Found Phys (2018) 48:1305–1332
https://doi.org/10.1007/s10701-018-0183-y

Towards the Limits of Cosmology

Joseph Silk[1,2,3]

Received: 8 January 2018 / Accepted: 2 June 2018 / Published online: 21 June 2018

Abstract One of our greatest challenges in cosmology is understanding the origin of the structure of the universe. I will describe how the cosmic microwave background has provided a window for probing the initial conditions from which structure evolved and seeded the formation of the galaxies, and the outstanding issues that remain to be resolved. I will address our optimal choice of future strategy in order to make further progress on understanding our cosmic origins.

Keywords Cosmology · Dark matter · Dark energy · Cosmic microwave background · Radio astronomy · Inflation

1 Introduction

We are at an impasse in cosmology. We seek definitive evidence for the nature of dark matter, dark energy physics, and robust tests of inflation. There has been no detection of dark matter particles despite decades of intensive searches. Increasingly precise probes of dark energy come up with improving confirmation of constant Λ and equation of state $p = -\rho c^2$. And our best theory of the beginning of the universe, inflation, awaits a definitive and falsifiable probe, in order to satisfy most physicists

✉ Joseph Silk
silk@astro.ox.ac.uk

1 Institut d'Astrophysique, 98 bis Boulevard Arago, 75014 Paris, France

2 Department of Physics and Astronomy, The Johns Hopkins University, 3701 San Martin Drive, Baltimore, MD 21218, USA

3 Department of Physics and Beecroft Institute for Particle Astrophysics and Cosmology, University of Oxford, Denys Wilkinson Building, 1 Keble Road, Oxford OX1 3RH, UK

that it is a trustworthy theory. I will summarise the present state of cosmology and review prospects for future progress.

2 The Dark Matter Saga

The dark matter saga began with the Swiss-American astronomer Fritz Zwicky in 1933 [1]. He realized that the Coma cluster of galaxies would fly apart were it not held together by unseen matter. He estimated that there had to be ten times more dark matter than visible matter. The only data available at that time were the measurements of the relative velocities of the cluster galaxies. We now have measurements of cluster masses from more refined and accurate techniques, most notably x-ray maps and spectra of the hot gas distribution and gravitational lensing of distant galaxies. Zwicky's original estimate is largely confirmed by a combination of x-ray studies, CMB detections and gravitational lensing maps.

Some two decades after Zwicky's discovery, observations of nearby spiral galaxies began to show evidence for dark matter. The best case was made for the Andromeda galaxy, a neighbor of the Milky Way galaxy. In the early 1950s, US astronomer Vera Rubin used optical data to study ionized nebulae surrounding massive stars in Andromeda [2]. These are very bright systems and can be observed throughout the luminous part of the galaxy and even well beyond. Her result was soon confirmed by radio astronomer Morton Roberts, who performed similar measurements for atomic hydrogen clouds [3]. These extend to much larger distances. There was now little doubt: most of the mass of disk galaxies like Andromeda was dark. Over the next decade, similar evidence was found for many other galaxies of all types.

The existence of dark matter has now been inferred over a much wider range of scales from its gravitational effects, spanning dwarf galaxies, clusters of galaxies and the large-scale distribution of galaxies. On galactic scales, we measure the velocities of stars and gas clouds around the centers of galaxies. On larger scales, we probe dark matter through the motions of galaxies in clusters. High precision mapping of dark matter is now performed by gravitational lensing. The distortions of distant galaxies can be measured, and because there are many galaxies behind nearby clusters, for example, one can build up a map of the dark matter distribution.

On the very largest scales, dark matter controls the formation of cosmic structure, and indeed the evolution of the universe as a whole. Results on the anisotropies in the cosmic microwave background radiation (from the Planck and WMAP satellites) suggest that about a quarter of the cosmic energy budget is in the form of dark matter. Dark matter provides the scaffolding on which the large-scale structure of the universe coalesces. The combination of satellite telescopes with larger, and consequently higher resolution terrestrial telescopes has resolved parameter degeneracies and led us to the era pf precision cosmology, at the 1% level.

Lets ignore for the moment various degeneracies between the secondary (calculated) and primary (measured) parameters. The CMB, most precisely via Planck, primarily probes curvature, or the combination $\Omega_m + \Omega_\Lambda$. Supernovae probe acceleration, or the combination $\Omega_m - \Omega_\Lambda$. Gravitational lensing and baryon acoustic oscillations are especially sensitive to Ω_m. Now we have redundancy in $\Omega_m - \Omega_\Lambda$

parameter space. This is what has led to the advent of precision cosmology, We have now measured the mean density of dark matter to a value, in terms of the critical density of an Einstein-de Sitter Universe, $\Omega_{DM} = 0.2589 \pm 0.0057$. Now we have a very specific dark matter target.

Dark matter supposedly consists of weakly interacting massive particles (WIMPs), that interact among themselves and with ordinary matter only via the weak nuclear force that controls scattering, and gravitationally, but usually not electromagnetically or via the strong nuclear force. WIMPs are the lightest of the so-called supersymmetric (SUSY) particles. These are all majorons, or their own antiparticles. Almost all of the partner SUSY particles decayed or annihilated with each other, but the lightest partner particle is stable, and survived to be the dark matter candidate particle.

Remarkably, since the particles once were in thermal equilibrium, one can calculate how many such particles freeze out of equilibrium and survive until the present epoch in the universe. The relic fraction determines just how much dark matter there is.

To obtain the relic baryon abundance, one needs to build in a baryon number, an excess of baryons over antibaryons, at the beginning, or more precisely, the epoch of baryogenesis. From the observed CMB photon density, a proxy for the number of particles and antiparticles early on, one infers the needed baryon asymmetry of order 10^{-8}. This could develop as early as the epoch of grand unification, or as late as the electroweak phase transition. A good model for baryogenesis is yet to emerge, but we are confident of its general ingredients as laid out by Sakharov in 1967 [4]. These consist of experimentally observed or inferred phenomena: baryon number violation, CP violation and out-of- equilibrium interactions due to the expansion of the universe, essentially providing an arrow of time.

Models identifying dark matter particles with the lightest stable survivor from SUSY naturally give the particle mass range of a neutralino, typically expected to be in the range tens to thousands of proton masses, along with the interaction strength, invariably weak. If the relic particles interacted too weakly, there would be too many survivors, if too strongly, too few. Like Goldilocks, the properties must be just right, if the dark matter particle is the lightest stable SUSY relic. It is this coincidence that has motivated a plethora of ever more sensitive direct detection experiments. While a mass of order the Higgs mass or a hundred GeV is favoured for the SUSY breaking scale, the failure to find SUSY has motivated exploration of less restrictive model building. The mass range is easily extended by theorists, appealing to non-minimal SUSY and related phenomenology, to span masses below a GeV to above a TeV, providing new challenges for experimentalists.

2.1 Direct Detection

The majority of the current experiments use either cryogenic detectors (that detect the heat released when a particle hits an atom in a crystal), or noble liquid (xenon or argon) detectors (that monitor the ionizations and scintillations produced in scattering and collisional excitations). No signal has been found to date. But it may be that detectors simply are not yet large and sensitive enough. Physicists are already planning detectors with several metric tons of liquid xenon. The next step, under design, is a 100 ton-scale

detector, and that should be the ultimate limit for any future efforts in direct detection about a decade from now. At this point, coherent scattering of neutrinos from the earth's atmosphere and from the sun give an impenetrable detector background: we can never do better, it seems for direct detection. However direct detection experiments generally lose sensitivity below a few GeV, and novel methods are being developed to explore sub-GeV particle scales.

A second type of search for dark matter particles involves colliding protons together at unprecedentedly high energies. One could thereby hope to produce dark matter particles directly. The Large Hadron Collider (LHC) in Geneva does precisely this at a center-of-mass energy of 14 TeV. No experimental evidence for the existence of the theoretically predicted supersymmetric particles - subatomic partners to the known elementary particles - has been seen so far. This weakens the case for the most popular dark matter candidate particle.

Failure so far to detect any hints of SUSY tells us that we may need to push towards even higher energies. At such a high energy scale, SUSY becomes increasingly unnatural as the SUSY scale moves away from the Higgs mass. Hence many particle physicists believe that a final push in energy would be decisive in exploring the last vestiges of plausible parameter space. A future 100 TeV collider merits serious consideration as the ultimate step in exploring the high energy physics frontier.

Of course there are different views on what is natural in terms of parameter space for SUSY, or for that matter, any other theory. Dark energy, as we will discuss below, is thought to be unnatural, because one has to somehow account for some 120 orders of magnitude. Consequently a trifling three orders of magnitude of tuning the Higgs scale, and the need for vacuum energy cancellations to say a 100 TeV, SUSY scale seems relatively natural. One does not really have any compelling reason to stop searching for evidence of SUSY at 100 TeV or beyond, apart from the overwhelming scale and corresponding cost of going to this or even higher energies using current accelerator technology.

Similarly, axions (another type of potential dark matter candidate) remain elusive despite a number of ongoing experiments. Axions are extremely low mass weakly interacting particles, of mass up to a trillionth the mass of a proton and motivated to solve a fine-tuning issue, the strong CP problem in quantum chromodynamics. Axions interact electromagnetically and can generate microwave photons inside strong magnetic field cavities, where they freely penetrate. A variety of experiments aiming at the detection of axion dark matter particles has been operating since the 1980s, without success. Unlike a plethora of other DM candidates, both axions and neutralinos are well motivated by particle physics.

