvin–Helmholtz timescale.

In order to discuss the possible additional sources of energy in stars, we first notice that the slow uniform gravitational contraction discussed above leads to a simultaneous increase in density and temperature in central stellar layers under non-degenerate conditions (see equation (2.42) with $\alpha = \delta = 1$ for the perfect gas case). When the stellar matter is in degenerate conditions, the situation is completely different and a contraction can then lead to a cooling of the matter instead of an increase in the temperature. This is of course directly related to the equation of state of a completely degenerate non-relativistic gas, and in particular to the fact that in this case the density depends only on the pressure (and not on the temperature) with $P \propto \rho^{5/3}$ (equation (2.42) with $\alpha = 3/5$ and $\delta = 0$). For a perfect gas, $P \propto \rho T$, so that the transition between the non-degenerate and degenerate cases occurs when the two pressures are equal (i.e., when $\rho T \propto \rho^{5/3}$). This leads to $T \propto \rho^{2/3}$ and hence to a slope of 2/3 for the transition between the non-degenerate and degenerate conditions that can be compared to the slope of 1/3 obtained above for a slow contraction in non-degenerate conditions. This means that a star will begin its evolution in non-degenerate conditions with its central values of temperature and density evolving along this 1/3 line, but that at a given point of its evolution, the frontier defined by the slope of 2/3 will be reached and the central stellar layers will then evolve in degenerate condition (see figure 2.43). Close to this transition from non-degenerate to degenerate conditions, the coefficients α and δ will then evolve from the value of 1 to tend to the values of 3/5 and 0 that characterize the completely degenerate case. The value of α becomes lower than 4/3 before the value of δ reaches 0; consequently, the coefficient of proportionality between the variations in temperature and density given by equation (2.42) becomes negative. A contraction is then found to produce a cooling of the stellar matter in the central regions so that an increase in the central temperature can no longer be obtained in degenerate conditions.

This different behaviour in non-degenerate and degenerate conditions is of fundamental interest in the discussion of the additional sources of energy needed in stellar interiors and for the evolution of a star in general. Indeed, thermonuclear reactions can produce this additional energy in stars. For these reactions to occur, specific physical conditions are needed and in particular high values of temperatures and densities must be reached. As discussed above, these conditions of high densities and temperatures can be obtained in central stellar layers from a slow contraction in non-degenerate conditions. A star then begins its evolution by extracting its energy solely from its gravitational potential on a characteristic timescale given by the Kelvin–Helmholtz timescale (equation (2.43)). This leads to a continuous increase of both the central temperature and density until the values needed for thermonuclear reactions to occur are met.

The first reaction to take place in stellar cores is the hydrogen burning during which the production of energy is obtained from nuclear binding energy by

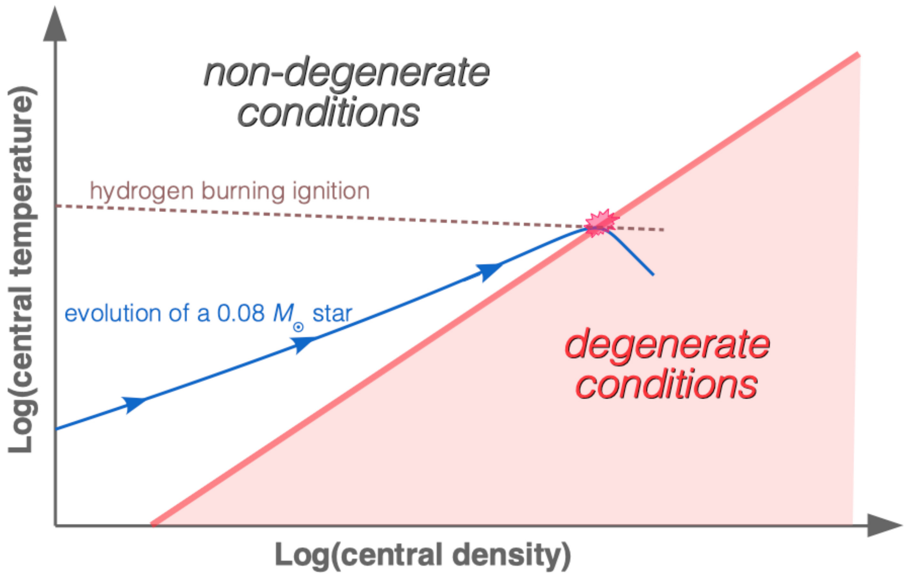


Figure 2.43. Illustration of the change in the central temperature and pressure during the evolution of a star (continuous blue line with a slope of about 1/3 in such a diagram). The transition from the perfect gas to the degenerate region (red shaded area) is indicated by the red line with a slope of about 2/3. The dashed line shows the values of temperature and density needed for the ignition of core hydrogen burning.

transforming four hydrogen nuclei into one helium nucleus. Hydrogen burning can occur through two main reactions: the proton–proton chain and the CNO cycle. For the proton–proton chain, two protons are first needed to form a deuterium nucleus, which then interacts with another proton to form ^{3}He. The last step in the formation of ^{4}He can be done in different ways, with either the interaction of two ^{3}He nuclei to obtain ^{4}He (referred to as the proton–proton1 chain), or the interaction with existing ^{4}He nuclei (referred to as the proton–proton2 and 3 chains). Another possibility for hydrogen burning is referred to as the CNO cycle. In this case, isotopes of carbon, nitrogen, and oxygen are needed in order to transform hydrogen into helium. Hydrogen burning through the proton–proton chain is found to dominate over the CNO cycle at relatively low temperatures (lower than about 1.5×10^7 K), while hydrogen burning through the CNO cycle dominates at higher temperatures.

An estimate of the characteristic timescale related to thermonuclear reactions as the source of energy of a star can then be obtained from this transformation of hydrogen into helium. A relative mass defect of about 0.7% is obtained for the transformation of four protons into one ^{4}He nucleus. The corresponding timescale is then simply obtained by dividing the total energy produced by nuclear reactions by the luminosity of the star. If nuclear reactions would be able to occur in the whole interior of a star of total mass M, this energy would then simply be equal to 0.007 Mc^2, where c corresponds to the speed of light. The values of temperature and density are, however, high enough for hydrogen burning to occur only in the central regions of a star. This leads to the introduction of a factor q (lower than 1) in the preceding expression to account for this fact. The characteristic nuclear timescale is then simply expressed as

$$t_{\text{nuclear}} \cong 0.007 \frac{qMc^2}{L}. \tag{2.44}$$

Introducing numerical values corresponding to the solar case in this equation together with $q = 0.1$ (thus assuming that nuclear reactions are only at work in the 10% of the total mass of the star corresponding to the central layers), one obtains a characteristic timescale of about 10 billion years. This is much longer than the typical lifetime of about 30 million years estimated from the sole gravitational energy and enables to reconcile the theoretical modeling of the evolution of the Sun with the different constraints on the age of the Solar System and of our planet. In addition to providing an important source of stellar energy, thermonuclear reactions play a key role in the chemical evolution of a star by changing the chemical composition in its central layers. This is of course at the basis of stellar nucleosynthesis and in particular of the creation of chemical elements until the iron-peak elements.

Similarly to the cases of slow contractions in non-degenerate and degenerate conditions discussed above, the impact of thermonuclear reactions is also completely different depending on the state of the stellar matter. In non-degenerate conditions, the nuclear energy production in a star is self-regulated. Indeed, when more energy is produced by nuclear reactions in the central regions, an expansion of these layers will follow as a result of the increase in temperature. This expansion will then produce a decrease of the temperature and a corresponding decrease in the energy production rate by nuclear reactions. In degenerate conditions, an increase in temperature will also occur in the stellar core as a result of an excess of energy production, but this will not lead to an increase in pressure due to the particular equation of state of a degenerate gas with the pressure that does not depend on the temperature. As a result, no regulating effect is at work in degenerate conditions so that nuclear reactions are unstable and will continue to produce an increasing amount of energy that will finally result in a flash or an explosion. Conversely, nuclear reactions are stable in non-degenerate conditions due to the dependency of the pressure on temperature.

2.7.4 Evolution as a function of the initial mass

The simple physical aspects of stellar structure discussed above enable to draw a global view of the different evolutions of stars as a function of their initial masses. This can be understood from figure 2.43. We begin by recalling the two key points discussed in the preceding section. First, a star will approximately evolve along a line with a slope of 1/3 in such a plot in non-degenerate conditions (see equation (2.42)) and this leads to the simultaneous increase in both temperature and pressure in the stellar core. When sufficiently high values of temperature and pressure are reached, nuclear burning can be ignited. Second, the separation between non-degenerate and degenerate conditions is given by a line with a slope of 2/3 in figure 2.43 so that a star globally evolves towards degenerate conditions. Once the star enters this domain of degenerate gas, a contraction does not result anymore in an increase in the temperature, but leads instead to a cooling of the central layers, so that the conditions for nuclear burning ignition cannot be reached anymore.

- Brown dwarfs and planets

As can be seen in figure 2.43, there exists a minimal initial mass for an object to reach central physical conditions that enable the ignition of hydrogen burning. This is directly related to the moment when the star will enter the degenerate phase. If an object is not massive enough, the slow gravitational contraction occurring on a Kelvin–Helmholtz timescale will increase its central temperature under non-degenerate conditions, but without being able to reach a sufficiently high value to start hydrogen burning before entering the degenerate phase. As discussed above, once central layers are in degenerate conditions, there is no possibility for a contraction to increase further the central temperature so that physical conditions favorable to the fusion of hydrogen cannot be reached. The mass limit for this to happen is typically of about $0.08M_{\odot}$ (see figure 2.43); this corresponds to the minimal mass for a star, and objects with initial masses below $0.08M_{\odot}$ are brown dwarfs ans planets.

Interestingly, there is also a minimal mass that can be reached for the formation of these objects following the contraction and the subsequent fragmentation of an interstellar cloud. This minimal mass results from the need to radiate away the energy related to the contraction of the cloud in order for fragmentation to proceed. This leads to a minimal mass of about $0.01M_{\odot}$ for an object to be formed from the initial contraction of a cloud. This fragmentation limit is lower than the limit of about $0.08M_{\odot}$ discussed above for hydrogen burning to be ignited in non-degenerate conditions. The mass intervalle between about 0.01 and $0.08M_{\odot}$ is typically the domain of brown dwarfs. This fragmentation limit also shows that objects with masses lower than about $0.01M_{\odot}$ cannot be formed by fragmentation. This means that another formation mechanism is needed for planets. The latter are formed by collision and accumulation of rocks, ice, and gas in protostellar disks surrounding young forming stars.

- Low-mass stars

When the initial mass of the object is larger than about $0.08M_{\odot}$, hydrogen burning can take place in the central layers and one leaves the domain of substellar objects to enter the stellar domain. As discussed in the preceding section, the typical timescale corresponding to this new source of nuclear energy is much larger than the one associated with the energy released by the sole gravitational contraction. Once hydrogen burning is ignited, the star is then in a long-standing evolutionary phase referred to as the main-sequence phase. This phase will last as long as hydrogen is present in the central stellar layers and will simply end once all the central hydrogen has been transformed into helium by nuclear reactions.

After the main-sequence phase, the central layers of the star contract to extract energy from the gravitational potential. This situation is then similar to the one preceding the ignition of hydrogen burning, with the contraction resulting in an increase in the central stellar temperature as long as central layers are in non-degenerate conditions. This leads to another mass limit that is defined by the possibility for a star to reach a sufficiently high central temperature for helium burning ignition before the helium core becomes too degenerate. This limit is found around $0.5M_{\odot}$ and stars with an initial mass lower than this value will then be able to ignite hydrogen burning, but not helium burning.

- Solar-type stars: the Sun as a reference star

Stars with initial masses above about $0.5M_{\odot}$ will then experience both hydrogen burning and helium burning. Our Sun is the perfect representative of stars evolving on the main sequence in this mass range. Owing to its proximity, the global and internal physical properties of the Sun can be determined with exquisite precision, which makes the Sun a fundamental reference for the modelling of the complex physical processes acting in stellar interiors. For instance, a precise determination of the mass and radius can be achieved for the Sun. Interestingly, this can be directly used to obtain a simple estimate of the values of the pressure and temperature in the solar core. To illustrate this, we first approximate the left-hand side of the equation expressing hydrostatic equilibrium (equation (2.37)) by:

$$\left|\frac{\mathrm{d}P}{\mathrm{d}r}\right| \cong \frac{P_{\odot}^{\mathrm{Core}} - P_{\odot}^{\mathrm{Surface}}}{R_{\odot}} \cong \frac{P_{\odot}^{\mathrm{Core}}}{R_{\odot}}. \tag{2.45}$$

Similarly, the right-hand side of equation (2.37) can be simply estimated using rough average values for the mass, radius, and density in the solar interior based on the total mass and radius of the Sun: $M_r \cong M_{\odot}/2$, $r \cong R_{\odot}/2$ and a mean density $\rho \cong 3M_{\odot}/(4\pi R_{\odot}^3)$:

$$\rho\frac{GM_r}{r^2}r^2 \cong \frac{3}{2\pi}\frac{GM_{\odot}^2}{R_{\odot}^5}. \tag{2.46}$$

Using the equation of hydrostatic equilibrium (equation (2.37)) with the approximations given in equations (2.45) and (2.46) leads to the following estimate for the pressure in the central layers of the Sun:

$$P_{\odot}^{\mathrm{Core}} \cong \frac{3}{2\pi}\frac{GM_{\odot}^2}{R_{\odot}^4}. \tag{2.47}$$

Introducing numerical values for the solar mass and radius in this expression leads to a central pressure in the Sun of about 5.4×10^{15} g s^{-2} cm^{-1}. An estimate of the central temperature of the Sun can be directly obtained from equation (2.47) by adopting an equation of state. The hypothesis of a perfect gas is well suited for a main-sequence star like the Sun and in this case the pressure P is related to density ρ and temperature T by:

$$P = \frac{k}{\mu m_{\mathrm{u}}}\rho T, \tag{2.48}$$

where m_u is the atomic mass unit and k corresponds to the Boltzmann constant. Combining equations (2.47) and (2.48) (using again that $\rho \cong 3M_\odot/(4\pi R_\odot^3)$) leads to the following estimate for the temperature of the solar core:

$$T_\odot^{\rm Core} \cong 2\frac{\mu m_u}{k}\frac{GM_\odot}{R_\odot}. \tag{2.49}$$

The chemical composition of the Sun being largely dominated by hydrogen, we can simply approximate the value of the mean molecular weight μ by the one of pure ionized hydrogen ($\mu = 1/2$) to obtain a temperature of about 2.3×10^7 K for the central layers of the Sun. While this value for the central temperature of the Sun is obtained from very crude approximations, it is nevertheless in relatively good agreement with the more realistic determination of 1.5×10^7 K based on the computation of solar models.

