PHY418 Particle Astrophysics

Susan Cartwright

Contents

Chapter 1

Introduction

1.1 What is particle astrophysics?

Particle astrophysics, also known as astroparticle physics, is essentially the use of particle physics techniques, either experimental or theoretical, to address astrophysical questions, or conversely the use of astrophysical data to constrain theories of particle physics. Examples of the former include gamma-ray astronomy and the development of the theory of inflation as an outgrowth from Grand Unified Theories; an example of the latter is the use of solar neutrinos to measure neutrino oscillation parameters.

Particle astrophysics as a discipline in its own right is a relatively recent development, and the topics included under its umbrella vary from place to place. The journal *Astroparticle Physics* defines its subject matter as[1]

- High-energy cosmic-ray physics and astrophysics;
- Particle cosmology;
- Particle astrophysics;
- Related astrophysics: supernova, AGN, cosmic abundances, dark matter etc.;
- High-energy, VHE and UHE gamma-ray astronomy;
- High- and low-energy neutrino astronomy;
- Instrumentation and detector developments related to the above-mentioned fields

(a somewhat unsatisfactory definition since it includes "particle astrophysics" as a topic in its own right!). The Science and Technology Funding Council (STFC) defines particle astrophysics as "that branch of particle physics that studies elementary particles of astronomical origin, and their relation to astrophysics and cosmology" [2], but its description of the activities funded under this heading[3] includes gravitational waves, which do not seem to fit this definition. The 2008 and 2011 Roadmap documents of the Astroparticle Physics European Consortium (ApPEC) [4] define their subject as "the intersection of astrophysics, particle and nuclear physics and cosmology. It addresses questions like the nature of dark matter and dark energy, the physics of the Big Bang, the stability of protons, the properties of neutrinos and their role in cosmic evolution, the interior of the Sun or supernovae as seen with neutrinos, the origin of cosmic rays, the nature of the Universe at extreme energies and violent cosmic processes as seen with gravitational waves.” The table of contents of the 2011 roadmap includes, as chapter or section headings,

- charged cosmic rays;
- gamma-ray astrophysics;
- high-energy neutrinos;
- dark matter;
- neutrino mass measurements (direct and via double beta decay);
- low-energy neutrino astronomy;
- proton decay;
- dark energy;
- gravitational waves.

Despite the minor variations, a fairly coherent picture of particle astrophysics emerges from these definitions. Essentially, the core disciplines of particle astrophysics are

1. early-universe cosmology;
2. the physics of dark energy;
3. high-energy processes in astrophysics;
4. neutrinos;
5. dark matter.

I have omitted gravitational waves from this list, despite their inclusion by both the STFC and ApPEC, because neither their production nor their detection involves particle physics (essentially, this is a classical phenomenon described by general relativity). However, the special case of primordial gravitational waves, detected via the imprint they leave on the polarisation of the cosmic microwave background, does belong in particle astrophysics because of its relevance to early-universe cosmology.

In the rest of this chapter, we will briefly introduce each of the topics listed above. The rest of the course, however, will focus almost exclusively on the third and fourth items. Particle cosmology and the physics of dark energy will not be discussed in detail because they are very technical theoretical topics which would require a whole module to cover in adequate depth, while dark matter is covered in PHY326/426 Dark Matter and the Universe [5], and therefore will only be summarised here.

1.2 Early-universe cosmology

The temperature of the universe now, as measured by the cosmic microwave background, is $2.72548 \pm 0.00057\,\mathrm{K}$ [6]. Temperature scales as $(1 + z)$, where z is the redshift (see PHY306/406 Introduction to Cosmology [7]), so the early universe was much hotter than this, and hence had higher characteristic energies ($E \simeq k_B T$, where Boltzmann’s constant $k_B = 8.65 \times 10^{-5}\,\mathrm{eV\,K^{-1}}$). Big Bang nucleosynthesis (see PHY306/406 and PHY320 Nuclear Astrophysics [8]) takes place a few minutes after the Big Bang, at temperatures of order 10^9 K and energies of order 0.1 MeV. Energies above a few MeV, corresponding to times before 1 s or so after the Big Bang, are too high for nuclear physics and fall into the domain of particle physics, so early-universe cosmology is in many ways an application of theoretical particle physics, and is often referred to as *particle cosmology*.