2.2 Indirect Detection

Indirect detection of dark matter has hitherto been equally inconclusive. There are several nearby environments known to be rich in dark matter, and these provide promising targets for signatures of annihilation products. One intriguing result from the Fermi satellite has somehow stood the test of time. Over the past several years, the Fermi telescope has reported a diffuse glow of gamma rays extending up to around 20 degrees

from the centre of our galaxy. This is just where the dark matter in our galaxy is most concentrated. There are other contributions to diffuse gamma rays, most notably from the interaction of high energy cosmic rays with dense interstellar gas clouds, that generate competing backgrounds.

Once all known gamma ray emission templates are subtracted from the observed flux, a diffuse excess of gamma rays remains. Curiously, it has a simple morphology and is distributed spherically symmetrically about the Galactic Center, just like the dark matter. Its intensity rises towards the center just as expected if the dark matter produces gamma rays by annihilations between dark matter particles. That is, the radial gradient of the gamma ray excess is reported to be proportional to the square of the dark matter density. The spectrum of the excess is measured and it peaks in the tens of GeV range. Calculations of the predicted spectrum from annihilations fit well with a leptophilic dark matter candidate of mass 30 GeV that self-annihilates with the predicted thermal cross-section, inferred from the relic density of dark matter particles.

Dark matter does indeed increase in density towards the Galactic Center. This is inferred from the galactic rotation curve. Within a kiloparsec, however, there is increasing uncertainty, as this is where stars dominate. Most notably, there is a stellar bar that influences stellar dynamics. Appeal to cold dark matter theory suggests that the innermost density profile of dark matter rises inversely with radial distance towards the center of our galaxy, but observations of stellar and gas kinematics seem to prefer a kpc-scale core [5,6].

While both the radial dependence with distance from the center, the intensity and spectrum of the excess gamma rays fit the simplest WIMP model, there is a second issue that has led to scepticism of the DM interpretation. Studies of the fluctuations in the diffuse gamma ray emission favour many sources, presumably millisecond pulsars, over the hypothesis of a truly diffuse dark matter-induced flux, even though one cannot detect the individual sources [7]. Thousands of these weak gamma ray sources in the central region would collectively contribute to the diffuse flux. Millisecond pulsars are conjectured to be a possible candidate [8].

Such sources are common in massive globular clusters. One scenario involves massive globular star clusters falling into the galactic centre region by dynamical friction and being tidally disrupted, leaving an enhanced population of millisecond pulsars [9].

Dark matter models can also produce point-like sources. These may be ultracompact relic dark matter clumps from the early universe, or even intermediate mass black holes, a possible sub-dominant component of dark matter and motivated by galaxy formation theory, that accumulated clouds of dark matter when they formed in the early universe. The jury is out on this question of whether the Galactic centre provides evidence for the presence of dark matter.

Dwarf galaxies are another dark matter-rich environment. If the Fermi hint of dark matter were correct, one should also detect a similar signal from the ultrafaint dwarf galaxies in the Milky Way halo, as these are completely dominated by dark matter. Dark matter-rich dwarfs have been extensively surveyed by the Fermi and HESS telescopes for gamma ray excesses. No gamma ray signal has yet been definitively detected from dwarf galaxies.

If dark matter particles collect in the sun, they may also annihilate in the solar core, provided the WIMP mass exceeds 4-5 GeV. Dark matter particle annihilations produce energetic neutrinos at GeV energies which escape from the sun. These are potentially smoking gun signals of dark matter [10]. The high energy neutrinos reach and traverse the earth where they produce muons. These short-lived particles in turn can be intercepted in underground detectors by looking for upgoing faint flashes of blue Cerenkov light of nanosecond duration. One experiment monitors 50,000 tons of purified water in a dark underground cavern. This is the Super-Kamiokande neutrino observatory located under Mount Ikeno near the city of Hida in Japan, which seeks light flashes from upgoing muons for evidence of energetic neutrinos from the direction of the sun. No candidate events have so far been reported. The experiment is being updated to Hyper-Kamiokande, whose fiducial target mass will be a megaton of purified water.

2.3 Other Directions

We certainly should be prepared to seek broader categories of dark matter particles. After all, ordinary, or baryonic, matter is also composed of many types of particles. Dark matter particles might, for example, not undergo annihilations but could have an enhanced scattering signal. If in addition they had light masses, below say ten proton masses, a range not easily probed by direct detection experiments, one would find that the Sun could be as powerful a collector and detector as any terrestrial experiment. Such particles would slightly modify the temperature profile of the sun and thereby affect helioseismology, for example, more effectively than their massive counterparts. Very massive particles cannot easily be seen by any direct or even indirect experiments because their flux is so low. But with some ingenuity, one can imagine that cold objects such as old neutron stars, white dwarfs and even planets, where they would collect, might potentially provide sources of novel signals.

One alternative to dark matter, known as Modified Newtonian Dynamics (MOND), was already proposed as early as 1983. Physicist Moti Milgrom suggested this approach as a phenomenological framework, in which rather than invoking dark matter to explain velocities around galactic centers, he modified Newton's law of gravity [11]. His idea is that the Newtonian force law must be tweaked at very low accelerations, such as the ones encountered at the very outskirts of galaxies. He showed that MOND accounts for the rotation curves of hundreds of nearby galaxies, without any need for dark matter, simply by choosing one parameter that represents the scale at which MOND-like gravity replaces that of Newton. Modern data shows that rotation curves, especially of dark matter-dominated dwarf galaxies, are often complex and may require a more complicated model [12]. Clusters of galaxies require some dark matter even if MOND were correct [13]. And there is a remarkable galaxy cluster known as the Bullet cluster in which the dominant concentration of baryons, in the form of hot x-ray emitting gas, and dark matter, mapped by gravitational lensing, are clearly spatially and morphologically separated. This tells us that baryons, even with a non-Newtonian force law, cannot simply trace the dark matter: strong lensing models demonstrate that the baryonic x-ray emitting gas, accounting for most of the baryons, and the dark matter are clearly displaced [14].

Perhaps the most severe critique of MOND is that it is incomplete as a theory, since it is not Lorentz invariant. A generalisation, tensor-vector-scalar gravity [15], is fully Lorentz invariant and accounts for the phenomenological issues for which MOND was developed. However it has recently been ruled out, along with essentially all massive gravity models, by the recent detection of an electromagnetic counterpart to a gravitational wave event involving merging neutron stars [16].

The advocates of modified gravity have not abandoned their pursuit of an alternative to Newtonian gravity. A crucial ingredient of Galilean mechanics and indeed Newton's laws of motion is the dependence on the observer's reference frame. Einstein considered that a universal theory of gravity should not be observer-dependent, and this motivated him to successfully develop the theory of relativity, in which there is no preferred reference frame. There are indeed observer-independent theories of modified gravity, but all suffer from faults. The worst of these are called ghosts, singularities in the theory. There are strong constraints from solar system measurements and the recent LIGO detection of a merging binary neutron star, along with its electromagnetic counterparts, that leave little scope for natural choices of parameters. Modified gravity theories rapidly become complex and ugly, if they work at all. Dark matter seems here to stay. But the path forward remains to be clarified.

We certainly should be preparing to seek broader categories of dark matter particles, or even unparticles, as dubbed by one theorist. Unparticles are massive dark scalar fields. Dark matter particles might even not undergo any annihilations but could have an enhanced scattering signal. Some theorists have started to wonder whether dark matter truly exists. The worst scenarios are threefold.

Firstly, there may indeed be no single solution to the identity of dark matter particles. There could be several contributors, perhaps even aided and abetted by modified gravity which growing numbers of theorists are invoking to address another equally grave astrophysics problem, that of the origin of the cosmological constant. Secondly, the dark matter particles may be almost completely elusive and interact only gravitationally, such as the gravitino, the SUSY counterpart of the graviton.

Perhaps indeed we do not understand gravity. Einstein's theory, despite the overwhelming evidence for it most recently manifested by the LIGO discovery of gravitational waves, may simple be wrong in the low density outer parts of galaxies where dark matter seems to be dominant.

Our choices are not simple, but there are simply too many astrophysical and cosmological observations in support of dark matter for the dark matter problem to disappear. The coming decade will provide a fascinating challenge for elucidating the nature of the dark matter particles. However at present, scientists are not hopeful of a breakthrough. If dark matter particles still are not detected within the next decade, we should be prepared for a serious re-evaluation of our options.

3 Dark Energy

The universe is accelerating thanks to an infinitesimal amount of negative energy. Now mass causes matter to collapse via gravity. It is an attractive force. But physics allows us to have negative energy. This seems weird, but that's exactly what potential energy

is, in a gravity field. Tension in a rubber band is an example of potential energy. A catapult has potential energy.

The source of gravity is matter. But it is really the sum of matter and pressure. Normally pressure is positive, it pushes a system apart. But it can also be negative, as in the example of a rubber band under tension. If positive pressure is dominant, it adds to the gravity, according to general relativity. But any contribution by negative pressure does the opposite. If there is more negative than positive energy in a system it will expand, not collapse. It will even accelerate if the positive contribution to energy, coming say from the density of matter, which accounts for gravity, is sufficiently subdominant, that is if $p < -\frac{1}{3}\rho c^2$.