The fact that the properties of the Sun can be precisely determined (see next section for the determination of the solar internal properties) is also of prime interest to discuss the internal structure of stars and in particular the relation between this internal structure and the transfer of energy. Indeed, above the limit of about 0.5 $M_\odot$, the whole stellar interior is not fully convective and both radiative and convective zones are present depending on the local physical conditions. In the specific case of the Sun, the internal region is radiative below about 70% of the total solar radius, while the remaining external layers constitutes the convective envelope. This leads us to briefly discuss here the internal transport of energy in a star. We first discuss the radiative transfer of energy, which is the transport of energy by photons. Energy is produced by nuclear reactions in the central layers and we thus consider a photon emitted in the stellar core. This photon will then travel on a typical distance between two interactions that is called the mean-free path. In the solar case, the length of this mean-free path is of the order of 10 mm in the deep interior and increases to about 1 cm in the envelope. This typical length for the travel of a photon between two interactions is thus much smaller than the solar radius: a photon emitted in the core will then undergo a very large number of interactions in its way to the surface that will only be reached after a time of the order of the Kelvin–Helmholtz timescale. One can also compare this mean free path to the typical variation of temperature in stellar interiors. Such a variation can be very simply estimated from the solar properties, and in particular from the temperature in the solar core obtained above (equation (2.49)):

$$\left|\frac{{\rm d}T}{{\rm d}r}\right| \cong \frac{T_\odot^{\rm Core} - T_\odot^{\rm Surface}}{R_\odot} \cong 10^{-4}\ {\rm K\ cm^{-1}}. \tag{2.50}$$

Adopting a value of the mean-free path of 1 cm, one then obtains that the variation of temperature is typically only of about 10^{-4} K during the travel of a photon between two interactions. This very small variation compared to typical temperatures of million of K in the solar interior is at the basis of the fundamental assumption of local thermodynamic equilibrium in stellar interiors, which implies that the radiation field at a given radius r inside a star can be simply approximated by the radiation emitted by a black body using the local temperature $T(r)$ at this radius r. Transport of energy by radiation is present in the whole stellar interior. However, depending on the different physical conditions characterizing stellar interiors, the efficiency of energy transport by radiation can be found to be insufficient to transfer a given heat excess. This happens when the opacity of the stellar matter is too high to enable an energy transfer by radiation alone. In this case, convection is active in order to transport efficiently energy through turbulent motions. As mentioned above, low-mass stars with initial masses below about $0.5M_\odot$ are fully convective, while more massive stars like the Sun are characterized by a radiative core and a convective envelope. The depth of this convective envelope decreases when the initial mass of the star increases, due to the corresponding increase in temperature and related decrease in the opacity of the external layers of the star. For a solar chemical composition, this convective envelope disappears for stars with initial masses typically larger than about $1.5M_\odot$, which exhibit radiative envelopes. Concerning central layers, a convective core appears for main-sequence stars with initial masses typically larger than about $1.2M_\odot$ (for a solar chemical composition) as a result of the transition from hydrogen burning through the proton–proton chains to hydrogen burning through the CNO cycle discussed in the preceding section.

The end of the main-sequence phase corresponds to the exhaustion of hydrogen in the central layers of these stars. At this stage, the physical conditions in the core are not favorable for the ignition of helium-burning: in particular the central temperature is still lower than the typical value of about 10^8 K required for helium burning. Consequently, there is a contraction of the central layers, where no nuclear reactions are taking place due to the lack of hydrogen fuel, together with hydrogen burning in more external layers where hydrogen is still present. This post-main sequence phase is then characterized by the contraction of the central layers on a Kelvin–Helmholtz timescale and the simultaneous expansion of the envelope that leads to a rapid decrease in the surface temperature of the star that becomes a red giant star. The subsequent evolution of the star as a red giant follows the contraction of the central layers and the corresponding increase in luminosity together with the expansion of the envelope. Due to contraction, the central values of density and temperature continue to increase: the density then becomes high enough for the gas to become degenerate, while the central temperature is high enough for helium burning to be ignited in the core. As discussed above, nuclear reactions are unstable in degenerate conditions: this ignition of helium burning leads then to an important increase in temperature on a very short timescale called the helium flash. However, this flash does not result in the disruption of the star, but is able to remove the degeneracy in the central stellar layers, and core helium burning can then proceed in non-degenerate conditions. The star is then in another long-standing evolutionary phase, the core helium burning phase, during which energy is obtained from the triple alpha reaction, which is the formation of ^{12}C from three ^{4}He nuclei. This phase ends once helium fuel is exhausted in the central stellar layers. The star evolves then on the asymptotic giant branch and will end up as a white dwarf for which the degeneracy pressure of electrons is able to counteract gravity.

- Intermediate-mass stars

As discussed above, stars with masses typically larger than about $1.2M_{\odot}$ (for a solar chemical composition) exhibit a convective core while evolving on the main sequence. For these stars, hydrogen burning takes place in well-mixed cores, because of the efficient transport of chemicals by the turbulent convective motions. This results in helium core of significant mass at the end of the main sequence, with an increase of the mass of the core with the initial mass of the star. For stars with an initial mass typically larger than about $2.3M_{\odot}$ (for a solar chemical composition), the contraction of this helium core after the main sequence always occurs in non-degenerate conditions until the ignition of core helium burning. Contrary to lower-mass stars, these stars will not experience the helium flash and this difference marks the transition to intermediate-mass stars. As for lower-mass stars, the ignition of core helium burning in intermediate-mass stars stops the contraction of the central layers and the star remains in this long-lasting phase until exhaustion of helium in the core. After the core-helium burning phase, stars with masses lower than about $8M_{\odot}$ evolve on the asymptotic giant branch and end up as white dwarfs. For more massive intermediate-mass stars with initial masses typically in the range $8–10M_{\odot}$, the evolution after the core-helium burning phase is more complicated. Depending on key physical parameters as the efficiency of mass-loss at the surface of the star or the modelling of convection, the end point of evolution could be a white dwarf or the collapse of the star due to electron capture.

- Massive stars

Stars more massive than about $10M_{\odot}$ (always for a solar chemical composition) are massive enough to be able to ignite all the different nuclear burning phases in their central layers during their evolution. This marks the transition from intermediate-mass to massive stars. Similarly to lower-mass stars, these massive stars will begin with core-hydrogen burning and core-helium-burning, but then the contraction of their central layers will continue in order to be able to successively ignite carbon burning, neon photodisintegration, oxygen, and finally silicon burning. After these different phases of nuclear burning, an iron core is present, which corresponds to the end of this series, because no exothermic fusion is then possible. At this stage, the internal structure of the star is characterized by different layers exhibiting heavier elements when the distance to the stellar centre decreases. This characteristic 'onion-skin' chemical structure directly results from the successive nuclear reactions that were at work in the stellar core as the central temperature increased during the evolution of the star. This shows that heavy elements until the iron-peak elements are natural products of nuclear reactions occurring in the core of massive stars. Of course, chemical elements heavier than iron are not produced in the central layers of stars, but can be processed during the supernova explosion that characterized the end of the evolution of a massive star. These supernova explosions will then result in the chemical enrichment of the interstellar medium, and successive generations of stars and planets can then be formed from an interstellar medium that contains a higher and higher fraction of heavy elements. The nature of the remnant of a massive star depends on the final mass of the iron core. If this mass is lower than a given threshold (typically of about $2 - 3M_{\odot}$), a neutron star is formed. Similarly to white dwarfs where the degeneracy pressure of electrons is able to counterbalance gravity, the pressure of degenerate neutrons is able to counteract gravity in the case of a neutron star. However, if the mass of the iron core is larger than this threshold, the degeneracy pressure of neutrons is not anymore able to stop the collapse, which results in the formation of a black hole.

2.7.5 Challenges and opportunities

In the preceding section, the basic physical ingredients needed to model stars were briefly described together with their evolution according to their initial mass. These results rely on models obtained by solving the equations of stellar structure. The computations of these models require of course many different physical inputs, going from microphysics with the equation of state and opacity of stellar matter or nuclear reaction rates, to macro-physics with the modelling of convection and magneto-hydrodynamic transport mechanisms in stably stratified radiative regions.

Observational constraints are then of fundamental importance in order to progress in the theoretical modelling of these various physical processes taking place in stellar interiors. Global stellar properties observed either in photometry such as the luminosity of the star or in spectroscopy with the surface temperature and surface gravity of the star are the first observational constraints used for a comparison with theoretical predictions. In addition to these global properties, surface chemical abundances can be obtained by spectroscopic measurements. The aim of stellar models is then to be able to reproduce these different surface properties either for one specific well-studied target or for a large population of stars.

The comparisons of theoretical models with observed surface properties are of course particularly useful to test different physical inputs of the models, but suffer from a key limitation: it's very difficult to constrain the theoretical modelling of a whole internal structure of a star or a planet, and hence the various physical processes at work in the deep layers of these objects, by only having access to its surface properties. In this respect, spectroscopic determinations of surface chemical composition can nevertheless be particularly useful for the modelling of transport processes in radiative zones. This is for instance the case of light elements like lithium and beryllium in solar-type stars and of nitrogen surface abundance in more massive stars that can reveal the transport of material processed by the CNO cycle from the stellar core to the surface. We can also enlight here the specific cases of intermediate-mass stars exhibiting peculiar surface abundances and of the massive Wolf-Rayet stars. The former stars reveal the role played by radiative forces on the relative surface abundances of chemical elements, while the latter stars nicely illustrate the key impact of mass loss on the evolution of massive stars. Indeed, strong mass loss, either due to stellar winds in single stars or to mass transfer in binary stars, are able to remove the surface layers of Wolf-Rayet stars, so that the chemical composition of internal layers, which has been modified by the different nuclear reactions, can then be observed at their surface.

While observing surface properties for these specific targets can bring some indirect constraints on stellar interiors, the main challenge consists of course in obtaining direct observational constraints on the internal physical properties of stars. To address this key question, we recall that the internal structure of the Earth can be studied thanks to the analysis of seismic waves. These waves can either be artificially generated by hitting the ground at fixed frequencies in order to probe the structure of layers close to the surface, or be produced by an earthquake and then travel through deeper layers of the Earth to give some information about the internal structure of our planet. By analogy to the seismic probing of our planet, one can then wonder whether a 'starquake' could happen and the analysis of the generated waves be used to probe the internal structure of a star.

2.7.6 The Sun: the solar modelling problem

In the case of the Sun, this question has been beautifully answered by the observation and analysis of the so-called solar five-minute oscillations. To understand how a 'sunquake' can happen, we recall that convection is active in the envelope of the Sun in order to efficiently transfer energy in these external layers (see previous section). Due to these turbulent convective motions, waves can be generated and will then travel in the solar interior. These travelling waves can then give raise to standing waves, or global oscillation modes, whose properties depend on the internal structure and composition of the Sun. The characterization and analysis of these solar oscillations can then be used to probe the interior of the Sun similarly to the analysis of seismic waves in the case of the Earth: this is the goal of helioseismology, which is the study of solar oscillations.

The analysis of solar oscillations has led to key results regarding the internal structure of the Sun [229, 230]. Thanks to precise determinations of the frequencies of pressure modes (global oscillation modes related to acoustic waves), the solar sound speed and density profiles have been determined. These helioseismic data have also enabled a precise determination of the location of the transition from the radiative interior of the Sun to the convective envelope [231], as well as the determination of the abundance of helium in the solar envelope [232–234]. In addition to the well-determined global parameters of the Sun (mass, radius, luminosity, and age), these helioseismic measurements constitute of course fundamental constraints for theoretical solar models. Interestingly, standard solar models (i.e., models that do not take into account hydrodynamic or magnetic transport processes) based on the solar photospheric abundances determined from spectroscopy in the 90s [235] were found to be in fairly good agreement with these different observational constraints [236]. In particular, the location of the base of the solar convective zone as well as the helium content of the solar envelope predicted by these standard models were in perfect agreement with the helioseismic determinations. Moreover, the sound speed and density profiles deduced from helioseismic data were satisfactorily reproduced by these solar models. The success of these standard solar models based on the solar surface abundances available in the 90s then suggested that our global understanding of the Sun was quite good.

This situation completely changed about 15 years ago with the redetermination of the solar surface abundances and the significant decrease by about 25% of the heavy element abundances of the Sun [237, 238]. This important change in the spectroscopic determination results from the use of three-dimensional models of the atmosphere (instead of the previous one-dimensional models), from a new consideration of the selection of the solar spectral lines and from the inclusion of effects related to non local thermodynamic equilibrium. Standard solar models computed by accounting for these revised photospheric abundances were then found to be unable to correctly reproduce the helioseismic constraints. In particular, the transition from the internal radiative zone to the convective solar envelope is predicted to be to close to the surface compared to helioseismic data, while the predicted helium surface abundance is too low compared to the helioseismic determination. Moreover, the sound speed and density profiles deduced from helioseismology are not correctly reproduced anymore [239–241]. This important tension between models based on the revised solar abundances and helioseismic constraints is referred to as the solar modelling problem. One of the main challenges concerning the Sun at the horizon 2050 will then be related to the study of the origins of this tension between solar abundances and helioseismic measurements with the final goal of solving the solar modelling problem.

Another key result of helioseismology is the determination of the internal rotation profile of the Sun. These measurements reveal an approximately uniform rotation in the radiative interior of the Sun with a rapid transition to differential rotation as a function of latitude in the solar convective envelope. These two key features of the internal rotation of the Sun have always been challenging to reproduce. Indeed, solar models that account for hydrodynamic transport processes related to rotation-induced instabilities predict a strong radial differential rotation in the solar radiative interior, which is in contradiction with the almost uniform rotation inferred from helioseismic measurements [242]. This shows that additional dynamical physical processes are missing in the solar interior, in particular regarding the efficient transport of angular momentum in the radiative zone. The effects of magnetic fields are of prime interest in this context. For instance, the strong coupling ensured by large-scale fossil fields can be invoked to reproduce the flat rotation profile in the radiative interior of the Sun [243–248]. However, such a scenario has some difficulties to simultaneously account for the second key observational constraint, which is the sharp transition from the approximately uniform rotation in the radiative interior to differential rotation in the convective zone [249, 250]. The effects of magnetic instabilities could then play an important role. In particular, the Tayler instability [251] could be at the basis of a small-scale dynamo leading to an efficient transport of angular momentum in stellar radiative zones [252]. Interestingly, such a process can simultaneously account for the approximately uniform rotation in the radiative interior of the Sun and the transition to differential rotation in its convective envelope [252, 253]. However, the existence of this kind of process has to be confirmed by simulations performed under realistic stellar conditions [249]. Obtaining numerical simulations of these complex hydrodynamic and magnetic processes under realistic stellar conditions is another main challenge for the years to come.

The different results obtained so far by helioseismology have been derived from the study of pressure oscillation modes. A big observational challenge for the Sun at the horizon 2050 is related to the detection of the solar gravity oscillation modes. This is a long and difficult quest owing to the extreme difficulty of detecting these very low-amplitude oscillation modes. Characterizing solar gravity modes would enable the determination of the rotation rates in the core of the Sun, which is not feasible from the rotational splittings of pressure modes. This is of course particularly interesting to better understand the specific case of the internal rotation of the Sun, but also for stars in general by directly constraining angular momentum transport in radiative interiors where strong chemical gradients able to inhibit the transport efficiency by turbulent processes are present [254]. In addition to the rotational properties of the central layers of the Sun, gravity modes would of course also offer an interesting possibility to better constrain the chemical mixture and opacities in the solar core that would perfectly complement the constraints coming from solar neutrinos [255].

2.7.7 From the Sun to distant stars

The beautiful successes obtained from the study of oscillation modes in the case of the Sun motivated the detection of solar-like oscillations for other stars. Using the same seismic techniques for more distant stars is of course of prime interest to probe the internal structure of stars with properties that can be quite different from the Sun: for instance, stars with different masses and chemical compositions, stars at various evolutionary stages, or even physical processes not at work in the Sun, such as mixing by convective cores. Having a direct observational access to the internal properties of different stars is needed to progress in our understanding of stellar evolution.