1.2.1 Inflation

One of the first applications of theoretical particle physics to early-universe cosmology was the development of the idea of *inflation* [9, 10], which used the physics of Grand Unified Theories at very high energies ($E \sim 10^{16}$ GeV, $t \sim 10^{-35}$ s after the Big Bang) to drive a brief period of extremely rapid expansion (usually assumed to be exponential, $a \propto \exp(Ht)$, although a steep power law, $a \propto t^n$ where $n > 1$, will also work). Inflation was originally postulated to account for two observations which are otherwise difficult to reconcile with the classical Big Bang model: the fact that the universe is observed to have a flat (Euclidean) geometry, and the extremely high level of isotropy displayed by the cosmic microwave background (see PHY306/406 for further details).

Inflation also accounts for the small ($\sim 10^{-5}$) anisotropies of the microwave background, which arise from quantum fluctuations of the vacuum "frozen in" and expanded to macroscopic size by the rapid expansion. Inflation models predict that the spectral index of the fluctations, n, should be about 0.95, in good agreement with the fitted value of 0.9603 ± 0.0073 from *Planck*[11] (and 0.968 ± 0.012 from the 9-year *WMAP* results[12]; this is not one of the parameters on which *Planck* and *WMAP* disagree).

Further support for inflation has recently been provided by the *BICEP2* experiment[13], which reported the detection of B-mode polarisation in the cosmic microwave background. In the early universe, B-mode polarisation must be generated by gravitational waves: density fluctuations can only produce so-called E-mode polarisation (patterns with even parity, i.e. symmetric under reflection, unlike the odd-parity patterns of B-mode). E-mode polarisation can be converted to B-mode later in the history of the universe, by the distortions introduced by gravitational lensing, but this occurs at much smaller angular scales than the primordial B modes. The existence of such *primordial gravitational waves* is a solid prediction of inflation—they arise because of fluctuations in the gravitational field being "blown up" to macroscopic scale, in much the same way as the density fluctuations—and is not expected in some competing models such as those based on extra dimensions, so the *BICEP2* results, if confirmed, will provide strong evidence for the reality of inflation. However, the level of polarisation observed by *BICEP2* is surprisingly high: if this is not an accident of statistics (the statistical error of the result is large), it will put severe constraints on many theoretical models of inflation.

The key theoretical ingredient of inflation is the existence of a scalar field, the *inflaton* ϕ, which has a non-zero potential energy $V(\phi)$ when $\phi = 0$, and a minimum (zero?) value at some non-zero value of ϕ. At high energies, the energy density of the universe is dominated by $V(\phi)$, but as the universe cools ϕ must eventually settle down to its minimum. For inflation to work, the high-V region near $\phi = 0$ must take the form of a nearly flat plateau, terminated by a sharp drop-off to the minimum: the inflationary period occurs while ϕ slowly rolls off its plateau, and ends at the sharp drop. The energy released at the drop *reheats* the universe, producing large numbers of particle-antiparticle pairs: this is essential, because the pre-inflation number density of particles has been diluted to essentially zero by the inflation (a visible universe containing only one particle is not consistent with observations!).