There is energy in nothing, so the quantum theory asserts. Heisenberg's uncertainty principle is at the core of quantum theory. Our inability to do a precise measurement of the position of an electron is because the act of observation itself, via transmission of a photon, induces an intrinsic uncertainty in position and time. This uncertainty is equivalent to tiny fluctuations in space and time. These fluctuations generate a tiny energy. Hence even the most perfect vacuum, as long as it is in space and time, contains an infinitesimal amount of energy.

3.1 Einstein's Prediction

But in the early twentieth century, Einstein did not know about dark energy. He developed a cosmology that contained only matter and energy, which acted as the source of gravity. For Einstein, space, matter and gravity were interchangeable. Energy, observable as radiation or thermal energy, was negligible. Matter could be replaced by the curvature of space, and the motion of matter or rays of light was directed by the curvature of space.

The prediction that matter curved space was confirmed in 1919. A war-weary world awoke to triumphant newspaper headlines reporting the results of two expeditions that verified Einstein's prediction that the positions of stars near the solar limb were found to be displaced when their positions were revealed during a total eclipse of the sun. Another major prediction of general relativity, that of black holes, was only confirmed a century later by the discovery of gravitational waves.

Einstein assumed the universe was the same everywhere and in all directions, a concept that he dubbed the cosmological principle. Einstein showed that his theory of gravity naturally led to an explanation of the universe of galaxies, which he assumed to be stationary. To see how this works, let us symbolically denote Einstein's equations of general relativity, the theory of gravity, as follows:

$$G_{\mu,\nu} = \frac{8\pi G}{c^4} T_{\mu,\nu},$$

which tells us that gravity is described by the local curvature of space, which in turn is sourced by the energy-momentum tensor, or in more user-friendly language is

$$\text{CURVATURE OF SPACE} = \text{MASS/ ENERGY}.$$

Gravity is defined by matter curving space, so Einstein told us in 1916, and mass-energy is its source. But there was a problem. Einstein soon realized that since the universe contains matter, and gravity is an attractive force, this equation implies a collapsing universe. Mass and energy are also attractive, as far as their contribution to gravity goes. Indeed, one revelation of Einstein's theory was even that ordinary pressure due to matter is primarily attractive in a gravity field. I could write this equation more rigorously in tensor form but it is not really necessary. Indeed one could argue that the tensor notation was effective at keeping most theologians out of modern cosmology.

Einstein was convinced the universe was static, as were essentially all of his contemporaries. He addressed the problem of the stability of the universe in his first paper on cosmology in 1917, and introduced a cosmological constant term, Λ, to counter gravity, because without this, the universe would be unstable and collapse. Now we have in Einstein's language, with gravity embedded into the curvature of space,

$$\text{CURVATURE OF SPACE} + \Lambda = \text{MASS/ENERGY}.$$

Mass and energy locally curve space, that's the essence of general relativity, proposed by Einstein in 1915. He added a constant Λ to oppose gravity and yield a static universe, one that did not collapse. Think of Λ as antigravity. It is positive and balances the (negative) gravitational potential energy. No need for a non-stationary universe. The situation however soon became more complex when in the same year the Dutch physicist Willem de Sitter noticed that the introduction of Λ also allowed the universe to be empty. Mass and energy would be irrelevant. There was one bonus from de Sitter's solution, as it predicted that radiation emitted by test particles in the empty universe would be redshifted. This result intrigued astronomers because of the emerging evidence being accumulated, in particular by US astronomer Vesto Slipher at Flagstaff, Arizona, that the spectra of the extragalactic nebulae, now called galaxies, seemed to be systematically shifted with increasing distance towards red wavelengths.

The Russian mathematical physicist Alexander Friedmann in 1922, at the age of 34, discovered expanding cosmological solutions of Einstein's equations of general relativity. Einstein at first refused to accept the new solutions, suggesting a mathematical error, then dismissed them in 1923 as being unrealistic. In fact it was Einstein who had made the mathematical error in overlooking the expanding solution.

3.2 Observations Play a Role

Unfortunately, Friedmann died prematurely in 1925, and his work was ignored in the west. It was left to Georges Lemaitre, a Belgian physicist and recently ordained priest, to republish in 1927, unaware of Friedmann's work, the solutions to the expanding universe. There was one notable difference. Lemaitre knew about Slipher's results on the recession of the extragalactic nebulae as well as of Hubble's measurements of their large distances.

But once expansion was discovered, there was a natural way of opposing the pull of gravity. Lemaitre presented the expanding universe models, and demonstrated that an

expanding universe fit the astronomical data without any recourse to a cosmological constant. He talked on this topic at the 1927 Solvay conference on physics. Einstein, in the audience, was unimpressed. He famously responded *Vos calculs sont corrects, mais votre physique est abominable.*

When Hubble announced the discovery of the recession of the galaxies in 1929, Einstein realized that the expansion of space would stop its collapse. There was no more need for Λ, which Einstein regretted. He is quoted by George Gamow in his posthumously-published autobiography as saying that Einstein *remarked that the introduction of the cosmological term was the biggest blunder of his life.*

By 1931, Einstein had become reconciled with the expanding universe model. Travelling with Lemaitre on a lecture tour to California, Einstein commented on Lemaitre's presentation: *This is the most beautiful and satisfactory explanation of creation to which I have ever listened.* But Lemaitre would never agree about the notion of creation: for him, the equations pointed to a physical beginning of the universe, and said nothing about any prior event.

It was Lemaitre in 1933 who interpreted Einstein's cosmological constant in terms of energy of vacuum fluctuations, with an equation of state $p = -\rho c^2$ [17]. But his proposal was shelved for some 70 years, pending evidence for the acceleration of the universe discovered in 1998. Lemaitre realised that energy can accelerate if it is negative. It is the opposite of a pressure, more like a tension. Lemaitre interpreted the cosmological constant as the energy of the vacuum, which turns out to be negative and hence accelerates the universe.

The next step, confirmation, took 65 years. Two independent teams of mostly US astronomers were involved. Both found that distant supernovae were about 20% fainter than they should have been. Supernovae, or exploding stars, of a certain type were believed to be incredibly precise beacons, exploding via the decay of half a solar mass or so of radioactive nickel created in the precursor stars' collapse. Because of their immense luminosity, such supernovae are visible far away in distant galaxies. Acceleration of the universe was the likely culprit.

The conclusion was not easy to justify. One had to be certain that the selection of distant supernovae gave objects identical to nearby ones. Then one could justify the standard bomb hypothesis, verified for nearby examples. Other explanations were soon systematically eliminated. For example, interstellar or intergalactic dust could be responsible for the diminution in light. The more distant supernovae were produced by a host population of stars that was younger and more metal-deficient. This could lead to appreciable evolutionary and environmental differences between nearby and remote supernovae.

One conclusion remained. The systematic dimming of the otherwise identical distant supernovae could only be explained if the universe was accelerating. There was indeed a non-zero value of the cosmological constant. For this discovery, the Nobel prize in physics was awarded in 2011 to Saul Perlmutter, Brian Schmidt and Adam Riess for the discovery of acceleration of ther universe. The inferred positive cosmological constant is interpreted as the energy of the vacuum, which turns out to be negative, as Lemaitre had advocated. Vacuum energy is identified, in Einsteins equations of gravity, as the cosmological constant. It acts against gravity. The new cosmology looks like

$$\text{CURVATURE OF SPACE} = \text{MASS /ENERGY} - \Lambda.$$

Einstein's erstwhile demon was resuscitated and this means that two-thirds of the mass-energy in the universe reappears as Λ and acts to counter gravity. The discovery of acceleration was a revolution in cosmology. It meant that two-thirds of the mass energy in the universe could not be accounted for in classical physics. For this new component, it soon became clear, was present in the vacuum, where there was no matter of any consequence. It is an energy field that does not possess any attractive gravity of its own, rather it is repulsive.

3.3 Where Next with Dark Energy?

In the past two decades since discovery, the acceleration has been measured with ever increasing precision. However the energy scale corresponding to the cosmological constant is smaller than any natural scale from particle physics by some 120 factors of 10. There is no physics explanation for the scale of the cosmological constant, nor is there any convincing prediction of any deviations. Hence searching for deviations from a constant may not yield a guaranteed return. Certainly, improved precision is very welcome. The search for increased precision enshrines much of experimental physics. It is always good to raise the hurdles in parameter space, but it is even better if one has a precise destination in mind.

Indeed since discovery in 1998 to now, the errors on Λ have decreased by an order of magnitude, from $\sim$10% to $\sim$1%. All results to date converge on the equation of state $p = w\rho$ with $w = -1$. Several new surveys are being developed, including DESI, EUCLID. and LSST, that will improve the precision of dark energy measurements by up to another order of magnitude. There is hope that if we detection deviations from $w = -1$, we will discover new physics that will lead us to an improved understanding of the observed value of Λ. But there is absolutely no guarantee, or even robust prediction, of such a finding. Hence it would be wise to have an alternative strategy for advancing cosmology. Our overlying goal is an improved understanding of our cosmic origins. To approach the beginning of the universe, this strategy necessarily must involve testing inflation.

4 Inflation

We accept inflation because it accounts for a number of fundamental questions. Reasons for our trust in inflation include explanation of the size of the universe, the flatness of space, and the seeds of galaxies as quantum fluctuations. Observational confidence is enhanced by demonstration of the infinitesimal temperature fluctuations in the microwave background as approximately gaussian. The Planck map of the CMB all-sky temperature fluctuations is described by only 6 numbers. These are the densities of dark matter, dark energy and baryons, the Hubble constant, the spectral index of the temperature fluctuations, and the curvature of the universe as manifested by the angular size of the last scattering surface. Inflation is certainly supported by the astronomical evidence.