A first important issue to achieve this goal is related to the difficulty of detecting these oscillations for a distant star. Indeed, contrary to the solar case, another star is seen as a point source and its disc is not resolved. This has a direct consequence on the type of oscillation modes that can be detected for a distant star. Figure 2.44 illustrates the deformation of the surface corresponding to different oscillation modes. These modes can be identified with the degree l and azimuthal order m of the spherical harmonics, where l and m indicate the total number of nodal lines at the surface and the number of these lines that intersect the equator, respectively. While all these oscillation modes can be detected for the Sun, only the lowest-degree modes can be observed for a distant stars (typically modes with $l \leqslant 3$) due to the geometrical average of the deformation on the stellar disc. Fortunately, these low-degree oscillation modes are of particular interest, because they penetrate deeply in stellar interiors and can then bring key constraints on both global stellar properties (e.g., masses and radii) and properties of the stellar cores (e.g., central hydrogen abundance during the main sequence, size of convective cores). A second observational difficulty is related to the very low amplitudes of solar-like oscillations

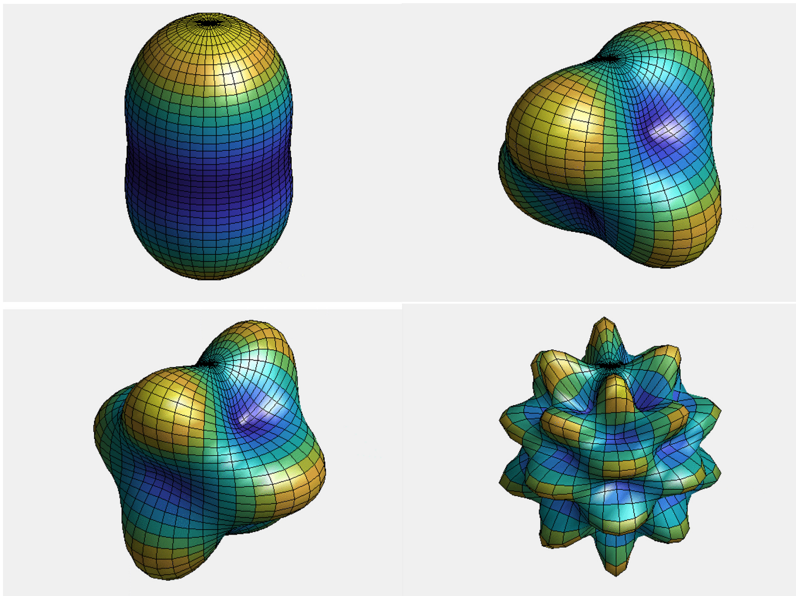


Figure 2.44. Deformation of the stellar surface due to global oscillations of the star. The amplitudes of these oscillations are largely exaggerated for illustrative purpose. Oscillations modes with 2 nodal lines at the surface ($l = 2$) and none of this nodal line that intersects the equator ($m = 0$) is shown in the top-left panel. A mode with 3 nodal lines at the surface ($l = 3$) and 2 nodal lines that intersect the equator ($m = 2$) is shown in the top-right panel. Oscillation modes with $l = 4$, $m = 2$ and $l = 10$, $m = 4$ are shown in the bottom left and right panels, respectively. (Courtesy of Sébastien Salmon, University of Geneva.)

(typically about 20 cm s^{-1} in radial velocity for the Sun): these oscillations can then only be detected in radial velocity from the ground by using high-accuracy spectrographs or from space in photometry in order to avoid the variations of luminosities induced by atmospheric turbulence.

The observational effort aiming at detecting solar-like oscillations for distant stars has led to the characterization of these oscillations for a large number of stars with various properties, in particular thanks to the space-based missions CoRoT, *Kepler* and TESS. The study of these stellar oscillations, or asteroseismology, can then bring different type of constraints for stellar evolution, from the determination of the global properties for a large population of stars to the detailed study of the internal structure of specific targets for which the most precise data are available. For instance, the masses and ages can be precisely determined for a large number of stars thanks to asteroseismology, which opens new perspectives for our understanding of stellar populations of the Galaxy and hence of the chemical evolution of the Milky Way [256, 257]. An important challenge for the years to come will be to extend these results in the metal poor regime and to dense fields of stars such as globular clusters [258].

Another key challenge for stellar physics is to progress in the modelling of the complex dynamical processes at work in stellar interiors. A first long-standing problem is of course related to the modelling of convection. The combined development of sophisticated numerical simulations of convection [259] and of asteroseismic constraints on the size of convective cores and shape of the border of these cores [260] is very promising to achieve significant progresses in this direction. A second key issue in stellar physics is related to the modelling of transport of angular momentum and chemicals in radiative stellar interiors. The observation of solar-like oscillations for subgiant and red giant stars has led to the measurements of rotational splittings of mixed oscillation modes. This has enabled the determination of the core rotation rates for these stars [247, 261–263]. The comparison with theoretical models has revealed that purely hydrodynamic processes do not provide an efficient enough transport of angular momentum to reproduce the observed core rotation rates [264–266] and that the efficiency of the missing transport processes can be precisely quantified thanks to these seismic data [267, 268]. This constitutes a unique opportunity to reveal the physical nature of these undetermined transport mechanisms in stellar radiative zone, which will be one of the main challenge of stellar physics for the years to come.

2.7.8 The star–planet connection

With the first detection of a planet around a solar-type star [269] and the following observations of numerous planets around various stars other than the Sun, the study of exoplanets rapidly became a central topic of modern astrophysics. These exoplanets show a great varieties of masses and sizes going from small rocky cores to massive gaseous planets on different kind of orbits (http://exoplanet.eu). Exoplanetary science is thus a very vast topic and we will only briefly touch here on the fundamental link between the stellar physical processes discussed above and the physics of the planets. With thousands of exoplanets with precisely characterized properties observed so far, one important challenge for the coming years is indeed related to the understanding of the evolution of these planetary systems studied as a whole.

A first obvious link between stars and planets is of course related to the need to have precise knowledge of the host-stars properties to derive the planetary parameters. For instance, a determination of the mass and the radius of the host star is needed to obtain the planetary mass from radial-velocity measurements and the planetary radius from transit observations, respectively. In order to study the evolution of planetary systems, one also needs a precise determination of the age of these systems. This is a particularly challenging goal, which can be successfully achieved thanks to the capability of asteroseismology to probe the deep stellar layers in order to reveal the central hydrogen abundance (i.e., the quantity of nuclear fuel still available for a main-sequence planet-host star). The observation of solar-like oscillations is then a powerful means to precisely determine the properties of the planet-host star (in particular its mass, radius, and age) needed to fully characterize the planetary system. This complementarity between asteroseismic and exoplanetary observations is well illustrated with space-based missions aiming at simultaneously detecting transiting exoplanets and performing asteroseismic measurements to characterize planetary systems such as TESS [270] or the future PLATO mission [271].

Being able to precisely determine the current state of a planetary system is the first mandatory step to study its past and future evolution. The different interactions between the host star and the planets have then to be taken into account to study the evolution of the system. A first important interaction between a star and a planet is the tidal interaction. To follow the orbital evolution of a planet due to tides implies to account for the transfer of angular momentum between the planetary orbit and the spin angular momentum of the star. The efficiency of tidal dissipation is also sensitive to the convective and radiative structure of stellar interiors as well as to the internal rotation of the star. Sophisticated rotating stellar models able to correctly predict the internal transport of angular momentum and the lost of angular momentum due to magnetized winds are then needed to coherently follow the rotational evolution of the host star together with the orbital evolution of the planet. This is another illustration of the link between exoplanetary science and stellar physics, in particular regarding the challenging question of the modelling of transport processes in stellar interiors discussed in the previous section. Developing such sophisticated models of rotating stars coupled with the orbital evolution of the planets due to tides constitutes one of the challenging questions to be addressed in the coming years to understand the evolution of planetary systems; preliminary attempts have thus been made in this direction [272–274].

Interestingly, these rotating stellar models including magnetic effects for the host star are also needed to study another fundamental star–planet interaction, namely the impact of the radiation emitted by the star on the planet properties. This is of particular relevance for the evolution of the high-energy flux received by a planet that will directly influence the photo-evaporation and the physical properties of its atmosphere [275, 276]. For solar-type and low-mass stars, the high-energy flux emitted by a star is indeed closely related to the rotational properties of its convective envelope, which determines the intensities of the magnetic fields that can be produced by a dynamo. The high-energy flux received by a planet can then change by orders of magnitudes during the evolution of a planetary system as a result of the rotational evolution of the host star [277, 278]. This evolution of the high-energy flux has to be coherently taken into account to study the evolution of planetary systems due to the photo-evaporation of the planets [240–281]. This nicely illustrates the complementarity between the physical description of stars and planets and the need for a multi-disciplinary approach to tackle the challenging questions related to the evolution of planetary systems at the horizon 2050.

2.8 Physics of the Earth's interior

Emmanuel Dormy[1]
[1] CNRS & Ecole Normale Supérieure, Paris, France

2.8.1 General overview

The Earth is, until now, the only planet we know on which lifeforms have developed and grown for several billions of years. Perhaps more surprisingly, it is also considered to be a living planet. By this we mean that it spontaneously evolves with time. Of course, we all know about the geothermal gradient: the temperature within the Earth increases with increasing depth. There are two reasons for that: first the Earth, very hot when it formed, is slowly cooling down with time, and second the presence of low radioactivity of its deep-seated rocks provides a source of internal heating.

When the Earth formed, it differentiated: the very heavy elements went to the centre, a place we now call the core of the Earth, mainly formed of iron (see figure 2.45). The comparatively lighter rocks floated and formed a deep layer

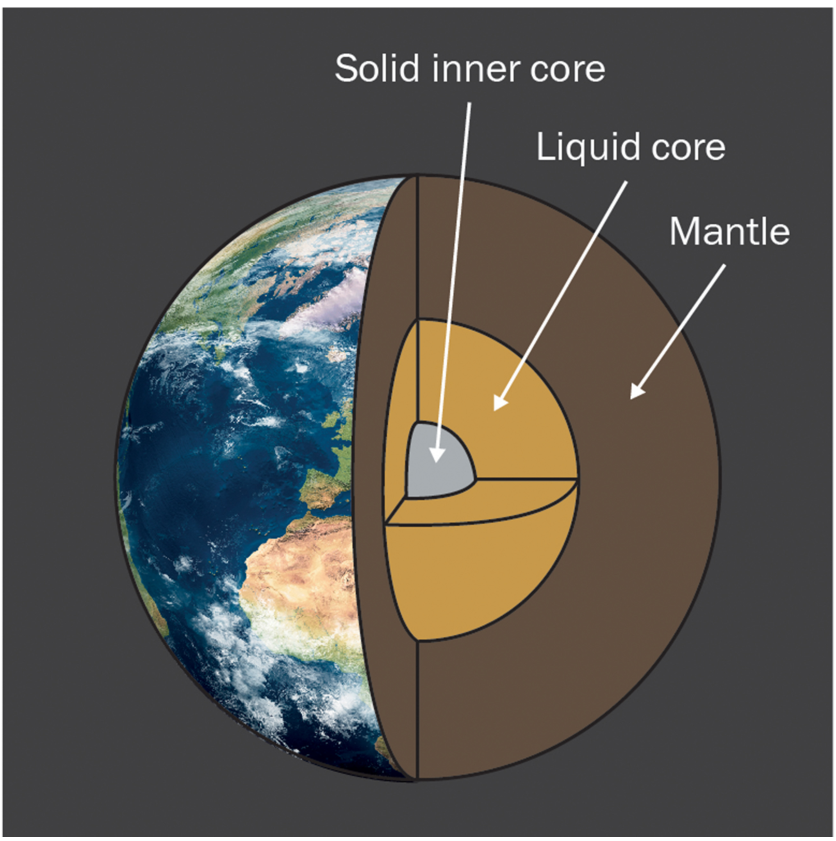


Figure 2.45. Cross-section of the Earth highlighting the simplified inner structure of our planet.

surrounding the core: the so-called mantle. The radioactivity in the mantle heats the rocks so as to keep them in a ductile state that allows the development of large-scale convective motion (through diffusion creep and dislocation creep). This slow convection is responsible at the surface for plate tectonics, the formation of mountains, earthquakes, and volcanism! About 3000 km below the surface, the core of the Earth is so hot that the iron is in a liquid phase, except for its very central part, where solid iron is met, as the pressure increase does more than compensate for the increase in temperature. Motions in the Earth liquid core are responsible for the Earth magnetic field, without which the compass would be of little use to find north.

During its long history, the geography of the Earth, its oceans, the composition of its atmosphere, its climate, and its biosphere (all forms of life) have undergone significant changes. The evolution of the Earth during its 4.5 billion years of existence is spectacular. It is only in the last century that we have begun to perceive the physical phenomena that have constantly altered it. Geophysics offers a venue for many branches of physics. When studying the interior of the Earth, the parameters are often extreme (in terms of pressure or temperature, for example) and the range of scales involved is (as we will see) phenomenal. The main specificity of this discipline being that theory, experiments, and simulations are completed by field work and *in situ* observations, which constitute the ultimate constraint over all other approaches. Of course, direct observations are limited to the surface of the Earth and a combination of physical insight, experiments, and indirect observations is needed to understand the processes at work within our planet.

In this short presentation, I will focus on three, and only three, open challenges in the physics of the Earth's interior. Such a choice is of course necessarily partial, yet I do hope that these questions can shed light on some aspects of research in geophysics and will motivate some of the readers to take an interest in the broad field of geophysics. Advanced texts such as the book by Turcotte and Schubert [282] can be referred to in order to develop a more complete picture of the current state of knowledge in this field.

The three subjects I wish to highlight concern the origin of the Earth magnetic field, of volcanic eruptions, and of earthquakes. If a motivation was needed to justify this choice, it would be only fair to say that the Earth magnetic field has played a major role through the centuries in the exploration of our planet and the discoveries of new worlds overseas; and since its appearance on Earth, mankind has been subjected to a changing environment, which includes deadly catastrophes such as volcanic eruptions and earthquakes.

2.8.2 The origin of the Earth's magnetic field

We have all played at least once in our life with a compass. The systematic orientation of the magnetised needle in the ambient field is mesmerising. The physical understanding of this observation is of course well established: the needle orientates itself according to the ambient magnetic field, which has both a strength and a direction. What is not so well understood, however, is where this field is coming from.

As we explained above, whereas the mantle is formed of silicate rocks, comparable to those found on the surface, the Earth's core is a spherical huge metallic ocean, mainly made of molten iron. The core probably formed 'rapidly' (in a geological sense, i.e., in 'only' some 100 million years) by differentiation. The temperatures in the core are over 3000 °C at the base of the mantle and about 5000 °C near the inner core. These temperatures keep the metal molten in spite of the considerable pressure, which varies from 1.3 to 3.5 million atmospheres, the latter being reached some 5200 km below the surface. At this depth, despite temperatures over 5000 °C, the pressure forces the crystallisation of a solid inner core.

At such high temperatures, permanent magnetism is not possible because of thermal agitation. This phenomenon was studied by Pierre Curie during his PhD thesis in 1895. The critical temperature at which permanent magnetism is lost is now known as the 'Curie temperature'. It is of the order of 800 °C for iron. Permanent magnetism in the core of the Earth can thus be ruled out. Only cold material, close to the surface can be magnetised (this is known as the crustal field, which is rather weak and small-scale).