The equation of state of a scalar field is given by

$$\begin{aligned}\mathcal{E}_\phi &= \frac{1}{2\hbar c^3}\dot{\phi}^2 + V(\phi),\\ P_\phi &= \frac{1}{2\hbar c^3}\dot{\phi}^2 - V(\phi),\end{aligned} \tag{1.1}$$

where $\mathcal{E}_\phi$ is the energy density, P_ϕ is the pressure, and $\dot{\phi} = \mathrm{d}\phi/\mathrm{d}t$. Exponential inflation corresponds to the case where $V\phi \gg \dot{\phi}^2/(2\hbar c^3)$, in which case the equation of state is approximately that of a cosmological constant, $P = -\mathcal{E}$. As shown in PHY306/406[7], a universe dominated by a cosmological constant expands exponentially, $a(t) \propto \exp(Ht)$ where $a(t)$ is the scale factor and $H = \dot{a}/a$ is the rate of expansion.

The properties of the inflaton field are reminiscent of those of the Higgs field [14], which is also a scalar field permeating all of space, and also has its minimum value at a non-zero value of the field. It is tempting to suggest that the inflaton field might actually *be* the Higgs field, which would be an elegant solution to the problem. Unfortunately, the constraints on the inflaton potential required for inflation to work lead to a naïve prediction that the mass of the inflaton should be around 10^{13} GeV, which is certainly not consistent with the Higgs. It *is* possible to persuade the Higgs field to drive inflation (see, for example, [15]) by giving it non-standard couplings, but the resulting model predicts a very low level of primordial gravitational waves, in contrast to the rather high level observed by *BICEP2*. However, extensions to the Standard Model generally require extensions to the Higgs sector—for example, supersymmetry has two Higgs doublets and five physical Higgs bosons, as opposed to one doublet and one boson in the Standard Model—so the lack of a fit with our one known Higgs boson is not a disaster.

Although inflation provides a conceptually elegant solution to the horizon and flatness problems of the classical Big Bang, and makes predictions (the geometry of the universe should be extremely close to flat; the spectral index of the anisotropies should be ∼0.95; there should be primordial gravitational waves) that are borne out by observation, the detailed particle physics underlying the idea appears problematic. The particular form of the inflaton potential necessary to make inflation work does not emerge naturally from the theory, but is put in "by hand," and the small coupling of the inflaton field makes it difficult to achieve thermal equilibrium. As the original motivation for introducing inflation was to avoid the need to fine-tune initial conditions, it is not satisfactory to find that one then has to fine-tune the properties of the inflaton field!

These problems are addressed by the *chaotic inflation* model (see, e.g., [16]), which works for a much wider range of potentials—the potential just has to be sufficiently flat—and initial conditions. The basic idea of chaotic inflation is that if the inital value of the scalar field ϕ is large, so that it dominates the energy density of the universe, the natural evolution of the Friedman equation will automatically lead to a quasi-exponential inflation (see [16], pp 6–7).

Unlike the original inflation models, chaotic inflation is not intimately linked to GUT phase transitions and does not require fine-tuning of the properties of the inflaton field; from the argument in the previous paragraph, nor should it require fine-tuning of the initial conditions (this point is highly debated, but Linde[16] claims that the criticisms are based on invalid assumptions).

As discussed by Guth in his original paper[9], the minimum amount of inflation needed to solve the horizon and flatness problems is about 60 e-foldings (i.e. an expansion factor of e^{60}, or 10^{26}). Chaotic inflation typically leads to *much* larger factors—Linde[16] quotes factors of order $10^{10^{10}}$! This implies that our visible universe is a *very* tiny part of a much larger cosmos. In addition, many inflation models lead to the scenario of *eternal inflation*, in which large quantum fluctuations during the inflation phase spawn separate "mini-universes," possibly with different low-energy physics, e.g. as a result of different compactification of the extra dimensions in string theories. This aspect of inflation provides an "escape" from fine-tuning problems such as the size of the cosmological constant: as a cosmological constant of "natural" size ($\sim 10^{120}$ times larger than what we observe) would make life impossible, we must necessarily live in a mini-universe with an unusually small value. (This is an example of the *Weak Anthropic Principle* [17]; such arguments are generally disliked by scientists because they are not very fruitful from a scientific perspective, but the basic logic of the argument—"we exist, therefore the laws of physics must be such as to permit us to exist"—is hard to fault.)