However critics maintain that it is built on poorly defined foundations, as there is no theory of quantum gravity. Inflation is the best theory we have, and merits our trust. It describes the universe back to 10^{-36} second after the beginning, an amazing accomplishment since the early years of modern cosmology. A similar, but much, much smaller, energy field to that postulated for inflation reasserted itself, a few billion years later. This is why the cosmological constant should not come as a total surprise. There is a key difference, however. The scale of inflation is a natural scale in terms of particle physics. It corresponds, at least in order of magnitude, to the energy scale where the fundamental forces of nature, apart from gravity, first became unified. Finding deviations from a constant Λ would help us understand its origin, most likely at the epoch of inflation. There is no robust prediction of this. Perhaps the most promising approach is quintessence, as this explains, or more precisely, phenomenologically models, the current magnitude of Λ while making testable predictions.

4.1 The Measure of Inflation

The weakest point in all models of inflation is in the initial conditions. There is no measure that defines the space of possibilities [18,19]. Hence predictions are meaningless. To assert that the flatness of the universe or the value expected for Λ is predicted by inflation is a meaningless prediction in the absence of a measure. It becomes a subjective game in which inflation proponents forcefully assert prejudices that lead to questionably reliable outcomes. This applies especially to "predictions" of the multiverse, accordimg to which one universe is as equally plausible as any other in the string theory prediction of 10^{500} Calaubi-Yau compactified manifolds, or even 10^{272000} if the counting of compactified manifolds is generalized to other string theories [20].

Let me focus on Λ. We have two choices. We can accept Λ as just one more number or constant of nature that describes the universe. Perhaps one day, an ultimate theory will emerge, sometimes dubbed the Theory of Everything, that will specify the fundamental constants of nature, and in particular the value of the cosmological constant. Or we can accept one of the precepts of string theory, that the constants of nature may vary according to the precise 3-dimensional universe that results after compactification from ten spatial dimensions. We then appeal to inflation theory in its generic form, which experts believe predicts an infinite number of separate universes. According to this argument, inflation is eternal, and it can occur at any time and place.

Our basic problem is that we cannot prove the theory of inflation is correct. In fact we cannot prove any theory is true. Falsifying a bad theory may be impossible. One can add parameters without limit. But we should be able to falsify a good theory. What I mean by a good theory is one that accounts for, explains and interprets the facts, as confirmed by observation and experiment, is consistent with other accepted good theories, and is parsimonious, that is it has a minimal number of free parameters to explain data with many more degrees of freedom. Inflation is a good theory by this definition. Our best way forward to understanding the beginning of the universe may be to probe falsification of inflation, for which there is a robust prediction, that is difficult to attain but nevertheless feasible, as I argue below.

4.2 The Future, According to Inflation

The universe is just beginning to accelerate and will continue expand ever more rapidly. The acceleration began about seven billion years ago, that's when we detect the first deviations from Hubble's law of the linear relation between distance and redshift or recession velocity. In 140 billion years from now, our universe will have accelerated so much that our Milky Way and its close companions will be alone in the visible universe. All more distant galaxies will have accelerated away outside our horizon.

Only if the acceleration were to stop would we renew contact with our neighbours. This happened with the first episode of inflation when the seed fluctuations destined to form galaxies reentered the horizon at a much later epoch. But we have no reason to believe we will be similarly saved in the distant future. Of course the earth will be destroyed in 5 billion years when the sun dies as a red giant, so there will be no terrestrial witnesses in the remote future.

Though it is encouraging to think that dwarf stars, some 20 percent of the mass of the sun, will live for a trillion years or more. Such stars are far more common in the Milky Way than are solar-type stars. And many are believed to possess habitable planets, with a triplet of exoplanets being recently discovered around the nearby dwarf star TRAPPIST 1, some 29 light years away. So cosmology will be very different for any civilization surviving in the distant future.

4.3 Where Next with the CMB: To B or Not to B?

The main target of current and future CMB experiments is the CMB polarization signal induced by gravity waves at the onset of inflation. There are enormous efforts underway to set stronger limits on the tensor-to-scalar ratio, the quantitative measure of the ratio of the primordial amplitude of the B-mode (or shearing) polarization component due to gravity waves to the scalar (or compressive) mode of CMB temperature fluctuations associated with the density fluctuations that seeded structure formation. The Planck telescope, which currently sets the best limit of tensor-to-scalar ratio (r) of 0.07, had some 32 detectors in its HRI instrument, Current experiments in the high Atacama desert, at the South Pole, and on balloons have tens of thousands of detectors and aim at reaching $r \sim 0.01$. Future experiments, planned over the next decade will incorporate up to a million detectors and extend the reach in r by another order of magnitude.

Yet there is no guaranteed prediction of any signal. The predicted tensor amplitude depends on the energy scale of inflation, to the fourth power. Inflation could well occur at an energy of 10^{15} GeV or less, and the primordial polarization signal would be unmeasurably small.

The cosmic microwave background (CMB) temperature fluctuations probe Fourier harmonics to $\ell_{max} \sim 2000$ and hence to $N \sim l^2_{max}/2$, or a million modes, with limiting precision $1/\sqrt{N} \sim 0.001$. The state-of-the-art upper limits from the Planck satellite, expressed as a quadratic correction to the inflationary potential of around 10, the nongaussianity parameter f_{NL}, correspond to a predicted temperature fluctuation $f_{NL} \sim (n-1)\delta T/T$, where $\delta T/T$ is the measured temperature fluctuation amplitude. The generic prediction expected in essentially all inflationary cosmology models is just

the scalar index deviation from invariance [21,22], best measured by Planck. Lower by some three orders of magnitude, f_{NL} is around 0.01 in generic inflation. This is the ultimate probe of inflation, primordial non-gaussianity at a level a thousand times better than ever achievable in the CMB, because one can never get more modes than Planck measures. Inflation is an excellent phenomenological fit to large-scale structure and the CMB, but we urgently need to understand whether it actually occurred.

Galaxy surveys will help. The most ambitious goal is that of LSST, which should obtain some 20 billion photometric galaxy redshifts over the next decade. Other surveys, especially with spectroscopic reach, including DESI and PFS, will help reduce galaxy bias. One may hope that 10^8 independent modes would be available at a redshift of order unity. This will improve the precision of cosmology by an order of magnitude relative to the CMB, and allow a f_{NL} limit of order unity. But we still need to improve on this by two orders of magnitude in order to really confront cosmology.

5 Issues with the Big Bang

There are some questions that simply do not go away. Ever-improving data sharpens the debate, but the perennial questions remain to haunt us. I have previously addressed the matter of the nature of the cosmological constant, which is a theoretical question that has few directly observable consequences. Here I summarize several noteworthy observational issues that continue to confront the standard model of the Big Bang, and its principal accomplice, the theory of structure formation.

5.1 Lithium

Lets begin with the prediction of the cosmological lithium abundance. Primordial nucleosynthesis has been remarkably successful in accounting for the abundances of helium and deuterium. There predictions were of immense significance for cosmology. The concept of primordial nucleosynthesis and the requirement of a hot Big Bang played a key role in the prediction of the cosmic blackbody radiation. It provided the first predictions of the number of neutrino species, later greatly refined by analysis of the CMB temperature fluctuations. It provides a precise prediction of the baryon fraction, verified independently by CMB fluctuation data, most recently that of the Planck satellite.

The measured baryon fraction is responsible for the strong inference that non-baryonic dark matter is needed to fill the gap between the baryon density and the total density of dark matter. Dark matter is measured most notably by gravitational lensing over a wide variety of scales, from dwarf galaxies to that of the entire universe.

However there is one cloud over the successes of primordial nucleosynthesis. Lithium is overpredicted by a factor of about three [23]. Lithium, along with deuterium, is destroyed by stars, hence it provides a measure of the primordial abundance after correcting for any astration. Attempts to reduce the lithium abundance to the predicted level include neutron or more exotic particle injection during the first 10^4 seconds, stellar destruction by rotationally-induced mixing, and mass transfer of astrated matter

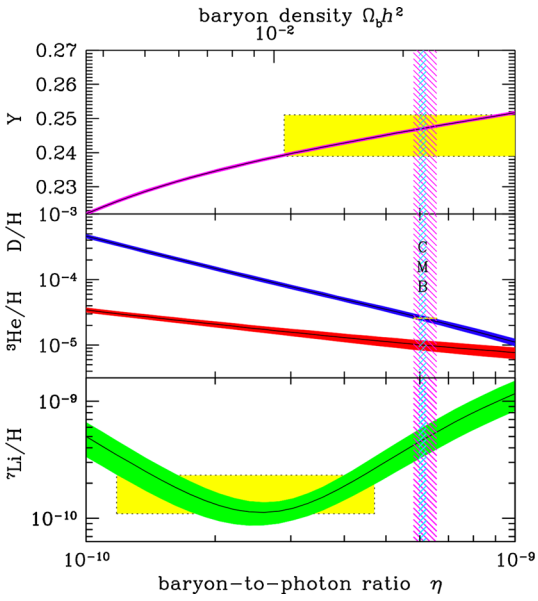

Fig. 1 Predicted abundances of BBN and observational constraints [23]

in a binary. Some relief in the tension is claimed, but there is no satisfactory resolution of the lithium problem (Fig. 1).