Yet the Earth exhibits a large-scale magnetic field, known as the main field, which can only be of internal origin. The rocky mantle being—to a good approximation—insulating, the magnetic field, as measured at the surface, can be downward continued to the core-mantle boundary using a harmonic potential. Since the core is free of permanent magnetism, the magnetic field we reconstruct at its surface is the signature of electrical currents crossing the core. The key open question lies in the generation mechanism for these electrical currents.

The electrical currents, and thus the internal magnetic field, are most likely due to a magnetohydrodynamic instability in the liquid core known as 'self-exciting dynamo action'. This instability is one of the rare examples of a mechanism acting at a scale which is neither very small (say nanophysics), nor very large (say cosmology), and which still lacks a proper physical understanding. How does a conducting fluid spontaneously transfer part of its kinetic energy into electrical energy?

The origin of the kinetic energy in the core is rather well understood. Since its formation, some 4.5 billion years ago, our planet has continuously been cooling down, by evacuating the heat it contains to the surface of the planet where temperatures are lower. In this process, heat is partly carried from the inner depths by thermal convection, that is to say by the transport of buoyant hot material toward the surface and of cold material to lower depths. In the mantle, this convective motion is rather slow. In the core, which is characterised by a kinematic viscosity comparable to that of water at ambient temperature and pressure, it is on the contrary very fast (again in a geological sense!). The flow is estimated to be of the order of a meter per hour. This may not seem terribly large at first, but it is almost a million times faster than the velocities of tectonic plates at the surface!

Besides thermal convection, a second source of energy is available to drive motions in the liquid core as the Earth is cooling down. High-pressure physics and geochemistry allowed to estimate that the liquid core of our planet is not made of pure iron. Instead, it contains approximately 80% in mass of iron, 5% of nickel (with a density comparable to iron), and of 15% 'light elements' including possibly silicon, sulphur, and oxygen. The density of the liquid core is lower by approximately 10% than that of pure iron under the same physical conditions. The solid inner core solidifies and grows as the planet is slowly cooling down. In this process, which is not at the eutectic, only pure iron solidifies and the lighter elements are released at the base of the liquid core. They will rise and mix through buoyancy in a way similar to thermal convection, known as ‘solutal convection’.

We know that an electric current appears by induction within a conductor moving in the presence of a magnetic field. We also know that an electric current generates a magnetic field. Sir Joseph Larmor, then Lucasian Professor of mathematics in Cambridge, proposed a little bit over a century ago a mechanism by which these two phenomena could be combined to yield an instability. This is to say that, if the conditions are met, the motion of the liquid metal in the presence of electrical currents in the core could act to regenerate these very currents against ohmic losses.

However, these ‘conditions’ are yet to be understood (we refer the reader to the book by Moffatt and Dormy [283] for a more advanced discussion). There exist, of course, theoretical bounds such as lower bounds on the strength of the flow (as measured by the magnetic Reynolds number) in order to be able to sustain the magnetic field. There are also theoretical results ruling out field amplification for too simple geometries of the flow (e.g., Zeldovich’s theorem, ruling out dynamo action from two-dimensional flows, and Cowling’s theorem from axisymmetric flows). But the detailed conditions for dynamo action are still challenging.

This is made even more complicated by the rotation of the Earth (which is remarkably fast: one spin a day!). The effect of the Coriolis force on the flow is determinant, as highlighted by the fact the axis of the large-scale field is nearly aligned with the axis of rotation of the planet (a decisive property for the use of a compass!). The day is a very short timescale compared, for example, to the time needed for the weak viscosity of the fluid to slow down the motion. It follows that very sharp boundary layers will form at the boundaries of the liquid core. These can be estimated to be about a meter thick, to be compared with the 3400 km radius of the core. Despite the complexity of the non-linear equations describing the induction mechanisms in the core, some theoretical advances were made possible thanks to the disparity of scales involved in the process.

Back in 1955, Eugene Parker introduced the idea that small-scale motions, influenced by rotation, will be of a helical nature (a property of rapidly rotating flows) and that ‘cyclonic events’ could be incorporated by an averaging procedure in equations for the mean magnetic field. But it was only some ten years later that the mathematical formalism was properly derived independently by Braginskii (1964) [284] and by Steenbeck, Krause, and Rädler (1966) [285].

The essential idea is that an average can then be taken to estimate the large-scale field resulting from small helicoidal vortices. It is found to be non-vanishing. In this two-scale analytical approach, some randomness is assumed on the small-scale flow, and the large-scale field (known as ‘mean field’) is constructed on the basis of this assumption. Braginskii considered the axisymmetric field as ‘large-scale’ and non-axisymmetric fluctuations as small scales.

The magnetic Reynolds number relevant to this scale separation can be shown to involve the square root of the product of the large scale and the small scale (i.e., the geometric mean of the two length-scales) [283]. Because of this, mean-field effects are possible down to a very small scale. In the liquid core of the Earth, the flow down to a scale of a few 100 m can contribute to the large-scale magnetic field generation. Such mean-field models usually assume either no-flow at the large-scale, or more simply a spectral gap between the small-scale dynamo vortices and the large-scale flow and field. This formalism has been used for decades, and was until recently the only way to tackle this challenging problem. The full non-linear set of equations not being prone to a complete analytical treatment.

Since 1995, direct numerical simulation has taken the lead. The full set of equations can be successfully discretised on a computer (usually computers working in parallel). The outcome is spectacular magnetic field lines rendering in three-dimensional visualisation (see figure 2.46), Earth-like magnetic fields, and occasionally reversals of the field polarity. Yet, it should be stressed that all of these models have to rely on non-physical filtering of the small scales (either through unrealistic parameters, or through additional terms in the equations). This does not come as a surprise when one realises the formidable disparity of scales involved (the size of the core being, for example, about a million times that of the boundary layers we

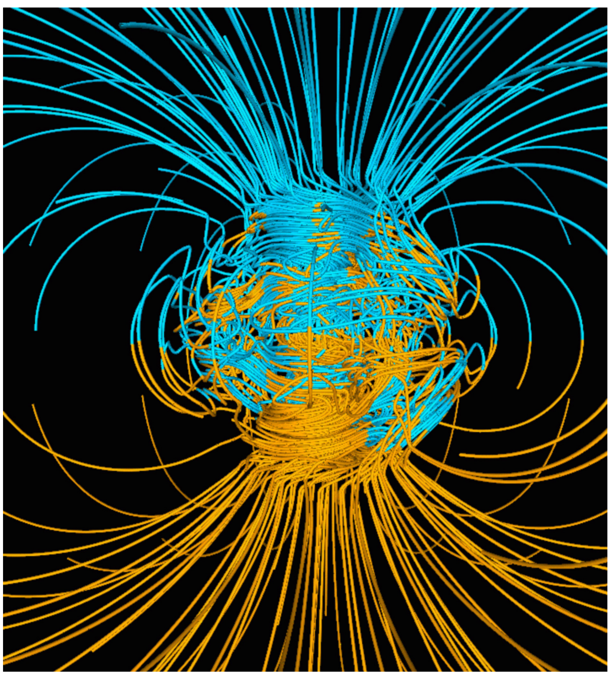


Figure 2.46. The celebrated Glatzmaier–Roberts model for the geodynamo. One of the first numerical models introduced in 1995. Computers and numerical models have improved since, but not to the point of closing the formidable gap between the smallest scales of the flow relevant to magnetic field generation and the size of the core. Reprinted from [286], Copyright (1995), with permission from Elsevier.

mentioned above). A proper understanding of the various scales at play and their interaction is sadly still missing despite the use of some of the largest computers currently available.

It is not at all easy to perform experiments of self-exciting dynamo action. Only three experiments have been successful so far. The first two in Karlsruhe and Riga involving severely constrained flows. The third one in Cadarache, allowing for an unconstrained turbulent flow, has highlighted Earth-like polarity reversals when the symmetry between the forcing blades was broken [287]. However, it needed ferromagnetic blades to operate.

It is then interesting to consider other planets to try and gain an insight from their magnetic fields. Some planetary magnetic fields are similar, such as Jupiter and the Earth (despite the fact Jupiter is a gaseous planet, with a very different inner structure). In other cases they are very different. Uranus and Neptune, for example, both exhibit an asymmetric multipolar field, whereas Saturn presents a nearly axisymmetric field. There is thus no simple picture to be drawn from such comparisons. The latter geometry raising interesting questions, since Cowling's theorem prevents dynamo action to sustain a purely axisymmetric field.

2.8.3 The physics of volcanic eruptions

Let us now turn our attention to convection in the mantle and some of its interactions with us: volcanoes. Contrary to our first problem, which was purely academic and not directly related to society concerns (see [288] for a discussion), this problem addresses a life-threatening hazard. We cannot discuss the physics of volcanoes without asking the question: to what extent can volcanic eruptions be predicted?

Volcanoes are one of the signatures at the Earth's surface of the deep-seated thermal convection in the Earth's mantle. We have already mentioned three 'layers' in the Earth's interior: the inner core, the liquid outer core, and the mantle. We should also introduce the thin shell at the surface: the Earth's crust, with thickness ranging from some 5–10 km under the oceans and 30–50 km under continents. Together with a thin portion of the upper mantle, it forms the Lithosphere (i.e., the 'rigid' outer layer of the Earth). Because of the underlying thermal convection in the mantle, it is broken into tectonic plates (the physical laws governing the sizes of these plates also constitute a challenge in geophysics).

A volcano is formed of one or more magma reservoirs that are connected to the surface by small pipes that contain a liquid, the magma, formed of molten silicate. The magma is produced by a deep source, located in the Earth's mantle (remember the core is formed of liquid metal, not magma!), or near the surface in the crust. Some volcanoes, known as 'hot spots', originate from very deep in the mantle (probably plumes of melted material emerging from the base of the mantle), while others are located at plates boundary and are of shallower origin.

The volume of a volcanic eruption is much greater than can be sustained by the flow through the deep plumbing system that brings the magma to the surface. A volcanic eruption is therefore only possible when a sufficient quantity of magma has accumulated in a reservoir, the so-called magma chamber. These reservoirs are formed because of the migration of magma to the surface, due to buoyancy. This migration stops in the vicinity of the surface, when the magma reaches rocks of a lower density. The magma then stops rising and starts forming the reservoir. The magma chamber does more than just storing the magma. During the period of stay in the chamber, the magma changes properties, in terms of crystallinity, viscosity, dissolved water content as the magma cools and partly solidifies. Crystallization is fractional, and the residual liquid chemically evolves and is enriched in volatile elements, such as carbon dioxide, sulphur dioxide, or hydrogen sulfide. These changes in the magma are essential, because they modify its physical properties and thus determine the type of the next eruption. The content in gazes of course depends on the volcano. In general, volcanoes associated with convergent plate boundaries tend to emit more water vapour than those at hot spots or at divergent plate boundaries (because of the seawater trapped in the subduction).

An eruption occurs when the magma chamber fails. This occurs when the overpressure in the chamber exceeds the strength of the rocks surrounding it. This is, of course, very hard to predict, if only because it is not possible to know the exact level of stress within a volcano.

When an eruption occurs and the magma moves up to the surface, the flow of magma is characterised by a complex rheology. Besides, the pressure drop causes both exsolution (i.e., the separation of the various constituents previously dissolved in a homogeneous phase) of volatiles and expansion of the gas phase present.

It turns out that the nature of the magma is essential in explaining the various observed eruptive behaviours. A given volcano, and even a given eruption, can evolve between different eruptive regimes. Some being far more dangerous than others and the changes being very difficult to forecast. The quantity of gas in the mixture, of course, depends on the initial content of the magma in the reservoir, but also on that quantity of gas lost during the ascent through the permeable walls and fractured rocks. The total content of volatile elements thus depends on the rising velocity. If the magma rises very quickly, gas losses can safely be neglected. If, on the contrary, the rising velocity is low, these losses can become very significant. Since the rate of ascent itself depends on the gas fraction, the non-linear coupling induces a strong sensitivity to small changes in the initial conditions in the reservoir.

The regimes can be classified in increasing proportion of gas. If the gas content is low, the magma contains a few gas bubbles in suspension. This is the regime corresponding to effusive lava flows. At a higher gas fraction, an intermittent rate is obtained, where the gas is collected in pockets whose diameter are close to that of the volcano conduit. With even more gas, the gas forms a continuous jet and the walls of the conduit are then covered with a thin layer of magma. The eruption then consists of high-velocity turbulent jets carrying fine droplets of magma. This is the so-called explosive regime. In this explosive regime, the volcanic jet can follow remarkably different dynamical evolutions depending on small variations in its mass discharge rate and in the amount of free gas in the eruptive mixture.

In the so-called ‘Plinian’ eruptions, the jet is buoyant enough for the eruptive columns to go as high as the stratosphere, where it spreads horizontally. However, if the turbulent jet is not light enough, it can collapse. This is due to the nature of the jet, but also to the amount of air being entrapped as the rising panache expands [289, 290]. If the jet collapses, it falls back on the slope of the volcano at formidable speeds. This is known as a pyroclastic flow. The dynamics of the flow is similar in many aspects (including its speed of some 100 km h^{-1}) to an avalanche, but its temperature is about 1000 °C. Pyroclastic flows are the most terrible of all volcanic hazards. When the conditions are close to those of collapse, the eruption can be transitional: a buoyant column can rise to high altitudes until it suddenly collapses (see figure 2.47).

Pyroclastic flows are notoriously life threatening. Famous examples of catastrophic events in the past include the terrible eruption of Mount Vesuvius in 79 AD which destroyed, among others, the city of Pompeii and caused a huge number of deaths (including the famous Pliny the Elder), or that of Mount Unzen, in Japan on June 3, 1991 which caused the death of two well-trained volcanologists, Katia and Maurice Krafft.

It may be surprising at first to note how the volatile content and the micro-bubbles forming in the magma will have such decisive large-scale consequences. Minute variations in the pressure of the reservoir can induce variations in velocity, pressure, and gas content, and shift from an explosive regime to an effusive regime. Initially, small bubbles are not buoyant enough to rise quickly through the viscous magma. However, as the magma ascends, the bubbles will grow, both because of expansion as the pressure drops, and because the solubility of volatiles in the magma decreases causing more gas to exsolve. This results in a complex rheology which prevents classical solutions of equations of fluid mechanics to describe the flow of magma. The presence of bubbles will drive a faster upward-propagation. What happens next largely depends on the gas concentration and on the viscosity of the magma. The bubbles may rise and merge to form bigger bubbles, or they may be stuck in the magma and slowly connect to form some sort of network. The

Figure 2.47. A Plinian column (on the left) characterised by a buoyant jet extending to high altitudes, and a collapsing fountain producing a devastating pyroclastic flow (on the right) which is not buoyant enough to rise high in the atmosphere and collapses on the flanks of the volcano.

overpressure in the conduit caused by these gases is directly related to the explosivity of an eruption [291].

Volcanoes do not only occur on Earth. The surface of Venus is covered by many volcanic structures. They are probably not active, though a recent study indicates that some of them may have been active during the past few million years [292]. The most active extra-terrestrial volcanoes were reported on moons, namely on Io (Jupiter's moon), Triton (Neptune's moon), and Enceladus (Saturn's moon). Io was the first moon for which evidence of active volcanism on a body of the Solar System other than Earth was obtained, some 40 years ago. It is now regarded as the most volcanically active body in the Solar System.