Inflation is certainly particle cosmology: scalar fields and quantum fluctuations belong to theoretical particle physics rather than classical cosmology. However, it is somewhat detached from the rest of theoretical particle physics: the inflaton field is introduced *ad hoc* rather than being deduced from the wider context of particle physics (except in so far as extensions to the Standard Model of particle physics do tend to predict additional scalar fields). This contrasts with other applications of theoretical particle physics to astrophysics and cosmology, such as baryogenesis (see next section) and non-baryonic dark matter (see below and PHY326/426), where the relevant theories also have implications for "traditional" experimental particle physics. Partly for this reason, and partly because the theory of inflation rapidly becomes very technical, we shall not cover inflation any further in this course. Interested students should refer to Linde's lecture notes[16], bearing in mind that Linde, as a co-inventor and vocal proponent of the theory, may not be exactly unbiased in his assessment of the arguments!

1.2.2 Baryogenesis

The other aspect of early-universe cosmology with clear links to theoretical particle physics is the problem of *baryogenesis*—why does the universe contain matter, but not antimatter?

When we create particles in terrestrial accelerators, we always create particle-antiparticle pairs (e.g. $e^+e^- \to q\overline{q}$), in accordance with the empirical conservation laws for baryon number B and lepton number L. However, the universe appears to contain baryons but no antibaryons[1], since (1) we do not observe any significant amount of antimatter locally—only a very small proportion of the cosmic-ray flux is antiparticles, consistent with recent production by high-energy particle collisions—and (2) nor do we observe the γ-ray flux from intergalactic space that would be expected if some galaxies were entirely matter while others were entirely antimatter.

In terms of number densities, though not of energy densities, the universe today is entirely dominated by the photons of the cosmic microwave background:

[1] Note that cosmologists tend to regard all Standard Model particles, with the possible exception of neutrinos, as "baryons". The reason for this is that the baryons completely dominate the mass: as the universe is electrically neutral on large scales, there's an electron for every proton, but the electron mass is only about 1/1800 of the proton mass.

there are about 1.6 billion photons for every proton. At temperatures where pair production ($\gamma\gamma \to f\bar{f}$) and annihilation ($f\bar{f} \to \gamma\gamma$) are in equilibrium, we would expect the numbers of photons and fermions to be approximately equal, so this huge disparity strongly suggests that *most* of the particles and antiparticles did indeed annihilate in the early universe, but some asymmetry in this process led to a remnant population of baryons and leptons which we now see (and of which we are made, so this one-in-a-billion imbalance is rather important to us!). The production of this remnant population is known as *baryogenesis*, and is one of the great unsolved problems of early-universe cosmology.

The conditions necessary for baryogenesis were laid out by Andrei Sakharov [18] in 1967: they are

1. interactions that violate baryon number conservation must exist;
2. C (charge conjugation) and CP (charge conjugation and parity) symmetries must both be violated;
3. the reactions must take place out of thermal equilibrium.

The first condition is obvious: if the universe starts from a matter-antimatter symmetric state in which $B = 0$, it cannot reach a state in which $B > 0$ without violating baryon-number conservation! The third is also obvious: in thermal equilibrium, forward and reverse reactions proceed at equal rates, so our hypothetical B-violating reaction would go equally in both directions, with no net gain. The argument for the second is similar to this: if C is conserved, reactions which increase B will be balanced by antireactions that decrease B, and if CP is conserved (even if C is violated), B-increasing reactions will be balanced by mirror-image B-decreasing antireactions.