5.2 Hubble Constant Tension

The value of the Hubble constant has historically been a source of controversy. For decades, stars were found to be older than the inferred expansion age of the universe. Improved measurements are now fully consistent with stellar ages, but there is a systematic offset between the value of H_0 obtained from supernovae in the local universe, and the value measured with the Planck satellite via the power spectrum temperature features in the cosmic microwave background, effectively at an age 380,000 years after the Big Bang. The local value is $H_0 = 73.2 \pm 1.2$km/s [24,25], whereas the distant value from Planck is $H_0 = 66.9 \pm 0.6$km/s. The source of the discrepancy is unknown. If not due to systematic errors, the difference might be a pointer towards new physics (Fig. 2) [26].

5.3 Dwarf Galaxies

Galaxy formation theory has made enormous progress in recent years, with advent of sophisticated numerical codes that address ever-increasing efforts to tackling unresolved star formation physics. An improved understanding of how stars form in diverse conditions is a key obstacle in preventing robust cosmological results from being obtained. This problem is present for massive galaxies, where supermassive black

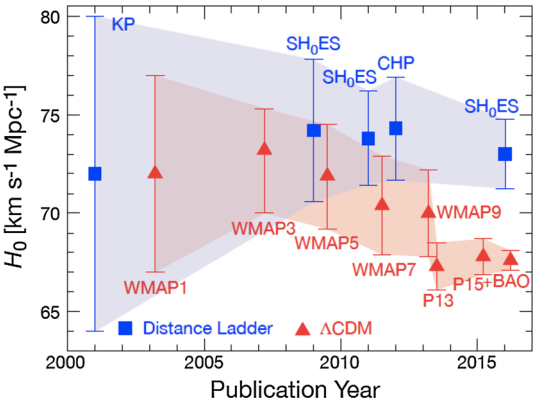

Fig. 2 History of Hubble constant measurements [26]

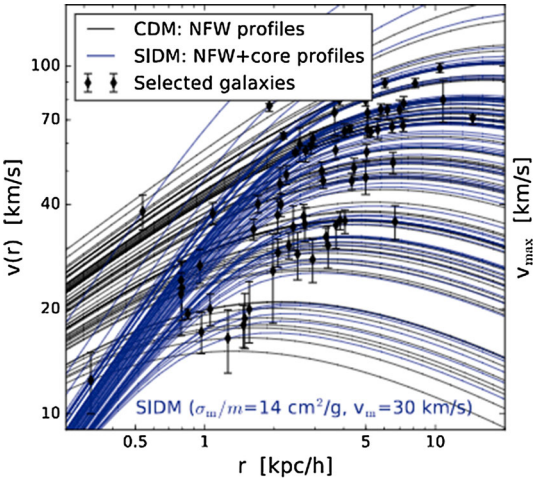

Fig. 3 Dwarf galaxy cores from rotation curves and fits for CDM and SIDM [29]

hole growth and feedback is unresolved, and equally for dwarf galaxies, where the observed frequency and core structure is poorly understood.

Most attention has been paid to dwarf galaxy issues, where particle physicists have flocked into the field by trying to remedy apparent astrophysical anomalies by inventing new particle physics models for the dark matter (Fig. 3).

Here are the key problems. Cold dark matter has had immense success in accounting for the large-scale structure of the universe, culminating with its incorporation into the standard ΛCDM cosmology that successfully accounts for the Planck all-sky map of temperature fluctuations, containing more than a million independent modes, with only six numbers. However CDM predicts many thousands more dwarf dark halos per host galaxy than are seen in the form of dwarf galaxies. It also predicts that the dark matter density profiles are generically spiked, with NFW cores proportional to $1/r$, whereas dwarf galaxy rotation curves almost universally require cored DM profiles.

These generic predictions, however, are for pure CDM simulations. When baryons are added, the situation becomes complex. Firstly, baryons must cool and dissipate

energy in order to form stars. This only happens in dwarfs with masses more than about $10^6 M_\odot$. These potential wells are so shallow that the first generation of stars, which includes some stars above $\sim 10 M_\odot$ that explode as supernovae, wreaks so much havoc in the form of feedback on the ambient gas that the efficiency of future star formation is severely curtailed. In this way, one produces many fewer dwarfs containing visible amounts of stars, and most of these are low surface brightness and detected with difficulty. This more or less solves the dwarf abundance problem [27].

The cusp/core problem is more difficult to resolve. Here one appeals to repeated episodes of bulk gas motions driven by successive episodes of supernovae with intermittent periods of gas accretion. The dynamical friction exerted by the dark matter on the gas clouds results in enough kinetic energy loss by the gas to heat up the dark matter and destroy the dwarf cores in the DM distribution [28]. It remains to be seen if this solution does not have .other byproducts such as overproduction of chemical elements, but it seems so far to be a promising solution. However the recent discovery that our MWG has a core may bring bad news, as the dwarf heating mechanism cannot work in the deeper gravity potential of our MWG. Here perhaps tumbling and buckling of the MWG bar does the dirty work of heating the DM cusp.

One final prediction that confronts data and is unresolved is the "too big to fail" problem. According to this, too many dwarfs are made in the mass range $10^8 - 10^9 M_\odot$ whose star formation is not suppressed by SNe and greatly outnumber the observed dwarfs in this mass range.

One's impression is that all of these are teething problems undergone by the new-born dwarf galaxies in confronting the realities of baryonic physics. As the phenomenology of star formation and feedback is improved, one expects to generate more realistic dwarfs. Of course the topics of magnetohydrodynamics and cosmic ray physics are only now beginning to be addressed. It seems reasonable to conclude that there are sufficient degrees of freedom that it will be very hard to use dwarf characteristics as a rigorous probe of CDM. Rather, improved observations at progressively higher resolution that elucidate the earliest phases of fragmentation, accretion and feedback are likely to help resolve many of the uncertainties. Improved numerical resolution will be essential, but it may be that observations hold the key to progress, as has seemed to emerge from the recent studies by ALMA of star-forming galaxies.

Compare this viewpoint with that advocated by the particle physicists eager to justify dark matter candidates, an especially urgent task as we fail to find the generic WIMPs predicted by SUSY. Warm dark matter, normalized to give the required number density of dwarfs and constrained by the Lyman alpha forest constraints, and characterised most eloquently by sterile neutrinos of mass $\sim$ 3keV, fails to account for kpc-scale dwarf cores, as also does a mixture of cold and hot dark matter. This has opened the way for more exotic dark matter scenarios, involving for example strongly self-interacting dark matter [29] or ultralight axions [30]. One has to raise the question of whether resolution of the dwarf galaxy issues justifies the introduction of new physics. Nor is this step especially favored by the fact that the various dark matter physics proposals narrowly address only a selective subset of the dwarf galaxy "anomalies".

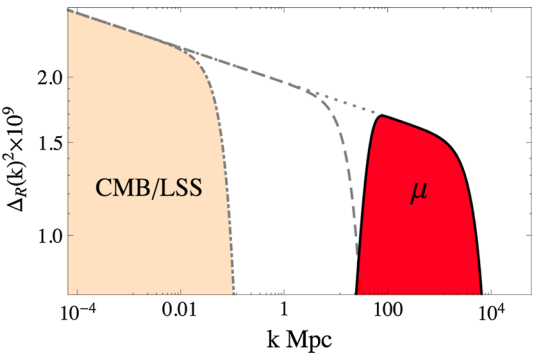

Fig. 4 The ultimate spectral distortion signal [34]

5.4 Why is the CMB a Perfect Blackbody?

The cosmic microwave background fits a blackbody spectrum to high significance, with $T = 2.728 \pm 0.004K$ [31]. At some point, this blackbody spectrum must reveal distortions due to energy injection in the early universe. Because the number of photons per baryon is $1.6 \pm 0.1 \times 10^9$, the naive estimate of expected distortions is of order one part in a billion. Small spectral distortions are characterized by two parametrised deviations from the blackbody spectrum, depending on whether photon energy injection processes add non-thermal energy, while failing to fully thermalize the photons yet conserving photon number. The first condition occurs after $\sim 3 \times 10^6$ seconds, the epoch of last thermalization via photon creation due to double Compton scattering and bremsstrahlung. Between the thermalization epoch and the epoch of matter domination at $\sim 10^{12}$ seconds, blackbody photons undergo rapid Compton scattering, with energy equilibration between electrons and photons but maintaining photon conservation, resulting in a chemical potential-like (or μ) distortion. This is a gray body distortion of a blackbody spectrum. At later epochs, especially after decoupling, infrequent Compton scattering plays a role and the result is also photon number conserving but now revealing up-scattering of photons from the Rayleigh-Jeans to the Wien spectral regions. This is called a y distortion. The expected level of spectral distortions from damping of dwarf galaxy-scale fluctuations, characteristic of the μ era, is of order 10^{-9} for minimal running of the scalar fluctuation spectral index [32], whereas the FIRAS upper limit on μ is 10^{-4} (Fig. 4). Another guaranteed signal that becomes visible at this sensitivity would be recombination lines from hydrogen and helium from the recombination epoch [33].

6 The Mysterious Role of Supermassive Black Holes in the Early Universe

Black holes involve extreme physics. Understanding their interface with the visible universe requires pushing our modeling and observations to span a dynamic range of unprecedented extent. The outlook is bright for observational advances but less clear

for the progress of simulations. Penetrating the interface of AGN activity and star formation is our current challenge.