2.8.4 Earthquakes

Let us now turn to yet another consequence of mantle convection at the surface of our planet: earthquakes. Earthquakes are, without a doubt, among the most devastating natural phenomena on Earth. They are also intimately related to our knowledge of the interior of our planet. Seismology has allowed the use of earthquakes as an indirect means of observing the Earth's inner depth (we would not know of the internal structure of the Earth introduced above, if it was not for seismology).

Earthquakes are a direct consequence of plate tectonics. Tectonics is driven by thermal convection in the mantle and by the fact the colder and thus heavier plates want to sink. Gravity pulls the slab downward in the form of cold downwelling. Seismic tomography reveals that some slabs appear to stop sinking and are horizontally deflected at a depth of some 660 km (i.e., the transition between the upper-mantle and the lower-mantle, corresponding to a mineral phase transition, that of Olivine). Other slabs are sinking further down in the mantle.

Earthquakes occur at plate boundaries, where the plates converge, diverge, or slide along each other. Brittle deformation causes the formation of faults in the lithosphere. Earthquakes are caused by a sudden slip on an active fault: tectonic plates are slowly moving, but they tend to 'stick' at their edges due to friction with neighbouring plates. Stress builds up and at some stage becomes strong enough to overcome friction. An earthquake then occurs, releasing a large amount of energy in the form of waves.

While predicting earthquakes remains exceedingly challenging, our physical understanding of these phenomena has significantly progressed since the understanding, over a century ago, that they result from the brutal sliding of rocks on faults, which are regions of weakness of the Earth's crust. Fault motion can be viewed as frictional sliding on a fault plane. From a mechanical point of view, friction will evolve as a result of slip (i.e., the relative displacement of the two sides of the fault plane), the velocity, and the history of contact.

The very nature of the trigger makes it extremely difficult to predict the likely ground motions during future earthquakes. It is currently impossible to predict their time, location, and magnitude. Observations of precursor phenomena are numerous but involve a significant risk of false alarms. This should be contrasted with the good efficiency of early warning systems, which rely on the finite propagation speed of seismic waves to provide real-time assessment of the hazard, and notification of people in the endangered area. Fatalities in many parts of the world have also been reduced through the development of building codes designed to withstand strong earthquakes.

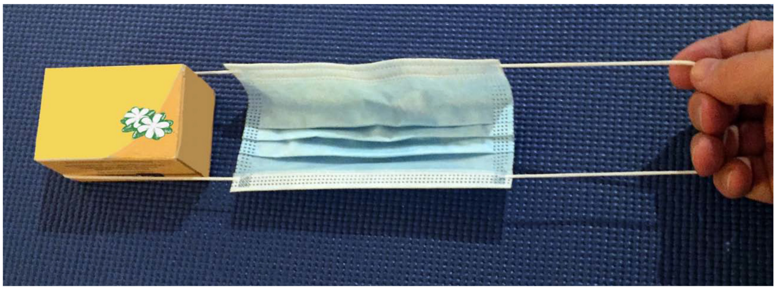
To understand the formation of earthquakes, we must think of the tectonic stress on a fault, slowly building up, until it reaches the critical stress necessary for failure. The frictional stress is essential as it controls the seismic motion. An earthquake can occur only if friction decreases rapidly with slip; in other words, if the 'dynamic friction' is much less than the 'static friction', a process often referred to as 'slip weakening'. In that case, the two walls of an earthquake fault are either sticking to each other or sliding over each other. They can be thought as rough surfaces with irregular contact points. The resulting sliding motion then takes a form of relaxation oscillations, common in non-linear physics, and known in this case as 'stick-slip'.

This is probably best understood using a physical analog: the slider-block model. In this model, originally introduced by Burridge and Knopoff [293] and often used to describe earthquake dynamics, a block is attached to a spring, which is pulled over a surface with a constant force. The friction between the block and the surface prevents the block from sliding. The area of the frictional surface in contact with the block is then analogous to the fault. As the spring is pulled, tension slowly builds up in the spring. At some stage the tension will be such that the critical stress is reached and the block will slip. On most surfaces, the block will simply gently slide at constant velocity. However, if we meet a velocity-weakening of the friction force, that is to say a decrease of friction with velocity, a 'stick-slip' behaviour will occur. The block will slide for a distance and then stop again as the tension in the spring has dropped too much (in the same way, after an earthquake occurs, it yields a sudden stress drop). The tension on the spring will then build again until the next event. A new 'earthquake' cycle subsequently begins.

This experiment is easily performed using a standard soap in a paper box as 'sliding block' and a yoga mat or any rubber mat as supporting surface. The spring can conveniently be replaced by one of these surgical masks that have sadly filled our lives in the recent years. The experiment is illustrated in figure 2.48(a).

More realistically, small irregularities at the surface will involve local variations of the critical stress, loading rate, and stress drop. So that the stick-slip cycle will not be periodic. Burridge and Knopoff went beyond this and introduced chains of such sliding blocks, all coupled to nearest neighbours by elastic springs (see figure 2.48(b)) and then a two-dimensional network of such blocks interconnected with springs. They performed both laboratory experiments and numerical simulations. Remarkably, their model accounts for the redistribution of stresses associated with a sliding event. A sequence of event occurs, much like 'aftershocks' in real earthquakes. This simplified physical system also produces a power law probability density function for event size, in agreement with observations of real earthquakes known as the Gutenberg-Richter law [293, 294]. Such confrontation of physical models to observations is an essential part in the understanding of geophysical processes [295].

Burridge and Knopoff's model is now over 50 years old. It cannot be considered as new! Yet, it highlights very clearly one of the key challenges in understanding

a.


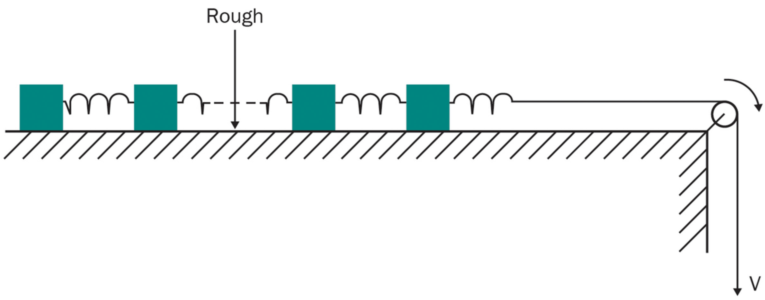
b.


Figure 2.48. A sliding block model for earthquake, highlighting the importance of small-scale contacts in large-scale events. Realised with a single block and low-cost equipment (a) and in the Burridge and Knopoff (1967) one-dimensional chain model (b).

earthquakes: the very complex mechanisms at their source and the importance of small-scale heterogeneities in the rupture process. In real earthquakes, ruptures in a fault with heterogeneous stress and rupture resistance distributions are very complex. The rupture front itself is not given. It can change directions because of small-scale heterogeneities. It will follow regions of strong stress concentration and avoid those with lower stress or higher rupture resistance. This dynamics will essentially depend on small-scale heterogeneities.

This complex behaviour of the fault network, governed by small-scale heterogeneities, hinders an accurate *a priori* estimate of the energy being released. Trying to forecast when a propagating rupture will or will not stop appears nearly impossible. The dynamics are controlled by terribly small-scale irregularities in the fault. In order to be able to estimate potential ground motions that may occur due to rupture on an active fault, one would need a remarkably precise description of the state of the rocks, both in terms of stress and resistance. The resulting complex rupture and earthquake clustering appear essential to achieving a better understanding of earthquakes [296].

Earthquakes can be viewed as terrible 'large-scale events' that appear to depend crucially on the dynamics of very small-scales heterogeneities within the rocks.

'Quakes' are not a specificity of the Earth. They were also recorded on our Moon some 50 years ago by Apollo's seismometers. These are known as 'moonquakes'. On Mars, the first 'marsquake' was recorded by the Insight mission in 2019, thus revealing the inner structure of the planet.

2.8.5 Challenges and opportunities

Only three of the many open questions in Earth and planetary physics have been highlighted here. This, of course, corresponds to an arbitrary selection. Many other issues offer a challenge to researchers. These include but are not limited to: the formation of tectonic plates at the surface of the Earth from mantle convection and the parameters controlling their characteristic sizes; why Earth exhibits plate tectonics but not other planets; the necessary ingredients for magnetic polarity reversals; the formation of banded structures in giant gaseous planets and what sets the depth of the associated flow; how the strong zonal winds in these bands might interact with flow in the electrically conducting interior and affect the magnetic field of those planets. These are only examples of the wide variety of open questions which offer a venue for physical investigation.

I thought nonetheless that these three questions had some merits to be put forward. They highlight both the wide variety of open questions in Earth sciences, ranging from fundamental physical processes (such as the dynamo instability) to phenomena of direct societal concern (such as volcanic hazard and earthquakes) and the impressive variety of scales involved in many geophysical processes.

Those three problems have more in common than may appear at first sight. They all belong the the broad field of non-linear physics. In all three problems, the basic physics is, to a large extent, reasonably well understood. It has been theoretically described, sometimes more than a century ago. The large-scale dynamics, however, is highly non-linear and hardly predictable.

The large-scale non-linear physics can often be efficiently tackled through high-performance numerical simulations. These have made impressive progresses over the last decades with the development of high-performance computing. Increasingly complex physics can be accounted for, with ever more realism and higher resolutions. Yet the problems associated with extreme ranges of scale remain challenging.

In many cases, current numerical simulations will have to rely on some parameterisation of the small scales. Small-scale dynamics, out of reach of current simulations, have to be either ignored or parameterised using ad hoc simplifying assumptions. Such a separation of scale is a very common approach in geophysics, including external geophysics. For example, the microphysics of droplet formation is essential to large-scale atmospheric phenomena such as tropical cyclones and has to be parameterised.

In view of the rapidly increasing computational power, it may be tempting to simply sit and wait for faster computers and increased numerical resolution to close this gap. Yet the disparity of scale is such, in many geophysical problems, that the wait could still be long. Some new physical insight is more likely to yield progress in the years to come. In many cases, numerical work does not outrun theory, but rather complements it or sometimes provides new ideas to be tested theoretically.

It should be stressed that the importance of small-scale physics goes here far beyond the theory of deterministic chaos. It is well established that a tiny perturbation of a non-linear system can yield a drastic change in behaviour. In the above examples, and in many others, however, the small scales are an intrinsic part of the physical mechanism. They cannot be viewed as some small-scale perturbations of a large-scale dynamics. Instead, they are an essential part in the dynamics itself.

The reason for which these open problems involve such a wide variety of scales is probably that, had it not been the case, they would have been solved long ago! Such an extreme range of scale is of course not restricted to geophysics, but it is surprisingly common in Earth sciences.

When facing such a wide range of scale, a physicist will usually rely on separation of scale to handle it. This is the usual theoretical formalism to apprehend such a wide range of phenomena, and it is a very powerful tool. Yet, it usually assumes a 'spectral gap' between the small scales and the large scales or in other words an irrelevance of the intermediate scales. Sadly this approximation is often not justified.

Rather than a simple coupling of the small- and large-scale physics, some new developments in the years to come are probably needed to offer a better description of physics of the 'intermediate scale'. Beyond the standard separation of scale, it becomes increasingly evident that some physical problems are associated with a huge variety of scales, all relevant to understanding the large-scale behaviour. The physics of the 'intermediate-scale' is of course challenging as it necessarily involve a combination of the non-linearities typical of large-scales and of the complex small-scale physics.

Acknowledgments

I would like to thank a few colleagues and friends whom have shared with me, over the years, their understanding of some of the problems mentioned above. In alphabetical order: Claude Jaupart, Keith Moffatt, Dmitri Pissarenko, Andrew Soward, and Albert Tarantola.

References

[1] Director: Mark Levinson 2013 Particle Fever (Documentary) (http://particlefever.com/)
[2] Kajita T and McDonald A B 2015 The Nobel Prize in physics was awarded for the discovery of neutrino oscillations, which shows that neutrinos have mass [online]
[3] Bednyakov A V, Kniehl B A, Pikelner A F and Veretin O L 2015 Stability of the electroweak vacuum: gauge independence and advanced precision *Phys. Rev. Lett.* **115** 201802
[4] Grange J *et al* 2015 Muon (G-2) technical design report 1 arXiv:1501.06858
[5] Rizzo T 2012 The BSM Zoo *Technical Report* (https://www-conf.slac.stanford.edu/ssi/2012/lectures.asp).
[6] Alimena J *et al* 2020 Searching for long-lived particles beyond the standard model at the large hadron collider *J. Phys. G: Nucl. Part. Phys.* **47** 090501
[7] *LHCb Trigger and Online Upgrade Technical Design Report* 5 2014
[8] Data Parking and Data Scouting at the CMS Experiment 9 2012
[9] Yoo J H 2019 The milliQan experiment: search for milli-charged particles at the LHC *PoS* ICHEP2018:520
[10] Lubatti H *et al* 2020 Explore the lifetime frontier with MATHUSLA *JINST* **15** C06026

[11] Ariga A *et al* 2019 FASER: ForwArd Search ExpeRiment at the LHC. 1 arXiv:1901.04468
[12] The LHCf CollaborationAdriani O *et al* 2008 The LHCf detector at the CERN Large Hadron Collider *J. Instrum.* **3** S08006–6
[13] The TOTEM CollaborationAnelli G *et al* 2008 The TOTEM experiment at the CERN large hadron collider *J. Instrum.* **3** S08007–7
[14] Pinfold J *et al* 2009 Technical design report of the MoEDAL experiment. 6
[15] Akerib D S *et al* 2015 LUX-ZEPLIN (LZ) conceptual design report. 9 arXiv:1509.02910
[16] Aprile E *et al* 2005 The XENON dark matter search experiment *New Astron. Rev.* **49** 289–95
[17] European Strategy Group 2020 Update of the European strategy for particle physics *Technical Report, Geneva* 2020 (https://cds.cern.ch/record/2720129)
[18] *The NuPECC Long Range Plan 2017* (http://nupecc.org/pub/lrp17/lrp2017.pdf)
[19] *The NuPECC Long Range Plan Brochure* (https://nupecc.org/pub/lrp17/nupecc_lrp_brochure_2017.pdf)
[20] *The 2015 Long Range Plan for Nuclear Science* (https://science.osti.gov/~/media/np/nsac/pdf/2015LRP/2015_LRPNS_091815.pdf)
[21] Durante M *et al* 2019 All the fun of the FAIR: fundamental physics at the facility for antiproton and ion research *Phys. Scr.* **94** 033001
[22] *JINR Long-Term Development Strategy up to 2030 and Beyond* (http://jinr.ru/posts/jinr-long-term-development-strategy-up-to-2030-and-beyond/)
[23] Gales S *et al* 2018 The extreme light infrastructure—nuclear physics (ELI-NP) facility: new horizons in physics with 10 PW ultra-intense lasers and 20 MeV brilliant gamma beams *Rep. Prog. Phys.* **81** 094301
[24] Savalle A *et al* 2018 SPIRAL2: une sonde de nouvelle génération pour explorer la matière nucléaire *Reflets Phys.* **59** 11
[25] Jokinen A 2014 The Jyväskylä Accelerator Laboratory *Nucl. Phys. News* **24** 4
[26] De Angelis G *et al* 2016 The SPES radioactive ion beam project of LNL: status and perspectives *EPJ Web Conf.* **107** 01001
[27] Calabretta L *et al* 2017 Overview of the future upgrade of the INFN-LNS superconducting cyclotron *Mod. Phys. Lett.* A **32** 1740009 (Special issue on cyclotrons and their applications)
[28] Colonna N *et al* 2018 Neutron physics with accelerators *Prog. Part. Nucl. Phys.* **101** 177
[29] Ashford M *et al* 2020 Exploratory study for the production of Sc beams at the ISOL facility of MYRRHA preliminary thermal investigations *Nucl. Instrum. Methods Phys. Res.* B **463** 244
[30] Warren S *et al* 2020 Offline 2, ISOLDE's target, laser and beams development facility *Nucl. Instrum. Methods Phys. Res.* B **463** 115
[31] Kuchi V *et al* 2020 High efficiency ISOL system to produce neutron deficient short-lived alkali RIBs on GANIL/SPIRAL 1 facility *Nucl. Instrum. Methods Phys. Res.* B **463** 163
[32] Minaya E *et al* 2020 New program for measuring masses of silver isotopes near the $N = 82$ shell closure with MLLTRAP at ALTO *Nucl. Instrum. Methods Phys. Res.* B **463** 315
[33] Kaminki G *et al* 2020 Status of the new fragment separator ACCULINA-2 and first experiments *Nucl. Instrum. Methods Phys. Res.* B **463** 504
[34] Berger N *et al* 2021 The MESA experimental program: a laboratory for precision physics with electron scattering at low energy *Nucl. Phys. News* **31** 5