As the early universe is expanding and cooling at a very rapid rate, the third condition is easily satisfied, as was first pointed out explicitly by Gamow [19] in 1946 (in the context of nucleosynthesis). Surprisingly, the first condition is also satisfied in the Standard Model: conservation of B and L is an "accidental" property of SM interactions, not a consequence of a fundamental symmetry of the Lagrangian. In 1976, Gerard 't Hooft [20] pointed out that a certain class of non-perturbative transitions violate B (though they conserve $B - L$). These non-perturbative processes, known as *sphalerons*, can convert three baryons into three antileptons or vice versa (the number has to be an integral multiple of the number of families, so the smallest possibility is 3). Sphalerons are a quantum tunnelling phenomenon: at today's low energies, they are suppressed to unobservably tiny levels, but they would have occurred readily at the very high energies of the early universe. This means that lepton-number-violating interactions can be bootstrapped into baryon production through such processes, a concept known as *leptogenesis*.

It is very possible that lepton number violation occurs and can be observed today, albeit at low levels. The key to this possibility is the existence of neutrinos—*electrically neutral* fundamental fermions. For charged particles, the particle and the antiparticle are observationally distinguished by their opposite charges, e.g. the electron and the positron. Even neutral baryons like the neutron are composed of charged constituents and are therefore distinguishable: electron scattering experiments would see a difference between the neutron (with one charge $+\frac{2}{3}$ up quark and two charge $-\frac{1}{3}$ down quarks) and the antineutron (one $-\frac{2}{3}$ anti-up and two $+\frac{1}{3}$ anti-downs). For the neutrino, on the other hand, there is no *obvious* difference between the particle and the antiparticle, except that one produces the charged lepton when it interacts, e.g. $\nu_\mu + n \to \mu^- + p$, and the other the charged antilepton, $\overline{\nu}_\mu + p \to \mu^+ + n$. This might seem like a perfectly adequate distinction, but the weak interaction has the interesting property of being *left-handed*: only particles with left-handed chirality, and antiparticles with right-handed chirality, can interact weakly. Therefore the apparent distinction between neutrino and antineutrino might really be a distinction between the two chiral states of the same particle, and thus the neutrino and the antineutrino would be different states of the same particle. Fermions with this property are called *Majorana particles*, after the Italian theoretical physicist Ettore Majorana[2].

This would be a purely academic distinction if the neutrino were massless, because a massless neutrino has a well-defined handedness. However, the neutrino is not massless, and therefore a neutrino which is produced as left-handed may have a very small probability of subsequently interacting as a right-handed object, i.e. an antineutrino. This could be observed through the rare process of *double beta decay*.

In nuclear physics, we find that even-A nuclei are more tightly bound, i.e. have lower masses, if they have even Z than they are if they have odd Z. This is a result of the pairing up of nucleons: odd-odd nuclei have two unpaired nucleons (one proton and one neutron), and thus a lower binding energy than even-even nuclei. As a consequence, it is possible for an even-even nucleus (A, Z) to have a lower mass than either of its immediate neighbours $(A, Z \pm 1)$, but a higher mass than a next-to-nearest neighbour $(A, Z \pm 2)$. An example of this is $^{76}_{32}$Ge (atomic mass 75.921402 u), which is lighter than $^{76}_{33}$As (75.922393 u) but heavier than $^{76}_{34}$Se (75.919212 u).

Isotopes like $^{76}_{32}$Ge are stable to single beta decay, but in principle unstable to double beta decay,

$$^{76}_{32}\mathrm{Ge} \to {}^{76}_{34}\mathrm{Se} + 2e^- + 2\,\overline{\nu}_e. \tag{1.2}$$

This is a perfectly legitimate decay mode, obeying all the rules of nuclear and particle physics, but the probability of two simultaneous weak decays is so small that $^{76}_{32}$Ge is to all intents and purposes an entirely stable isotope (it makes up 7.8% of natural germanium). The *two-neutrino double beta decay* ($2\nu\beta\beta$) described by equation 1.2 has in fact been observed for this isotope: the measured half-life is $(1.74 \pm 0.01^{+0.18}_{-0.16}) \times 10^{21}$ years[21].