We see very massive black holes in the early universe. They are quasars, the most luminous objects in the universe, visible because they accrete interstellar gas. The gas heats up, glows in x-rays, and the black hole grows. The mass doubling time for a typical massive black hole is about fifty million years. This has two implications. Firstly, we expect to find massive black holes a few hundred million years after the Big Bang. And indeed black holes as massive as ten billion solar masses are found when the universe was a billion years old. Secondly, we need seeds, smaller black holes, to accrete gas or merge to make the massive ones. If we started from typical stellar black holes, around ten or twenty solar masses, there would not be time enough to grow the monsters we find, at least in the standard approach to black hole growth.

6.1 Black Hole Seeds

How the seeds are formed is somewhat of a mystery. We believe that they are formed from the first generation of million solar mass clouds formed after the Big Bang. These clouds were chemically pure, no heavy elements yet formed in supernovae. This means that cooling occurs by hydrogen atom excitations. Electrons jump from one atomic orbit to a higher energetic level, due to a collision with another atom or electron, then the energy level deexcites by emitting a photon. This is how cooling occurs. Atomic cooling provides a powerful channel for losing energy, and guarantees that the clouds will undergo direct collapse to form black holes, typically of ten thousand solar masses.

The only obstacle is that too much cooling may occur. Hydrogen molecules are a catalyst for cooling as they are more easily excited than hydrogen atoms. Trace amounts of hydrogen molecules can form and encourage fragmentation into stars. Moreover once a small black hole forms, it generates enough x-rays to destroy any molecules in its vicinity, and grows to form the first generation of intermediate mass black holes. These are expected to be present in all dwarf galaxies [35].

It is these intermediate mass black holes that undergo catastrophic accretion as dwarf mass halos merge to eventually form supermassive black holes in massive galaxies. Every galaxy most likely has a lurking giant in its center. Long ago, they played a crucial role in galaxy formation. Star formation in the forming galaxy was quenched. The radiation from the central black hole limited the amount of infallng gas that fragmented into stars. Long ago, galaxies formed stars prolifically, while today, most massive galaxies are not forming any stars.

The black hole eventually reaches maturity. If the black hole were to get any more massive, the gas is mostly blasted away, thereby maintaining a universal ratio of black hole to stellar mass. Remarkably, we observe such a relation from the smallest to the most massive galaxies, telling us that old stars and central black holes self-regulate and formed early in the history of galaxies [36].

6.2 Black Hole Outflows

Supermassive black holes at the centers of galaxies are generally dormant giants. They were hyperactive in their youth, long ago. They become reactivated only when fresh

gaseous fuel is provided. This may happen after billions of years when a merger occurs with a nearby galaxy. Merger fuelling is often observed when we peer into the distant universe with our largest telescopes, and observe galaxies in their youth.

Black hole activity can be a dramatic event. We occasionally see powerful jets of plasma that drive giant radio-emitting lobes. These are vigorous outflows that collide with and eject interstellar gas clouds into the surrounding circumgalactic medium, where the gas eventually cools and ends up as intergalactic clouds, to enrich new generations of galaxies.

Powerful radio jets are produced in the black hole ergosphere by release of energy arising from the rotation and winding-up of magnetic fields, The supermassive black hole acts like a gigantic flywheel., with the spin of space providing the momentum. Gripping the flywheel is presumably done with powerful magnetic fields, thought to be omnipresent. The fields initially have a dipole pattern but soon tangle up because of the differential spin and turbulence. Magnetic reconnection releases huge amounts of energy that is initially channeled along the axis of rotation, emerging as a collimated jet and continuing for thousands of parsecs.

6.3 The Supermassive Black Hole: Star Formation Connection

Perhaps the most intriguing result of studies with the largest telescopes show that intense black hole activity and extreme rates of star formation occur in the same objects. They are also driving powerful outflows of gas. This is one of the biggest mysteries of structure formation. Why are these phenomena coinciding at early epochs, especially reaching extreme rates of star formation and mass loss that are rarely seen in the absence of supermassive black holes?

Is black hole feeding, with its immense energy release, a consequence of star formation, with the stellar debris feeding the black hole? If so, this fails to address the extreme intensity of star formation seen in many of the most distant galaxies. Both AGN activity and star formation bursts may be the collateral damage from a merger between two galaxies that helps shed angular momentum and allows gas to pour into the vicinity of the most massive central black hole? The galaxy merger provides the gas supply responsible both for star formation and supermassive black hole fuelling. In a merger, gas cloud orbits are perturbed, and some clouds are directed into the capture zone of the supermassive black hole, to refuel its activity. The increase in gas mass stimulates star formation at the same time. Or have the powerful outflows from the black holes compressed nearby clouds and triggered an intense burst of star formation in the surrounding clouds? Of course the outflows show that the star formation rate is being quenched. But this may have been preceded by a phase of triggering that induced the winds.

The jury is out on any resolution to these questions. New observations with instruments such as the ALMA radio interferometer, now taking data at high angular resolution, and the James Webb Space Telescope, to be launched in 2020 and able to take exquisite images in the near infrared, will eventually elucidate many of these issues.

6.4 Gravitational Waves

As for the origin of supermassive black holes, the ultimate window on building and observing them is gravity waves, detected from stellar mass black holes for the first time in 2015. Gravitational wave experiments are being planned to search for traces of the elusive signatures of supermassive black holes. These include the approved ESA space experiment LISA, a three satellite interferometer with million km-long arms. Many of these supermassive black holes are in binary pairs and will orbit together and eventually merge, emitting gravity waves as they do. The LISA window includes mergers of black holes in the mass range $10^4 - 10^7 M_\odot$ [37].

This will provide the ultimate proof. of the existence of supermassive black holes. We may have to wait for LISA, currently scheduled for launch in 2034, before our understanding of the formation of supermassive black holes is ultimately achieved.

6.5 What if the Black Holes were Primordial?

Black holes could be produced very early in the universe. Perhaps the earliest time is at the Planck epoch, some 10^{-43} second after the Big Bang. This is the epoch of quantum gravity, when quantum gravity forces match gravitational forces that at later epochs are incredibly weak on particle scales. One natural component of the quantum foam that fills the universe is tiny black holes of Planck mass, that is of a particle whose minimum size in general relativity, the Schwarzschild radius of a black hole of mass m, or $2Gm/c^2$, is equal to its minimum size in quantum theory, defined to be such that a smaller particle is wave-like. Here G is Newton's constant of gravity and c is the speed of light. The latter scale is the Compton wavelength, or $\hbar/mc$, where $\hbar$ is Planck's constant (more correctly, Planck's constant divided by 2π). Set the two scales equal, and we obtain 22 micrograms as the mass of a particle in the domain where gravity meets quantum theory. This is the Planck mass, the maximum mass of a point-like particle that is allowed by nature. At the Planck time, the universe was a quantum foam of appearing and disappearing short-lived Planck mass black holes. As the universe expands and cools, these all promptly disappear.

More massive black holes live longer and should survive until today as early universe relics [39]. And there is no reason why these would not be produced as well from the ubiquitous density fluctuations that are generated by inflation some 10^{-36} second after the Big Bang. It is the long wavelength end of the spectrum of these fluctuations that form the galaxies, and there is every reason to believe that there would be fluctuations present primordially on much smaller scales.

Primordial black holes are very much an option to play a role in the contents of today's universe. This is because once formed, their contribution to the dark matter of the universe grows with time. Most of the early universe is radiation, and this redshifts away, since not only does the number of photons per unit volume diminish as the universe volume expands, but the typical energy of a photon, which is inversely proportional to its wavelength, also decreases as its wavelength redshifts with the expansion. However the mass of non-evaporating primordial black holes is unchanged after birth, indeed it can only grow if they accrete matter. So the mass density of

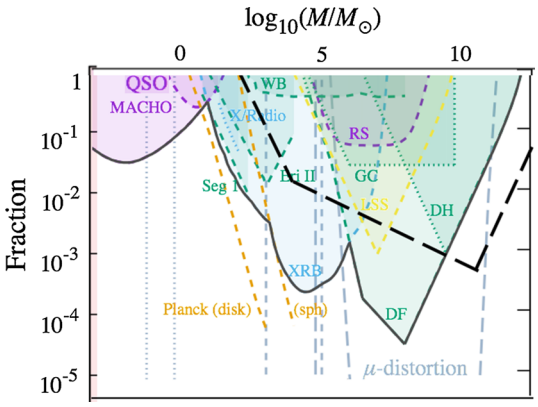

Fig. 5 Constraints on the fraction of the dark matter in massive PBHs from a variety of lensing, dynamical and accretion constraints [38]

primordial black holes acquires a huge boost compared to the energy density of the early universe. A very rare production rate can have huge consequences today, in particular for the dark matter content of the universe [40].

Massive primordial black holes, with masses in the planetary to stellar mass range could certainly contribute to the dark matter content of the universe and would be hard to detect. For example, the MACHO/EROS microlensing experiments limited the mass fraction over the mass range 0.3-30 $M_\odot$ to be at most 10-40% of the galactic halo dark matter [41,42].

The LIGO events have revitalized our interest in stellar mass black holes, of around thirty solar masses. This is in part because such events were not predicted from astrophysical sources, although this situation rapidly changed post-discovery. Another motivation is that despite intensive efforts, the dark matter of the universe has not yet been identified. Tthere are many constraints on the contribution of primordial black holes to the dark matter density, many using gravitational microlensing as a search technique, But primordial black holes at a level of 10% of the dark matter are still possible over a wide range of masses. For intermediate mass black holes of mass $10^3 M_\odot$ a dark matter mass fraction of only 0.1% still allows important consequences for structure formation (Fig. 5).