[35] Facility for Rare Isotope Beams 2021 Rare isotopes aplenty at FRIB *CERN Cour.* https://frib.msu.edu/

[36] Shimizu Y *et al* 2020 Database of radioactive isotope beams produced at the BigRIPS separator *Nucl. Instrum. Methods Phys. Res.* B **463** 158

[37] Marchetto M *et al* 2020 Status of the CANREB high-resolution separator at TRIUMF *Nucl. Instrum. Methods Phys. Res.* B **463** 227

[38] Kim Y J 2020 Current status of experimental facilities at RAON *Nucl. Instrum. Methods Phys. Res.* B **463** 408

[39] Burkert V D 2018 Jefferson Lab at 12 GeV: the science program *Annu. Rev. Nucl. Part. Sci.* **68** 405–28

[40] Ohnishi H *et al* 2020 Hadron Physics at J-Park *Prog. Part. Nucl. Phys.* **113** 103773

[41] Drischl C *et al* 2021 Towards grounding nuclear physics in QCD *Prog. Part. Nucl. Phys.* **121** 103888

[42] Arbuzova A *et al* 2020 On the physics potential to study the gluon content of proton and deuteron at NICA SPD *Prog. Part. Nucl. Phys.* **113** 103858

[43] Peset C 2021 The proton radius (puzzle?) and its relatives *Prog. Part. Nucl. Phys.* **121** 103901

[44] Ireland D G *et al* 2020 Photoproduction reactions and non-strange baryon spectroscopy *Prog. Part. Nucl. Phys.* **111** 103752

[45] Constantinou M *et al* 2021 Parton distributions and lattice-QCD calculations: toward 3D structure *Prog. Part. Nucl. Phys.* **121** 103908

[46] Bartmann W *et al* 2018 The ELENA facility *Philos. Trans. A: Math. Phys. Eng. Sci.* **376** 20170266

[47] Tino G M *et al* 2020 Precision gravity tests and the Einstein equivalence principle *Prog. Part. Nucl. Phys.* **112** 103762

[48] Dong X and Greco V 2019 Heavy quark production and properties of Quark–Gluon plasma *Prog. Part. Nucl. Phys.* **104** 97 and the ALICE Publications https://alice-publications.web.cern.ch/publications

[49] ALICE Collaboration 2020 Unveiling the strong interaction among hadrons at the LHC *Nature* **588** 232–8

[50] Chapon E *et al* 2022 Prospects for quarkonium studies at the high-luminosity LHC *Prog. Part. Nucl. Phys.* **122** 103906

[51] Oganessian Y T *et al* 2017 Superheavy nuclei: from predictions to discovery *Phys. Scr.* **92** 023003

[52] Wilson J N *et al* 2021 Angular momentum generation in nuclear fission *Nature* **590** 566–70

[53] Giovinazzo J *et al* 2021 4D-imaging of drip-line radioactivity by detecting proton emission from Ni pictured with ACTAR TPC *Nat. Commun.* **12** 4805

[54] Mougeot M *et al* 2021 Mass measurements of 99–101In challenge *ab initio* nuclear theory of the nuclide 100Sn *Nat. Phys.* **17** 1099

[55] Leimbach D *et al* 2020 The electron affinity of astatine *Nat. Commun.* **11** 3824

[56] Burgio G F *et al* 2021 Neutron stars and the nuclear equation of state**120** 103879

[57] Arnould M and Goriely S 2020 Astronuclear physics: a tale of the atomic nuclei in the skies *Prog. Part. Nucl. Phys.* **112** 103766

[58] Mossa V *et al* 2020 The baryon density of the Universe from an improved rate of deuterium burning *Nature* **587** 210–3

[59] The Borexino Collaboration 2020 Experimental evidence of neutrinos produced in the CNO fusion cycle in the Sun *Nature* **587** 577–82

[60] Korten W *et al* 2020 Physics opportunities with the advanced gamma tracking array: AGATA *Eur. Phys. J.* A **56** 137

[61] Bracco A *et al* 2021 Gamma spectroscopy with AGATA in its first phases: new insights in nuclear excitations along the nuclear chart *Prog. Part. Nucl. Phys.* **121** 103887 and AGATA Publication list: https://agata.org/

[62] Fallon P *et al* 2016 GRETINA and its early science *Rev. Nucl. Part. Sci.* **66** 321 and GRETINA Publication list; http://gretina.lbl.gov/publications

[63] Milner R G 2021 The electron ion collider: the 21st-century electron microscope for the study of the fundamental structure of matter *MIT Physics Annual 2020 and Nuclear Physics News*

[64] EicC working group 2021 EicC white paper *Front. Phys.* **16** 64701

[65] Kiefer C 2012 *Quantum Gravity* (Oxford: Oxford University Press) 3rd edn

[66] Schlosshauer M 2007 *Decoherence and the Quantum-to-Classical Transition* (Berlin: Springer)

[67] Joos E, Zeh H D, Kiefer C, Giulini D, Kupsch J and Stamatescu I-O 2003 *Decoherence and the Appearance of a Classical World in Quantum Theory* 2nd edn (Berlin: Springer)

[68] DeWitt C 1957 *Proc. of the Conf. on the Role of Gravitation in Physics (University of North Carolina, Chapel Hill, January 18–23, 1957)* WADC Technical Report 57–216 (unpublished). These proceedings have recently been edited inD Rickles and C M DeWitt Edition Open Sources, http://www.edition-open-sources.org/sources/5/

[69] Bose S, Mazumdar A, Morley G W, Ulricht H and Toroš M 2017 Spin entanglement witness for quantum gravity *Phys. Rev. Lett.* **119** 240401 6

[70] Marletto C and Vedral V 2017 Witness gravity's quantum side in the lab *Nature* **547** 156–8 see also the Correspondence in *Nature* 549

[71] Albers M, Kiefer C and Reginatto M 2008 Measurement analysis and quantum gravity *Phys. Rev.* D **78** 064051 17

[72] Hehl F W and Lämmerzahl C 2019 Physical dimensions/units and universal constants: their invariance in special and general relativity *Ann. Phys. (Berlin)* **531** 1800407 10

[73] Carlip S 2001 Quantum gravity: a progress report *Rep. Prog. Phys.* **64** 885–942

[74] Woodard R P 2009 How far are we from the quantum theory of gravity? *Rep. Prog. Phys.* **72** 126002 42

[75] Kiefer C 2013 Conceptual problems in quantum gravity and quantum cosmology *ISRN Math. Phys.* **2013** 509316

[76] Zeh H D 2007 *The Physical Basis of the Direction of Time* 5th edn (Berlin: Springer)

[77] Oriti D (ed) 2009 *Approaches to Quantum Gravity* (Cambridge: Cambridge University Press)

[78] Carr B (ed) 2007 *Universe or Multiverse?* (Cambridge: Cambridge University Press)

[79] Penrose R 2007 *The Road to Reality: A Complete Guide to the Laws of the Universe* (New York: Vintage)

[80] Page D N 2013 Time dependence of Hawking radiation entropy *J. Cosmol. Astropart. Phys.* **9** 28 28

[81] Feynman R 1990 *The Character of Physical Law* (Cambridge, MA: MIT Press)

[82] Carlip S 2019 *General Relativity: A Concise Introduction* (Oxford: Oxford University Press)

[83] Bertone G and Hooper D 2018 History of dark matter *Rev. Mod. Phys.* **90** 045002

[84] Zyla P A *et al* 2020 Review of particle physics *Prog. Theor. Exp. Phys.* **2020** 083C01
[85] Zwicky F 1933 Die Rotverschiebung von extragalaktischen Nebeln *Helv. Phys. Acta* **6** 110 127
[86] Milgrom M 1983 A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis *Astrophys. J.* **270** 365 370
[87] Einstein A 1916 The foundation of the general theory of relativity *Ann. Phys.* **49** 769–822
[88] Gaillard M K, Grannis P D and Sciulli F J 1999 The standard model of particle physics *Rev. Mod. Phys.* **71** S96–S111
[89] Blumenthal G R, Faber S M, Primack J R and Rees M J 1984 Formation of galaxies and large scale structure with cold dark matter *Nature* **311** 517–25
[90] Audren B, Lesgourgues J, Mangano G, Serpico P D and Tram T 2014 Strongest model-independent bound on the lifetime of dark matter *J. Cosmol. Astropart. Phys.* **12** 028
[91] Rubin V C and Ford W K 1970 Rotation of the Andromeda nebula from a spectroscopic survey of emission regions *Astrophys. J.* **159** 379–403
[92] White M J, Scott D and Silk J 1994 Anisotropies in the cosmic microwave background *Ann. Rev. Astron. Astrophys.* **32** 319–70
[93] Wayne Hu and Dodelson S 2002 Cosmic microwave background anisotropies *Ann. Rev. Astron. Astrophys.* **40** 171–216
[94] Ade P A R *et al* 2016 Planck 2015 results. XIII. Cosmological parameters *Astron. Astrophys.* **594** A13
[95] Lifshitz E 1946 Republication of: on the gravitational stability of the expanding universe *J. Phys. (USSR)* **10** 116
[96] Vogelsberger M, Marinacci F, Torrey P and Puchwein E 2020 Cosmological simulations of galaxy formation *Nature Rev. Phys.* **2** 42–66
[97] Alpher R A, Follin J W and Herman R C 1953 Physical conditions in the initial stages of the expanding universe *Phys. Rev.* **92** 1347–61
[98] Wagoner R V, Fowler W A and Hoyle F 1967 On the synthesis of elements at very high temperatures *Astrophys. J.* **148** 3–49
[99] Cerdeno D G and Green. A M 2010 Direct detection of WIMPs arXiv:1002.1912 [astro-ph. CO] 2
[100] Bullock J S and Boylan-Kolchin M 2017 Small-scale challenges to the Λ CDM paradigm *Ann. Rev. Astron. Astrophys.* **55** 343–87
[101] Baer H, Choi K-Y, Kim J E and Roszkowski L 2015 Dark matter production in the early Universe: beyond the thermal WIMP paradigm *Phys. Rept.* **555** 1 60
[102] Steigman G and Turner M S 1985 Cosmological constraints on the properties of weakly interacting massive particles *Nucl. Phys.* B **253** 375–86
[103] Hu W, Rennan B and Gruzinov A 2000 Cold and fuzzy dark matter *Phys. Rev. Lett.* **85** 1158–61
[104] Hooper D, Leane R K, Tsai Y-D, Wegsman S and Witte. S J 2020 A systematic study of hidden sector dark matter:application to the gamma-ray and antiproton excesses *J. High Energy Phys.* **07** 163
[105] Griest K 1991 Galactic microlensing as a method of detecting massive compact halo objects *Astrophys. J.* **366** 412–21
[106] Villanueva-Domingo P, Mena O and Palomares-Ruiz S 2021 A brief review on primordial black holes as dark matter *Front. Astron. Space Sci.* **8** 87

[107] Carr B, Kohri K, Sendouda Y and Yokoyama J 2021 Constraints on primordial black holes *Rept. Prog. Phys.* **84** 116902
[108] Bekenstein J D 2005 Modified gravity vs dark matter: relativistic theory for MOND *PoS* JHW2004:012
[109] McGaugh S S 2015 A tale of two paradigms: the mutual incommensurability of Λ CDM and MOND *Can. J. Phys.* **93** 250–9
[110] Gaskins J M 2016 A review of indirect searches for particle dark matter *Contemp. Phys.* **57** 496–525
[111] Kahlhoefer F 2017 Review of LHC dark matter searches *Int. J. Mod. Phys.* A **32** 1730006
[112] Teresa Marrodán UndagoitiaRauch L 2016 Dark matter direct-detection experiments *J. Phys.* G Nucl. Phys. **43** 013001
[113] Silk J *et al* 2010 *Particle Dark Matter: Observations, Models and Searches* (Cambridge: Cambridge University Press)
[114] Lewin J D and Smith P F 1996 Review of mathematics, numerical factors, and corrections for dark matter experiments based on elastic nuclear recoil *Astropart. Phys.* **6** 87 112
[115] Agrawal P *et al* 2021 Feebly-interacting particles: FIPs 2020 workshop report *Eur. Phys. J.* C **81** 1015
[116] Marsh D J E 2016 Axion cosmology *Phys. Rept.* **643** 1 79
[117] Abbott B P *et al* 2016 Observation of gravitational waves from a binary black hole merger *Phys. Rev. Lett.* **116** 061102
[118] Bertone G *et al* 202 Gravitational wave probes of dark matter: challenges and opportunities *SciPost Phys. Core* **3** 007
[119] Arbey A and Mahmoudi F 2021 Dark matter and the early Universe: a review *Prog. Part. Nucl. Phys.* **119** 103865
[120] Carney D *et al* 2021 Mechanical quantum sensing in the search for dark matter *Quantum Sci. Technol.* **6** 024002
[121] Donzelli M and Carpineti M 2017 The underground laboratories global research infrastructures *Case Studies Report, GSO—Group of Senior Officials on Global Research Infrastructures, August, 2017. Presented to the G7 Science Ministers' Meeting (Turin, 27–28 September)*
[122] Angloher G *et al* 2019 Exploring with NUCLEUS at the Chooz nuclear power plant *Eur. Phys. J.* C **79** 1018
[123] Bowen M and Huber P 2020 Reactor neutrino applications and coherent elastic neutrino nucleus scattering *Phys. Rev.* D **102** 053008
[124] Dodelson S and Schmidt F 2020 *Modern Cosmology* (London: Academic)
[125] van Albada T S, Bahcall J N, Begeman K and Sancisi R 1985 The Distribution of dark matter in the spiral galaxy NGC-3198 *Astrophys. J.* **295** 305 313
[126] Springel V, White S D M, Frenk C S, Navarro J F, Jenkins A, Vogelsberger M, Wang J, Ludlow A and Helmi A 2008 A blueprint for detecting supersymmetric dark matter in the Galactic halo arXiv:0809.0894 [astro-ph] 9
[127] Einstein A 1916 Die Grundlage der allgemeinen Relativitätstheorie *Ann. Phys. (Berlin)* **49** 769
[128] Einstein A 1916 Sitzungsber *Preuss. Akad. Wiss. Berlin* **1** 688
[129] Einstein A 1918 Sitzungsberichte *Preuss. Akad. Wiss. Berlin* **1** 154
[130] Abbott B P *et al* (LIGO Scientific Collaboration and Virgo Collaboration) 2016 *Phys. Rev. Lett.* **116** 061102