From the point of view of baryogenesis, the interesting process is not $2\nu\beta\beta$ but *neutrinoless double beta decay* ($0\nu\beta\beta$), a variant which can occur *only* if neutrinos are Majorana particles. In this case, the neutrino is an *internal* line in the Feynman diagram, being produced at one vertex as a neutrino and absorbed at the other as an antineutrino. The result is

$$^{76}_{32}\mathrm{Ge} \to {}^{76}_{34}\mathrm{Se} + 2e^-, \tag{1.3}$$

which violates lepton number by 2. The signature of $0\nu\beta\beta$ is that the two electrons come out back to back, each with energy equal to half the Q-value of the decay, since there are no neutrinos to carry off energy and momentum. Unfortunately, since neutrinos are very nearly purely left-handed particles, the probability of the right-handed (antineutrino) interaction is extremely small, so the half-life of the $0\nu\beta\beta$ decay mode is expected to be long even compared to the $2\nu\beta\beta$ mode. No confirmed positive results have yet been reported: the most recent limit for $^{76}_{32}$Ge, from the GERDA experiment[22] is $t_{1/2} > 2.1 \times 10^{25}$

[2] Another example of a—hypothetical—Majorana particle is the neutralino, a leading dark matter candidate

years (at 90% confidence level). Calculated rates depend on the nuclear matrix element for the decay (which is a challenging theoretical calculation and subject to large errors) and the effective neutrino mass; if neutrino masses follow the inverted hierarchy, where at least two neutrino mass eigenstates must have masses of order 0.05 eV/c^2, the expected rates should be observable with the next generation of detectors. Such an observation would both disprove lepton number conservation and establish that the neutrino is indeed a Majorana particle, as well as providing the first ever measurement of an absolute neutrino mass (as opposed to the squared mass differences measured in neutrino oscillation experiments).

Although $0\nu\beta\beta$ violates lepton number, this process itself cannot be responsible for baryogenesis: it is much too slow. We want a process that will generate baryon number efficiently in the early universe, and then shut down (since we do not currently observe large-scale violation of B or L). It turns out that the concept of neutrinos as Majorana particles not only predicts $0\nu\beta\beta$, but also leads to such a mechanism.

One of the most appealing aspects of the Majorana picture of neutrinos is that it provides a natural explanation for the fact that their masses, while non-zero, are many orders of magnitude less than the masses of the other fundamental fermions (tritium beta-decay currently limits the effective mass of the electron neutrino to < 2.2 eV/c^2, and *Planck* finds that the sum of all 3 neutrino masses must be < 0.23 eV/c^2, though the latter limit has some model dependence). The trick relies on the existence of *right-handed* (and therefore non-interacting) neutrinos, which decouple from the rest of the fundamental particles when the grand unified theory breaks down to the Standard Model. They would therefore naturally have a mass M corresponding to the GUT scale of 10^{15} GeV or so. If we assume that the mass term for neutrinos in the Standard Model Lagrangian contains both a Dirac term like those for the charged fermions *and* a Majorana term, we wind up with a combined mass term[23]

$$\begin{pmatrix} \overline{\nu_L} & \overline{\nu_R^C} \end{pmatrix} \begin{pmatrix} 0 & m \\ m & M \end{pmatrix} \begin{pmatrix} \nu_L^C \\ \nu_R \end{pmatrix}, \tag{1.4}$$

where m is the Dirac mass, which we assume is similar to those of the other fermions, say 1–100 GeV/c^2, M is the Majorana mass and ν and ν^C are the neutrino and antineutrino wavefunctions respectively (the C superscript stands for charge conjugation). The off-diagonal Dirac terms couple left- and right-handed states, while the on-diagonal Majorana terms couple particle and antiparticle. Therefore, for a purely Dirac particle like the electron, the left- and right-handed states must have equal mass, whereas for a Majorana particle their masses can be quite different—in this case, 0 and M.

To get from (1.4) to the physical neutrino mass eigenstates, we need to diagonalise the matrix, which gives us one predominantly right-handed state with mass M and one predominantly left-handed state with mass m^2/M. If M is large, the mass of this second state (which is the one that couples to the weak interaction) is therefore automatically very small. This is the *seesaw mechanism* (so called because the higher the right-handed mass M goes, the lower the left-handed mass m^2/M becomes).