7 The Ultimate Limits of Cosmology

To proceed to robustly test inflation, we need to search for primordial non-gaussianity. We need to improve on CMB constraints by up to 3 orders of magnitude. The only solution is to dramatically increase the number of independent modes by targeting the many hydrogen clouds that are the precursors of the galaxies and mapping them in absorption against the cosmic microwave background. Radio interferometers arrays on the far side of the Moon are the unique way to work towards this goal, by probing the dark ages at redshifts of around 50, when the best probe of cold hydrogen clouds, the precursors of the galaxies, is the 21cm line of atomic hydrogen which however is redshifted by the expansion of the universe to the decametric band. This is the only means of acquiring enough independent modes on the sky.

7.1 Lunar Radio Astronomy

Only a lunar low frequency array can meet this challenge by exploring the dark ages where the number of independent modes on the sky opens up dramatically. Here is why. The CMB has a million modes, since beyond a Fourier harmonic $\ell \sim 2000$, strong damping occurs of the primordial fluctuations. Galaxy surveys of billions of galaxies potentially have a hundred million independent modes. But 21 cm surveys in the dark ages open up the building blocks of galaxies, and there are a million such hydrogen clouds per typical galaxy, to trillions of observable modes, providing some two orders of magnitude potential improvement in precision over large-scale galaxy surveys.

The caveat: we have to catch the precursor clouds in absorption before the first stars formed, and that means a redshift beyond 30, and well after last scattering, so that the hydrogen, and its spin temperature in particular, is cooler than the CMB. In the redshift range 30-80, we can see pristine clouds in absorption against the CMB. This is the promised land. But the challenge is radio astronomy at or even below 30 MHz. Only the far side of the Moon offers a suitably radio-quiet environment in the entire inner solar system to achieve the required sensitivity for a lunar radio array.

There is only one option to truly enhance the accuracy of future cosmology experiments. 21 cm hydrogen line probes of the dark ages provide the ultimate precision by sampling Jeans mass neutral hydrogen gas clouds at very early epochs and still in the linear regime. These clouds are predicted to be millions of solar masses, there are millions of such precursors per massive galaxy. If we could detect them early on, one could improve cosmological precision by a factor of 100 to 1000. This is feasible with a sufficiently sensitive radio interferometer.

These clouds have spin temperatures colder than the cosmic microwave background (CMB), when we observe them before any stars, galaxies or quasars have formed, at very high redshift $z \sim 30$ to 50. Using the unique advantage of redshifted 21 cm line tomography but at observed frequencies below 50 MHz, one can approach the very large number of independent modes, up to a trillion or more , that would allow 1000 times better precision for f_{NL} than attainable with the CMB even at Stage IV (by 2025). It is a clean probe: at these early epochs, there are no stars or other sources to heat up the gas.

Terrestrial telescopes, even as large as the Square Kilometer Array (the low frequency phase 2 to be completed by 2025), are plagued by man-made interference and the earth's ionosphere at these cosmologically obligatory low frequencies. The far side of the Moon offers a unique environment, the most radio-quiet in the inner solar system. It is our best option for realistically exploring the one guaranteed prediction of inflationary cosmology, primordial non-gaussianity. Only failure to find such evidence can falsify inflation.

There is immense pressure developing from the international space agencies, including CNSA, ESA and NASA, to develop bases on the Moon. It would be highly desirable to include science projects, most notably telescopes, in lunar exploration. Over the next few decades, we should be able to exploit the potential offered by the Moon as the ultimate setting for the development of cosmology. The only site capable of reaching the needed sensitivity at very low frequencies is a radio array on the far side of the

Moon. One will need novel techniques for building a sufficiently fast and sensitive dipole interferometer. Design will need to be guided by theory. Science goals will include searching for spectral running, inferred from the Planck results at large angular scales, and the three-point correlation functions, the latter being non-zero only if there is primordial non-gaussianity. The computing issues for correlating the antennae are challenging and may need to use lunar orbiting relay satellites. An initial goal as a pilot project that should be feasible simultaneously with lunar village construction will be an order of magnitude improvement on anything ever achievable with CMB or galaxy surveys, terrestrial or even in space.

This would provide a factor of 10 improvement on existing CMB limits on primordial non-gaussianity from Planck. No CMB experiment can attain this goal because of the limited number of modes. Only future galaxy surveys can do this, but in the relatively nearby universe. The lunar low frequency radio array will illuminate the dark ages by extracting precision cosmological parameters at this early epoch, and the cosmological leverage will complement projects like EUCLID and WFIRST in space, or LSST and DESI on the ground, that operate at much lower redshifts, of order unity. Later refinements could nclude full deployment of an optimally filled lunar array over up to 100km in diameter that will bring us towards the ultimate goal of two orders of magnitude improvement on all future galaxy surveys, and within the range of robustly testing inflation. The far side of the moon provides the only sites where one might build a telescope capable of studying the dark ages, prior to $z \sim 30$, before any galaxies had formed.

7.2 Lunar Infrared Astronomy

There is more for the Moon. Another goal in cosmology will be to detect the first stars, galaxies and active galactic nuclei in the Universe at the end of the dark ages. Many space telescopes, including WFIRST and LUVOIR, complemented by other telescopes on the Earth, such as E-ELT and SUBARU-PFS, are being planned. There is no guarantee that their reach will be sufficiently powerful. We need to look ahead to the future, and develop lunar IR astronomy in order to provide orders of magnitude improvement in sensitivity and in angular resolution on all currently feasible telescopes, whether terrestrial or in space. The goal would be to use the ultimate power of IR telescopes and interferometry on the Moon to help unlock the mysteries of the dominant components of the Universe, dark energy and dark matter and to map out the first objects in the Universe.

The Moon provides a unique platform for constructing futuristic telescopes. Lunar polar sites have unique potential for infrared astronomy. Current terrestrial telescopes are limited in size to ∼40 m, because of mechanical issues. A 100m infrared (or optical) telescope on the Moon is perfectly feasible. The seismological stability and lack of atmosphere are unique advantages that motivate detailed study in the context of the discovery of permanently dark and cold lunar sites, the potential availability of continuous solar power in the proximity, and the likely presence of extensive amounts of ice, along with the renewed international emphasis on development of a lunar base. The cold polar craters on the moon offer an environment beyond anything achiev-

able on the Earth or foreseeably in space: areas of 10 km or more in size, thermally stable at 30K, and within reach of perpetual solar power from the crater rims. Seismically stable platforms, unconstrained by atmospheric opacity, these permanently cold and dark polar craters could host 100m scale infrared telescopes or even 10km scale interferometers able to form high resolution images at unprecedented sensitivity.

The potential science reach includes the detection of the first galaxies and the physics of accreting supermassive black holes and its nearby surroundings (dust torus, broad emission line region), and the astrochemistry of the first episodes of chemical enrichment of the universe. There would be unparalleled opportunities for tackling anthropic issues, including direct detection, imaging and spectroscopy of exoplanets and their atmospheres, analyzing their morphology and searching for biosignatures. Lunar infrared interferometry with very large telescopes would probe our cosmic origins and evolution.

8 Summary

Future science returns must address the most important problem in cosmology, namely understanding how the universe began, and equally shed light on one of the key questions in cosmogony, namely is life unique to our local environment? These questions are intimately related, and provide the unifying basis for profound multidisciplinary discussions of broad interest to humanity on topics such as whether we began from something or nothing, whether the universe is finite or infinite, whether structure formation began via the first stars or via the first massive black holes, whether earth twins have biosignatures in their atmospheres, and whether anthropic arguments, based on our exoplanet searches, play a role in cosmology.

One noteworthy development in tackling the most intractable of these problems has been the emergence of philosophy of science as a tool in providing new approaches to seemingly intransigent questions in cosmology. Perhaps its major and most controversial contribution has been to argue that a non-empirical framework may be a substitute for experimental probes of physics theories that are notoriously difficult to test empirically, such as string theory or the origin of the cosmological constant [43].

I have argued that this approach challenges the integrity of physics [44]. Let me lay down the essential requirements of a "good" theory. I will use the multiverse "solution" to the cosmological constant problem, pioneered by Weinberg [45], as an example of a "bad" theory. According to Polchinski, [46], *there are four arguments for the multiverse: the failure of conventional methods for understanding why the cosmological constant is not large, the success of environ- mental theories for doing so, the successful prediction of the nonzero cosmological constant, and the string landscape.*

My principal gripe here is with the notion of a successful prediction. Let's face it, Weinberg's prescient prediction, based on the necessary requirement of forming galaxies, is used to argue that Λ is environmentally bounded, as it is allowed to vary in space and time. But what does "successful" mean? This prediction fails as an upper bound by two or three orders of magnitude if we accept the measured value of primordial density fluctuations via the CMB, bearing in mind that galaxy-scale

fluctuations are only weakly constrained. Few physicists would accept such a weak prediction as conclusive evidence for a theory.

But it gets much worse. This is the case only if Λ is positive, acting against gravity. by no means generically guaranteed in string theory. For example, the accepted solution to the mass hierarchy problem at the Planck scale requires an anti-de Sitter space-time [47], that is, a cosmic beginning with negative Λ. Once the sign of Λ is allowed to have anthropic freedom, the uncertainty in any prediction is degenerate with the amplitude of the primordial density fluctuations and vast domains of parameter space become anthropically compatible with the existence of galaxies. The concept of anthropic pressure, pushing Bayesian constraints from below to give a non-zero value of Λ, must be thrown away, and there is no longer any prediction worthy of the name.