[131] Abbott R *et al* 2020 (LIGO Scientific, Virgo) (*Preprint*)

[132] Aasi J *et al* 2015 Characterization of the LIGO detectors during their sixth science run *Class. Quantum Grav.* **32** 074001

[133] Acernese F *et al* 2015 Advanced Virgo: a second-generation interferometric gravitational wave detector *Class. Quantum Grav.* **32** 024001

[134] Abbott B P *et al* 2020 Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA *Living Rev. Rel.* **23** 3 (KAGRA, LIGO Scientific, Virgo)

[135] Affeldt C *et al* 2014 Advanced techniques in GEO 600 *Class. Quantum Grav.* **31** 224002

[136] Somiya K 2012 Detector configuration of KAGRA—the Japanese cryogenic gravitational-wave detector *Class. Quantum Grav.* **29** 124007

[137] Unnikrishnan C S 2013 IndIGO and LIGO-India: scope and plans for gravitational wave research and precision metrology in India *Int. J. Mod. Phys.* D **22** 1341010

[138] Punturo M *et al* 2010 The Einstein Telescope: a third-generation gravitational wave observatory *Class. Quantum Grav.* **27** 194002

[139] Reitze D *et al* 2019 Cosmic explorer: the U.S. contribution to gravitational-wave astronomy beyond LIGO *Bull. Am. Astron. Soc.* **51** 141 (*Preprint*)

[140] Amaro-Seoane P *et al* 2017 Laser interferometer space antenna arXiv:1702.00786 [astro-ph.IM] *ArXiv e-prints* (*Preprint*)

[141] Arzoumanian Z *et al* 2020 The NANOGrav 12.5 yr data set: search for an isotropic stochastic gravitational-wave background *Astrophys. J. Lett.* **905** L34 (*Preprint*)

[142] Smits R, Kramer M, Stappers B, Lorimer D, Cordes J and Faulkner A 2009 Pulsar searches and timing with the square kilometre array *Astron. Astrophys.* **493** 1161–70

[143] Lazio T 2013 The Square Kilometre Array pulsar timing array *Class. Quantum Grav.* **30** 4011

[144] Stappers B W, Keane E F, Kramer M, Possenti A and Stairs I H 2018 The prospects of pulsar timing with new-generation radio telescopes and the Square Kilometre Array *Phil. Trans. R. Soc.* A **376** 20170293

[145] Taylor J and Weisberg J 1989 Further experimental tests of relativistic gravity using the binary pulsar PSR 1913+16 *Astrophys. J.* **345** 434

[146] Abbott B P *et al* 2016 GW150914: the advanced LIGO detectors in the era of first discoveries *Phys. Rev. Lett.* **116** 131103 (LIGO Scientific, Virgo) (*Preprint*)

[147] Huterer D and Shafer D L 2018 Dark energy two decades after: observables, probes, consistency tests *Rept. Prog. Phys.* **81** 016901 [arXiv:1709.01091 [astro-ph.CO]]

[148] Zhao G B *et al* 2017 Dynamical dark energy in light of the latest observations *Nat. Astron* **1** 627–32 [arXiv:1701.08165 [astro-ph.CO]]

[149] Saridakis E N *et al* [CANTATA] Modified gravity and cosmology: an update by the CANTATA network [arXiv:2105.12582 [gr-qc]]

[150] Fender R and Belloni T 2012 Stellar-mass black holes and ultraluminous X-ray sources *Science* **337** 540 (*Preprint*)

[151] Abbott B P *et al* 2019 GWTC-1: a gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs *Phys. Rev.* **X9** 031040 (LIGO Scientific, Virgo)(*Preprint*)

[152] Abbott R *et al* 2020 GW190521: a binary black hole merger with a total mass of 150 $M_\odot$ *Phys. Rev. Lett.* **125** 101102 (LIGO Scientific, Virgo)(*Preprint*)

[153] Abbott R *et al* 2020 Properties and astrophysical implications of the 150 $M_\odot$ binary black hole merger GW190521 *Astrophys. J. Lett.* **900** L13 (LIGO Scientific, Virgo)(*Preprint*)
[154] Ghez A M *et al* 2008 Measuring distance and properties of the Milky Way's central supermassive black hole with stellar orbits *Astrophys. J.* **689** 1044–62 (*Preprint*)
[155] Genzel R, Eisenhauer F and Gillessen S 2010 The Galactic Center massive black hole and nuclear star cluster *Rev. Mod. Phys.* **82** 3121–95 (*Preprint*)
[156] Akiyama K *et al* 2019 First M87 event horizon telescope results. I. The shadow of the supermassive black hole *Astrophys. J.* **875** L1 (Event Horizon Telescope)(*Preprint*)
[157] Schwarzschild K 1916 On the gravitational field of a mass point according to Einstein's theory *Sitzungsber. Preuss. Akad. Wiss. Berlin (Berlin)* **189** 96
[158] Penrose R 1965 Gravitational collapse and space-time singularities *Phys. Rev. Lett.* **14** 57–9
[159] Penrose R 1969 Gravitational collapse: the role of general relativity *Riv. Nuovo Cim.* **1** 252
[160] Wald R M 1999 *Gravitational Collapse and Cosmic Censorship* (Dordrecht: Springer) pp 69–86
[161] Israel W 1967 Event horizons in static vacuum space-times *Phys. Rev.* **164** 1776–9
[162] Carter B 1971 Axisymmetric black hole has only two degrees of freedom *Phys. Rev. Lett.* **26** 331–3
[163] Bekenstein J D 1972 Black holes and the second law *Lett. Nuovo Cim.* **4** 737–40
[164] Hawking S W 1974 Black hole explosions? *Nature* **248** 30–1
[165] Abbott B PLIGO Scientific Collaboration and Virgo Collaboration *et al* 2016 GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence *Phys. Rev. Lett.* **116** 241103
[166] Abbott B PLIGO Scientific Collaboration and Virgo Collaboration *et al* 2016 Binary black hole mergers in the first advanced LIGO observing run *Phys. Rev.* X **6** 041015
[167] Abbott B PLIGO Scientific Collaboration and Virgo Collaboration *et al* 2017 GW170814: a three-detector observation of gravitational waves from a binary black hole coalescence *Phys. Rev. Lett.* **119** 141101
[168] Abbott B PLIGO Scientific Collaboration and Virgo Collaboration *et al* 2017 GW170817: observation of gravitational waves from a binary neutron star inspiral *Phys. Rev. Lett.* **119** 161101
[169] Abbott B PVirgo, Fermi-GBM, INTEGRAL, LIGO Scientific *et al* 2017 Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A *Astrophys. J.* **848** L13
[170] Goldstein A *et al* 2017 An ordinary short gamma-ray burst with extraordinary implications: *Fermi*-GBM detection of GRB 170817A *Astrophys. J.* **848** L14 (*Preprint*)
[171] Abbott B P *et al* 2017 Multi-messenger observations of a binary neutron star merger *Astrophys. J.* **848** L12
[172] Christensen N and Meyer R 1998 Markov chain Monte Carlo methods for Bayesian gravitational radiation data analysis *Phys. Rev.* D **58** 082001
[173] Veitch J *et al* 2015 Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library *Phys. Rev.* D **91** 042003 (*Preprint*)
[174] Abbott RLIGO Scientific, Virgo *et al* 2020 Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog arXiv:2010.14533 [astro-ph. HE] (*Preprint*)

[175] Belczynski K *et al* 2020 Evolutionary roads leading to low effective spins, high black hole masses, and O1/O2 rates for LIGO/Virgo binary black holes *Astron. Astrophys.* **636** A104 (*Preprint*)
[176] Ivanova N *et al* 2013 Common envelope evolution: where we stand and how we can move forward *Astron. Astrophys. Rev.* **21** 59 (*Preprint*)
[177] Spera M and Mapelli M 2017 Very massive stars, pair-instability supernovae and intermediate-mass black holes with the SEVN code *Mon. Not. Roy. Astron. Soc.* **470** 4739–49 (*Preprint*)
[178] Farmer R, Renzo M, de Mink S E, Marchant P and Justham S 2019 Mind the gap: the location of the lower edge of the pair-instability supernova black hole mass gap *Astrophys. J.* **887** 53
[179] Rodriguez C L, Zevin M, Pankow C, Kalogera V and Rasio F A 2016 Illuminating black hole binary formation channels with spins in advanced LIGO *Astrophys. J. Lett.* **832** L2 (*Preprint*)
[180] Vitale S, Lynch R, Sturani R and Graff P 2017 Use of gravitational waves to probe the formation channels of compact binaries *Class. Quant. Grav.* **34** 03LT01 (*Preprint*)
[181] Carr B J and Hawking S W 1974 Black holes in the early Universe *Mon. Not. R. Astron. Soc.* **168** 399–415 (*Preprint*)
[182] Carr B, Kühnel F and Sandstad M 2016 Primordial black holes as dark matter *Phys. Rev.* D **94** 083504
[183] Bird S, Cholis I, Muñoz J B, Ali-Hamoud Y, Kamionkowski M, Kovetz E D, Raccanelli A and Riess A G 2016 Did LIGO detect dark matter? *Phys. Rev. Lett.* **116** 201301 (*Preprint*)
[184] Sasaki M, Suyama T, Tanaka T and Yokoyama S 2016 Primordial black hole scenario for the gravitational-wave event GW150914 *Phys. Rev. Lett.* **117** 061101
Erratum 2018 Primordial black hole scenario for the gravitational-wave event GW150914 *Phys.Rev.Lett.* **121** 059901 (*Preprint*)
[185] Jenkins A C and Sakellariadou M 2020 Primordial black holes from cusp collapse on cosmic strings arXiv:2006.16249 [astro-ph.CO] (*Preprint*)
[186] Boehm C, Kobakhidze A, O'Hare C A J, Picker Z S C and Sakellariadou M 2020 Eliminating the LIGO bounds on primordial black hole dark matter arXiv:2008.10743 [astro-ph.CO] (*Preprint*)
[187] Fernandez N and Profumo S 2019 Unraveling the origin of black holes from effective spin measurements with LIGO-Virgo *J. Cosmol. Astropart. Phys.* **08** 022 (*Preprint*)
[188] Abbott B PLIGO Scientific, Virgo *et al* 2019 Search for subsolar mass ultracompact binaries in advanced LIGO's second observing run *Phys. Rev. Lett.* **123** 161102 (*Preprint*)
[189] Knox L, Christensen N and Skordis C 2001 The age of the Universe and the cosmological constant determined from cosmic microwave background anisotropy measurements *Astrophys. J. Lett.* **563** L95–8 (*Preprint*)
[190] Kulier A, Ostriker J P, Natarajan P, Lackner C N and Cen R 2015 Understanding black hole mass assembly via accretion and mergers at late times in cosmological simulations *Astrophys. J.* **799** 178 (*Preprint*)
[191] Begelman M C, Volonteri M and Rees M J 2006 Formation of supermassive black holes by direct collapse in pre-galactic haloes *Mon. Not. Roy. Astron. Soc.* **370** 289–98 (*Preprint*)
[192] Banados E *et al* 2018 An 800-million-solar-mass black hole in a significantly neutral Universe at a redshift of 7.5 *Nature* **553** 473–6 (*Preprint*)
[193] Wang F *et al* 2021 A luminous quasar at redshift 7.642 *Astrophys. J. Lett.* **907** L1 (*Preprint*)

[194] Greene J E, Strader J and Ho L C 2020 Intermediate-mass black holes *Annu. Rev. Astron. Astrophys.* **58** 257–312 (*Preprint*)

[195] Abbott B PLIGO Scientific, Virgo *et al* 2020 A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals *Class. Quant. Grav.* **37** 055002

[196] Abbott B PLIGO Scientific Collaboration, Virgo Collaboration *et al* 2017 GW170104: observation of a 50-solar-mass binary black hole coalescence at redshift 0.2 *Phys. Rev. Lett.* **118** 221101

[197] Unnikrishnan C S 2013 IndIGO and LIGO-India: scope and plans for gravitational wave research and precision metrology in India *Int. J. Mod. Phys.* D **22** 1341010 (*Preprint*)

[198] Abbott B PLIGO Scientific Collaboration, Virgo Collaboration *et al* 2017 GW170817: observation of gravitational waves from a binary neutron star inspiral *Phys. Rev. Lett.* **119** 161101 (*Preprint*)

[199] Abbott B PLIGO Scientific Collaboration, Virgo Collaboration, Fermi-GBM, INTEGRAL *et al* 2017 Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A *Astrophys. J.* **848** L13 (*Preprint*)

[200] Savchenko V *et al* 2017 *INTEGRAL* detection of the first prompt gamma-ray signal coincident with the gravitational-wave event GW170817 *Astrophys. J.* **848** L15 (*Preprint*)

[201] Goldstein A *et al* 2017 An ordinary short gamma-ray burst with extraordinary implications: *Fermi*-GBM detection of GRB 170817A *Astrophys. J. Lett.* **848** L14 (*Preprint*)

[202] Abbott B PLIGO Scientific Collaboration, Virgo Collaboration *et al* 2019 Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1 *Phys. Rev.* **D100** 104036 (*Preprint*)

[203] Bekenstein J D and Mukhanov V F 1995 Spectroscopy of the quantum black hole *Phys. Lett.* B **360** 7–12 (*Preprint*)

[204] Cardoso V, Foit V F and Kleban M 2019 Gravitational wave echoes from black hole area quantization *J. Cosmol. Astropart. Phys.* **2019** 6

[205] Abbott B PLIGO Scientific, Virgo *et al* 2016 Observation of gravitational waves from a binary black hole merger *Phys. Rev. Lett.* **116** 061102 (*Preprint*)

[206] Abedi J, Dykaar H and Afshordi N 2017 Echoes from the abyss: tentative evidence for Planck-scale structure at black hole horizons *Phys. Rev.* D **96** 082004 (*Preprint*)

[207] Westerweck J, Nielsen A B, Fischer-Birnholtz O, Cabero M, Capano C, Dent T, Krishnan B, Meadors G and Nitz A H 2018 Low significance of evidence for black hole echoes in gravitational wave data *Phys. Rev.* D **97** 124037

[208] Schutz B F 1986 Determining the Hubble constant from gravitational wave observations *Nature* **323** 310–1

[209] Nissanke S, Holz D E, Hughes S A, Dalal N and Sievers J L 2010 Exploring short gamma-ray bursts as gravitational-wave standard sirens *Astrophys. J.* **725** 496–514 (*Preprint*)

[210] Christensen N, Meyer R, Knox L and Luey B 2001 Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements *Class. Quantum Grav.* **18** 2677–88

[211] Riess A G *et al* 2016 A 2.4% determination of the local value of the Hubble constant *Astrophys. J.* **826** 56 (*Preprint*)

[212] Abbott B PLIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, MASTER *et al* 2017 Hungary rewards highly cited scientists with bonus grants *Nature* **551** 425–6 (*Preprint*)