If the seesaw mechanism is the correct explanation of the light neutrino masses, it necessarily implies the existence of at least two massive, predominantly right-handed "neutrino" states N (because neutrino oscillation experiments guarantee that at least two of the three light, predominantly left-handed neutrinos have non-zero masses). Being very massive, these states couple strongly to the Higgs field and will decay by $N \rightarrow \ell(\bar{\ell}) + h$, where ℓ is a lepton and h is a Higgs boson. These are lepton-number-violating decays (the N, being a Majorana particle, does not have a well-defined lepton number), and leptogenesis occurs if the decay rates to ℓ and $\bar{\ell}$ are different. (In general, these decays also violate CP symmetry, so all three Sakharov conditions are satisfied.) Because the masses of the Ns are large, these decays occur in the very early universe when sphaleron transitions are common, so the lepton number asymmetry is partially transformed into a baryon number asymmetry.

Leptogenesis is an attractive way to generate the baryon asymmetry because of its close link to the seesaw mechanism, and because the proposition that lepton number conservation can be violated in the neutrino sector is experimentally testable, as is the existence of CP violation in the neutrino sector (although, in general, the CP-violating phase in neutrino oscillations, which is measurable, does not provide any useful constraints on the CP-violating phases in the heavy sector that are relevant to leptogenesis). There is also a link between leptogenesis and axions[24], relating the mass of the lightest right-handed neutrino to the axion symmetry-breaking scale. Axions are a possible candidate for cold dark matter (see below), so this might provide a link between two of the major unsolved problems of cosmology—the origin of the baryon asymmetry and the nature of dark matter.

Despite these advantages, leptogenesis is not without problems—for example, if the dark matter is supersymmetric particles rather than axions, leptogenesis tends to be associated with overproduction of gravitinos—and is far from the only available model of baryogenesis. CP-violating processes are known to occur in the quark sector, and sphaleron transitions in the early universe can generate non-zero B and L (while conserving $B - L$) as discussed above. Therefore it is possible to envision processes which generate non-zero B directly, in hadronic interactions, rather than indirectly through the neutrino sector, and indeed all the earliest models of baryogenesis were of this type. (Leptogenesis is only viable if neutrinos have mass, and therefore was not seriously considered until neutrino oscillations were established in the late 1990s.) There are three main mechanisms for hadronic baryogenesis: *GUT baryogenesis*, *Standard Model baryogenesis* and *electroweak baryogenesis*.

In GUT baryogenesis, the baryon number violation occurs through the GUT interactions rather than via sphalerons. Because grand unified theories explicitly unite the quark and lepton sectors of the Standard Model, there are heavy gauge bosons X (with spin 1) and Higgs bosons Y (with spin 0) which directly couple quarks to leptons, producing B and L violating interactions such as $X \rightarrow q_L e_R$. The Y decays in particular can occur at temperatures low compared to the mass of the Y, and therefore out of thermal equilibrium as required by the Sakharov conditions.

The trouble with GUT baryogenesis is that it obviously takes place at GUT-scale energies. This requires that the reheating phase after inflation reach high enough energies to produce X and Y bosons (since any such production *before* inflation gets diluted to nothing by the inflationary expansion). But one of the original motivations for inflation was to dilute away undesirable relics of the GUT scale such as magnetic monopoles, which are massive enough to overclose the universe and result in a rapid Big Crunch (clearly contrary to observation). Therefore, we would rather have baryogenesis taking place at a lower energy scale, too low to risk producing large numbers of such GUT relics.