Here is what I expect a physics theory to surmount in order to be a "good" theory, that is one that we can trust. There are four stages, all indispensable. These are credibility, explicability, predictability and falsifiability. In particle physics, such theories abound, from quantum electrodynamics, the standard model of particle physics, supersymmetry, and more. These theories are sharp and precise. In cosmology, the situation involves many shades of grey: from Einstein gravity, dark matter, dark energy, inflation to string theory and beyond. We need to develop more powerful probes than any currently envisaged in order to decide on a trustworthy explanation of the physical phenomena that we observe. We should not abandon hope, even for the multiverse, but I believe that there is a long road ahead before it can make the transition from metaphysics to physics.

Remarkably, development of the Moon as an astronomical site can potentially answer many questions that might otherwise seem beyond the range of current physics. I focus on testing inflation, via falsifiability or even verifiability. Admittedly, inflationary cosmology is a framework of many scenarios, but there is one generic prediction. This is primordial nongaussianity.

Consider a staged process, where a lunar far side radio array will initially improve on the desired sensitivity to primordial non-gaussianity over that achievable from the CMB by factors of 10 to 100, with the ultimate goal being the 1000 improvement that can robustly test the inflationary origin of the universe. Cosmogonically focussed programs on exoplanets and black holes could be developed for infrared interferometry and hypertelescope prototyping that are capable of significantly surpassing the reach of 10m class telescopes in space or 20–40 m very large terrestrial telescopes, before attaining ultimate goals that will improve resolution and spectroscopic sensitivity by orders of magnitude.

The space agencies need to include astronomical telescopes as a central part of a future lunar village. Now is the ideal time to develop this initiative so that cosmology can become an integral and fundamental component of lunar activities. Visionary plans for lunar telescope design, deployment and operation can unlock the greatest secrets of the universe, in a way that no terrestrial or even space telescope can do. The inspirational goals of cosmology and cosmogony can influence lunar base planning at this crucial moment in human history, when we are on the verge of developing manned bases on the Moon.

References

1. Zwicky, F.: Die Rotverschiebung von extragalaktischen Nebeln. Helv. Phys. Acta **6**, 110 (1933)
2. Rubin, V.C., Ford Jr., W.K.: Rotation of the Andromeda nebula from a spectroscopic survey of emission regions. Astrophys. J. **159**, 379 (1970)
3. Roberts, M.S., Whitehurst, R.N.: The rotation curve and geometry of M31 at large galactocentric distances. Astrophys. J. **201**, 327 (1975)
4. Sakharov, A.D.: The initial stage of an expanding universe and the appearance of a nonuniform distribution of matter. Sov. J. Exp. Theor. Phys. **22**, 241 (1966)
5. Cole, D.R., Binney, J.: A centrally heated dark halo for our Galaxy. MNRAS **465**, 798 (2017)
6. Portail, M., Gerhard, O., Wegg, C., Ness, M.: Dynamical modelling of the galactic bulge and bar: the Milky Way's pattern speed, stellar and dark matter mass distribution. MNRAS **465**, 1621 (2017)
7. Lee, S.K., Lisanti, M., Safdi, B.R., Slatyer, T.R., Xue, W.: Strong support for the millisec- ond pulsar origin of the Galactic center GeV excess. Phys. Rev. Lett. **116**, 051103 (2016)
8. Bartels, R., Krishnamurthy, S., Weniger, C.: Strong support for the millisecond pulsar origin of the Galactic center GeV excess. Phys. Rev. Lett. **116**, 051102 (2016)
9. Fragione, G., Antonini, F., Gnedin, O.Y.: Disrupted globular clusters and the gamma-ray excess in the Galactic Centre. arXiv:1709.03534 (2017)
10. Silk, J., Olive, K., Srednicki, M.: The photino, the sun, and high-energy neutrinos. Phys. Rev. Lett. **55**, 257 (1985)
11. Milgrom, M.: A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. Astrophys. J. **270**, 365 (1983)
12. Randriamampandry, T.H., Carignan, C.: Galaxy mass models: MOND versus dark matter haloes. MNRAS **439**, 2132 (2014)
13. Pointecouteau, E., Silk, J.: New constraints on modified Newtonian dynamics from galaxy clusters. MNRAS **364**, 654 (2005)
14. Paraficz, D., Kneib, J.-P., Richard, J., et al.: The Bullet cluster at its best: weighing stars, gas, and dark matter. A&A **594**, A121 (2016)
15. Bekenstein, J.D.: Tensor-vector-scalar- modified gravity: from small scale to cosmology. Philos. Trans. R. Soc. Lond. Ser. A **369**, 5003 (2011)
16. Ezquiaga, J., Zumalacárregui, M.: Dead Ends and the Road Ahead. arXiv:1710.05901 (2017)
17. Lemaître, G.: The beginning of the world from the point of view of quantum theory. Nature **127**, 706 (1931)
18. Linde, A., Noorbala, M.: Measure problem for eternal and non-eternal in ation. JCAP **9**, 008 (2010)
19. Olum, K.D.: Is there any coherent measure for eternal inflation? Phys. Rev. D **86**, 063509 (2012)
20. Taylor, W., Wang, Y.-N.: The F-theory geometry with most ux vacua. JHEP **12**, 164 (2015)
21. Maldacena, J.: Non-gaussian features of primordial uctuations in single field in ationary models. J. High Energy Phys. **5**, 013 (2003)
22. Cabass, G., Pajer, E., Schmidt, F.: How Gaussian can our Universe be? J. Cosmol. Astropart. Phys. **1**, 003 (2017)
23. Particle Data Group: (2016 and 2017 update), Rev. Part. Phys. Chin. Phys. C **40**, 100001 (2016)
24. Riess, A.G., Macri, L.M., Hoffmann, S.L., et al.: A 2.4% determination of the local value of the hubble constant. Astrophys. J. **826**, 56 (2016)
25. Mather, J.C., Cheng, E.S., Cottingham, D.A., et al.: Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. Astrophys. J. **420**, 439 (1994)
26. Freedman, W.L.: Cosmology at a crossroads. Nat. Astron. **1**, 0121 (2017)
27. Wetzel, A.R., Hopkins, P.F., Kim, J.-H., et al.: Reconciling dwarf galaxies with? CDM cosmology: simulating a realistic population of satellites around a Milky Way-mass galaxy. Astrophys. J. **827**, L23 (2016)
28. Pontzen, A., Governato, F.: Cold dark matter heats up. Nature **506**, 171 (2014)

29. Schneider, A., Trujillo-Gomez, S., Papastergis, E., Reed, D.S., Lake, G.: Hints against the cold and collisionless nature of dark matter from the galaxy velocity function. MNRAS **470**, 1542 (2017)
30. Schive, H.-Y., Chiueh, T., Broadhurst, T., Huang, K.-W.: Contrasting galaxy formation from quantum wave dark matter, ΨDM, with ∧CDM, using Planck and Hubble Data. Astrophys. J. **818**, 89 (2016)
31. Fixsen, D.J., Cheng, E.S., Gales, J.M., et al.: The cosmic microwave background spectrum from the full COBE FIRAS Data set. Astrophys. J. **473**, 576 (1996)
32. Cabass, G., Melchiorri, A., Pajer, E.: μ distortions or running: a guaranteed discovery from CMB spectrometry. Phys. Rev. D **93**, 083515 (2016)
33. Desjacques, V., Chluba, J., Silk, J., de Bernardis, F., Doré, O.: Detecting the cosmological recombination signal from space. MNRAS **451**, 4460 (2015)
34. Pajer, E., Zaldarriaga, M.: New window on primordial non-gaussianity. Phys. Rev. Lett. **109**, 021302 (2012)
35. Silk, J.: Feedback by massive black holes in gas-rich dwarf galaxies. Astrophys. J. **839**, L13 (2017)
36. Silk, J., Rees, M.J.: L1 Quasars and galaxy formation. A&A **331**, L1 (1998)
37. Amaro-Seoane, P., Audley, H., Babak, S., et al.: Laser Interferometer Space Antenna. arXiv:1702.00786 (2017)
38. Carr, B., Silk, J.: In press Primordial Black Holes as Generators of Cosmic Structures (2018)
39. Carr, B.J., Hawking, S.W.: Black holes in the early Universe. MNRAS **168**, 399 (1974)
40. Dolgov, A., Silk, J.: Baryon isocurvature uctuations at small scales and baryonic dark matter. Phys. Rev. D **47**, 4244 (1993)
41. Alcock, C., Allsman, R.A., Alves, D.R., et al.: MACHO project limits on black hole dark matter in the 1–30 Msolar range. Astrophys. J. **550**, L169 (2001)
42. Tisserand, P., Le Guillou, L., Afonso, C., et al.: Limits on the macho content of the galactic halo from the EROS-2 survey of the magellanic clouds. A&A **469**, 387 (2007)
43. Dawid, R.: The Significance of Non-empirical Confirmation in Fundamental Physics. arXiv:1702.01133 (2017)
44. Ellis, G., Silk, J.: Scientific method: defend the integrity of physics. Nature **516**, 321 (2014)
45. Weinberg, S.: Anthropic bound on the cosmological constant. Phys. Rev. Lett. **59**, 2607 (1987)
46. Polchinski, J.: Why trust a theory? Some further remarks. arXiv:1601.06145 (2016)
47. Randall, L., Sundrum, R.: Large mass hierarchy from a small extra dimension. Phys. Rev. Lett. **83**, 3370 (1999)