[213] Ade P A R *et al* 2016 *Planck* 2015 results *Astron. Astrophys.* **594** A13 (*Preprint*)

[214] Abbott B PKAGRA, LIGO Scientific Collaboration and Virgo Collaboration *et al* 2019 Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA *Living Rev. Relativ.* **23** 3 (*Preprint*)
[215] https://gw-openscience.org
[216] Armano M *et al* 2016 Sub-femto-*g* free fall for space-based gravitational wave observatories: LISA pathfinder results *Phys. Rev. Lett.* **116** 231101
[217] Armano M *et al* 2018 Beyond the required LISA free-fall performance: new LISA pathfinder results down to 20 μHz *Phys. Rev. Lett.* **120** 061101
[218] Kamionkowski M and Kovetz E D 2016 The quest for B modes from inflationary gravitational waves *Annu. Rev. Astron. Astrophys.* **54** 227–69 (*Preprint*)
[219] Maggiore M *et al* 2020 Science case for the Einstein telescope *J. Cosmol. Astropart. Phys.* **03** 050 (*Preprint*)
[220] Christensen N 2019 Stochastic gravitational wave backgrounds *Rep. Prog. Phys.* **82** 016903
[221] Biscoveanu S, Talbot C, Thrane E and Smith R 2020 Measuring the primordial gravitational-wave background in the presence of astrophysical foregrounds *Phys. Rev. Lett.* **125** 241101
[222] Martinovic K, Meyers P M, Sakellariadou M and Christensen N 2021 Simultaneous estimation of astrophysical and cosmological stochastic gravitational-wave backgrounds with terrestrial detectors *Phys. Rev.* D **103** 043023
[223] Regimbau T, Evans M, Christensen N, Katsavounidis E, Sathyaprakash B and Vitale S 2017 Digging deeper: observing primordial gravitational waves below the binary-black-hole-produced stochastic background *Phys. Rev. Lett.* **118** 151105
[224] Sachdev S, Regimbau T and Sathyaprakash B 2020 Subtracting compact binary foreground sources to reveal primordial gravitational-wave backgrounds *Phys. Rev.* D **102** 024051 (*Preprint*)
[225] Sesana A 2016 Prospects for multiband gravitational-wave astronomy after GW150914 *Phys. Rev. Lett.* **116** 231102 (*Preprint*)
[226] Adams M R and Cornish N J 2014 Detecting a stochastic gravitational wave background in the presence of a galactic foreground and instrument noise *Phys. Rev.* D **89** 022001 (*Preprint*)
[227] Boileau G, Christensen N, Meyer R and Cornish N J 2020 Spectral separation of the stochastic gravitational-wave background for LISA: observing both cosmological and astrophysical backgrounds arXiv:2011.05055 (*Preprint*) arXiv e-prints
[228] Mingarelli C M F 2019 Improving binary millisecond pulsar distances with Gaia arXiv:1812.06262 [astro-ph.IM] (*Preprint*)
[229] Buldgen G, Salmon S and Noels A 2019 Progress in global helioseismology: a new light on the solar modelling problem and its implications for solar-like stars *Front. Astron. Space Sci.* **6** 42
[230] Christensen-Dalsgaard J 2021 Solar structure and evolution *Living Rev. Sol. Phys.* **18** 2
[231] Basu S and Antia H M 1997 Seismic measurement of the depth of the solar convection zone *Mon. Not. R. Astron. Soc.* **287** 189–98
[232] Basu S and Antia H M 1995 Helium abundance in the solar envelope *Mon. Not. R. Astron. Soc.* **276** 1402–8
[233] Vorontsov S V, Baturin V A, Ayukov S V and Gryaznov V K 2013 Helioseismic calibration of the equation of state and chemical composition in the solar convective envelope *Mon. Not. R. Astron. Soc.* **430** 1636–52

[234] Vorontsov S V, Baturin V A and Pamiatnykh A A 1991 Seismological measurement of solar helium abundance *Nature* **349** 49–51

[235] Grevesse N and Noels A 1992 Cosmic abundances of the elements *Origin and Evolution of the Elements: Proceedings of a Symposium in Honour of H. Reeves held in Paris, June 22–25, 1993* ed N Prantzos, E Vangioni-Flam, M Casse, N Prantzos, E Vangioni-Flam and M Casse (Cambridge: Cambridge University Press) p 14

[236] Christensen-Dalsgaard J *et al* 1996 The current state of solar modeling *Science* **272** 1286

[237] Asplund M, Grevesse N and Sauval A J 2005 The solar chemical composition *Cosmic Abundances as Records of Stellar Evolution and Nucleosynthesis, of Astronomical Society of the Pacific Conf. Series* ed III Barnes, G Thomas and F N Bash (San Francisco, CA: Astronomical Society of the Pacific) vol 336 p 25

[238] Asplund M, Grevesse N, Jacques Sauval A and Scott. P 2009 The chemical composition of the Sun *Annu. Rev. Astron. Astrophys.* **47** 481–522

[239] Basu S and Antia H M 2008 Helioseismology and solar abundances *Phys. Rep.* **457** 217–83

[240] Pezzotti C, Eggenberger P, Buldgen G, Meynet G, Bourrier V and Mordasini C 2021 Revisiting Kepler-444. II. Rotational, orbital, and high-energy fluxes evolution of the system *Astron. Astrophys.* **650** A108

[241] Serenelli A M, Basu S, Ferguson J W and Asplund M 2009 New solar composition: the problem with solar models revisited *Astrophys. Lett.* **705** L123–7

[242] Pinsonneault M H, Kawaler S D, Sofia S and Demarque P 1989 Evolutionary models of the rotating sun *Astrophys. J.* **338** 424–52

[243] Charbonneau P and MacGregor K B 1993 Angular momentum transport in magnetized stellar radiative zones. II. The solar spin-down *Astrophys. J.* **417** 762

[244] Gough D O and McIntyre M E 1998 Inevitability of a magnetic field in the Sun's radiative interior *Nature* **394** 755–7

[245] Mestel L 1953 Rotation and stellar evolution *Mon. Not. R. Astron. Soc.* **113** 716

[246] Mestel L and Weiss N O 1987 Magnetic fields and non-uniform rotation in stellar radiative zones *Mon. Not. R. Astron. Soc.* **226** 123–35

[247] Mosser B *et al* 2012 Spin down of the core rotation in red giants *Astron. Astrophys.* **548** A10

[248] Spada F, Lanzafame A C and Lanza A F 2010 A semi-analytic approach to angular momentum transport in stellar radiative interiors *Mon. Not. R. Astron. Soc.* **404** 660

[249] Braithwaite J and Spruit H C 2017 Magnetic fields in non-convective regions of stars *R. Soc. Open Sci.* **4** 160271

[250] Brun A S and Zahn J-P 2006 Magnetic confinement of the solar tachocline *Astron. Astrophys.* **457** 665–74

[251] Tayler R J 1973 The adiabatic stability of stars containing magnetic fields-I. Toroidal fields *Mon. Not. R. Astron. Soc.* **161** 365

[252] Spruit H C 2002 Dynamo action by differential rotation in a stably stratified stellar interior *Astron. Astrophys.* **381** 923–32

[253] Eggenberger P, Maeder A and Meynet G 2005 Stellar evolution with rotation and magnetic fields. IV. The solar rotation profile *Astron. Astrophys.* **440** L9–L12

[254] Eggenberger P, Buldgen G and Salmon S J A J 2019 Rotation rate of the solar core as a key constraint to magnetic angular momentum transport in stellar interiors *Astron. Astrophys.* **626** L1

[255] Salmon S, Buldgen G, Noels A, Eggenberger P, Scuflaire R and Meynet G 2021 Standard solar models: perspective from updated solar neutrino fluxes and gravity-mode period spacing *Astron. Astrophys.* **651** A106
[256] Miglio A *et al* 2013 Galactic archaeology: mapping and dating stellar populations with asteroseismology of red-giant stars *Mon. Not. R. Astron. Soc.* **429** 423–8
[257] Miglio A, Montalbán J, Baudin F, Eggenberger P, Noels A, Hekker S, De Ridder J, Weiss W and Baglin A 2009 Probing populations of red giants in the galactic disk with CoRoT *Astron. Astrophys.* **503** L21–4
[258] Miglio A *et al* 2021 HAYDN *Exp. Astron.* **51** 963–1001
[259] Arnett W D, Meakin C, Viallet M, Campbell S W, Lattanzio J C and Mocák M 2015 Beyond mixing-length theory: a step toward 321D *Astrophys. J.* **809** 30
[260] Aerts C 2021 Probing the interior physics of stars through asteroseismology *Rev. Mod. Phys.* **93** 015001
[261] Beck P G *et al* 2012 Fast core rotation in red-giant stars as revealed by gravity-dominated mixed modes *Nature* **481** 55–7
[262] Deheuvels S *et al* 2012 Seismic evidence for a rapidly rotating core in a lower-giant-branch star observed with kepler *Astrophys. J.* **756** 19
[263] Gehan C, Mosser B, Michel E, Samadi R and Kallinger T 2018 Core rotation braking on the red giant branch for various mass ranges *Astron. Astrophys.* **616** A24
[264] Ceillier T, Eggenberger P, Garca R A and Mathis S 2013 Understanding angular momentum transport in red giants: the case of KIC 7341231 *Astron. Astrophys.* **555** A54
[265] Eggenberger P, Montalbán J and Miglio A 2012 Angular momentum transport in stellar interiors constrained by rotational splittings of mixed modes in red giants *Astron. Astrophys.* **544** L4
[266] Marques J P *et al* 2013 Seismic diagnostics for transport of angular momentum in stars. I. Rotational splittings from the pre-main sequence to the red-giant branch *Astron. Astrophys.* **549** A74
[267] Eggenberger P *et al* 2019 Asteroseismology of evolved stars to constrain the internal transport of angular momentum. I. Efficiency of transport during the subgiant phase *Astron. Astrophys.* **621** A66
[268] Eggenberger P *et al* 2017 Constraining the efficiency of angular momentum transport with asteroseismology of red giants: the effect of stellar mass *Astron. Astrophys.* **599** A18
[269] Mayor M and Queloz D 1995 A Jupiter-mass companion to a solar-type star *Nature* **378** 355–9
[270] Ricker G R *et al* 2015 Transiting Exoplanet Survey Satellite (TESS) *J. Astron. Teles., Instrum., Syst.* **1** 014003
[271] Rüdiger G and Kitchatinov L L 1996 The internal solar rotation in its spin-down history *Astrophys. J.* **466** 1078
[272] Privitera G, Meynet G, Eggenberger P, Vidotto A A, Villaver E and Bianda M 2016 Star–planet interactions. I. Stellar rotation and planetary orbits *Astron. Astrophys.* **591** A45
[273] Privitera G, Meynet G, Eggenberger P, Vidotto A A, Villaver E and Bianda M 2016 Star–planet interactions. II. Is planet engulfment the origin of fast rotating red giants? *Astron. Astrophys.* **593** A128
[274] Rao S, Meynet G, Eggenberger P, Haemmerlé L, Privitera G, Georgy C, Ekström S and Mordasini C 2018 Star–planet interactions. V. Dynamical and equilibrium tides in convective zones *Astron. Astrophys.* **618** A18

[275] Lopez E D and Fortney J J 2013 The role of core mass in controlling evaporation: the Kepler radius distribution and the Kepler-36 density dichotomy *Astrophys. J.* **776** 2
[276] Owen J E and Lai D 2018 Photoevaporation and high-eccentricity migration created the sub-Jovian desert *Mon. Not. R. Astron. Soc.* **479** 5012–21
[277] Johnstone C P, Bartel M and Güdel M 2021 The active lives of stars: a complete description of the rotation and XUV evolution of F, G, K, and M dwarfs *Astrophys. J.* **649** A96
[278] Tu L, Johnstone C P, Güdel M and Lammer H 2015 The extreme ultraviolet and X-ray Sun in time: high-energy evolutionary tracks of a solar-like star *Astrophys. J.* **577** L3
[279] Johnstone C P, Güdel M, Stökl A, Lammer H, Tu L, Kislyakova K G, Lüftinger T, Odert P, Erkaev N V and Dorfi E A 2015 The evolution of stellar rotation and the hydrogen atmospheres of habitable-zone terrestrial planets *Astrophys. Lett.* **815** L12
[280] Kubyshkina D *et al* 2019 The Kepler-11 system: evolution of the stellar high-energy emission and initial planetary atmospheric mass fractions *Astron. Astrophys.* **632** A65
[281] Rao S, Pezzotti C, Meynet G, Eggenberger P, Buldgen G, Mordasini C, Bourrier V, Ekström S and Georgy C 2021 Star–planet interactions VI. Tides, stellar activity, and planetary evaporation *Astron. Astrophys.* **651** A50
[282] Turcotte D and Schubert J 2014 *Geodynamics* (London: Cambridge University Press)
[283] Moffatt K and Dormy E 2019 *Self-Exciting Fluid Dynamos* (London: Cambridge University Press)
[284] Braginskii S I 1964 Kinematic models of the Earth's hydromagnetic dynamo *Geomagn. Aeron.* **4** 572–83
[285] Steenbeck M, Krause F and Radler K-H 1966 Berechnung der mittleren Lorentz-Feldstärke für ein elektrisch leitendes Medium in turbulenter, durch Coriolis-Kräfte beeinflusster Bewegung *Z. Naturforsch.* **21a** 369–76 (English translation: Roberts and Stix (1971) pp 29–47)
[286] Glatzmaier G A and Roberts P H 1995 A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle *Phys. Earth Planet. Inter.* **91** 63–75
[287] Pétrélis F, Fauve S, Dormy E and Valet J-P 2009 Simple mechanism for reversals of Earth's magnetic field *Phys. Rev. Lett.* **102** 144503
[288] Dormy E 2006 The origin of the Earth's magnetic field: fundamental or environmental research? *Europhys. News* **37** 2
[289] Kaminski E and Jaupart C 1998 The size distribution of pyroclasts and the fragmentation sequence in explosive volcanic eruptions *J. Geophys. Res.* **103** 29759–79
[290] Michaud-Dubuy A, Carazzo G, Kaminski E and Girault F 2002 A revisit of the role of gas entrapment on the stability conditions of explosive volcanic columns *J. Volcanol. Geotherm. Res.* **357** 349–61
[291] Geshi N, Browning J and Kusumoto S 2020 Magmatic overpressures, volatile exsolution and potential explosivity of fissure eruptions inferred via dike aspect ratios *Sci. Rep.* **10** 9406
[292] D'Incecco P, Filiberto J, López I, Gorinov D and Idunn Mons K G 2021 Evidence for ongoing volcano-tectonic activity and atmospheric implications on Venus *Planet. Sci. J.* **2** 215
[293] Burridge R and Knopoff L 1967 Model and theoretical seismicity *Bull. Seismol. Soc. Am.* **57** 341

[294] Clancy I and Corcoran D 2006 Burridge-Knopoff model: exploration of dynamic phases *Phys. Rev.* E **73** 046115

[295] Pétrélis F, Chanard K, Schubnel A and Hatano T 2023 Earthquake magnitude distribution and aftershocks: A statistical geometry explanation *Phys. Rev. E* **107** 034132

[296] Morell K D, Styron R, Stirlin M, Griffin J, Archuleta R and Onur T 2020 Seismic hazard analyses from geologic and geomorphic data: current and future challenges *Tectonics* **39** e2018TC005365