At the other extreme, the observed hadron-sector CP violation in the Standard Model, together with B violation via sphaleron processes, should produce some level of baryon asymmetry in the early universe. This obviously has the advantage of requiring no new physics whatsoever, and therefore being in principle testable at existing energies. Unfortunately, the level of CP violation in the Standard Model appears to be too low to produce the observed baryon asymmetry[25], so this minimalist approach does not work: new physics is required to introduce new sources of CP violation. At least there is no requirement that the new physics must live at GUT energy scales, so this could still be subject to experimental verification.

The other problem with Standard Model baryogenesis is the out-of-equilibrium requirement imposed by the Sakharov conditions. This usually means that the strength of the interactions has to be $\ll m/M_{\mathrm{Pl}}$, where M_{Pl} is the Planck mass, $\sim 10^{19}$ GeV/c^2[25] (this arises because the Hubble parameter $H \propto 1/M_{\mathrm{Pl}}$). Given that the Standard Model mass scale is of order 100 GeV/c^2, corresponding to the masses of the W, Z and Higgs, this implies an unreasonably weak interaction (recall that the electromagnetic coupling constant $\alpha = 1/137$). A possible way round this is provided by the *electroweak phase transition*, i.e. the point at which, as the universe cools, the combined electroweak interaction breaks down into separate weak and electromagnetic components. If this occurs sufficiently abruptly—that is, if it's a first order phase transition—it can provide the necessary departure from equilibrium; this is electroweak baryogenesis.

A first order phase transition tends to proceed by forming bubbles of the new phase—e.g., boiling water. The bubble walls can provide sites for out-of-equilibrium reactions. In contrast, second-order phase transitions are typically smooth and continuous, and are much less likely to induce out-of-equilibrium conditions.

As the electroweak phase transition is closely related to the behaviour of the Higgs field[25] (it is, after all, the Higgs field that generates the masses of the W and Z), the critical issue in determining the order of the transition is the shape of the Higgs potential. Unfortunately it appears that a first-order transition requires a Higgs mass of < 75 GeV/c^2 or so[25], so a Higgs mass of 125 GeV/c^2 suggests a smooth transition and no baryogenesis. Rescuing the situation requires new physics, such as an additional Higgs doublet.

Adding supersymmetry to the Standard Model changes the picture, because we now have to consider not only an additional Higgs doublet, but also the effects of sparticle interactions. There are also many different versions of supersymmetry, some much more strongly constrained than others. The most studied version is the *Minimal Supersymmetric Standard Model* (MSSM). As its name suggests, this model contains only the minimum number of additional particles needed to provide supersymmetric partners (one per Standard Model particle, and one additional Higgs doublet, which produces four additional Higgs bosons and their SUSY partners).

Electroweak baryogenesis in the MSSM turns out to be difficult to realise[25]: generally the SUSY particle masses need to be rather high (several TeV/c^2), which is not "natural" since the motivation for SUSY—keeping the Higgs mass light by cancellation of correction terms—suggests that SUSY masses should be closer to the electroweak scale. On the other hand, the failure to find SUSY at LHC argues for higher masses, so perhaps this is less of a problem than it was perceived to be in 2006.

"Next-to-minimal supersymmetry" (NMSSM), which adds an extra scalar field to the MSSM particle content, has considerably more freedom of manoeuvre than minimal SUSY and can surely accommodate baryogenesis, but it is not obvious that this theory is well motivated. All of these arguments would become much more concrete if SUSY particles were actually discovered, either at the LHC or by direct dark matter searches.

Considered as particle astrophysics, baryogenesis has much clearer links to the rest of theoretical particle physics than inflation, and tests of baryogenesis models often involve "conventional" particle physics such as LHC experiments. Here we have focused on leptogenesis, because the connection with the neutrino sector links it more closely to the rest of particle astrophysics, but the various models of hadronic baryogenesis sketched above are by no means ruled out (see [25] for more information), and in many cases offer testable predictions for LHC physics or dark matter searches. A title search for "baryogenesis" in the arXiv preprint server[26] will demonstrate the wide range of baryogenesis models still under active consideration. Unfortunately, doing justice to baryogenesis requires at least graduate-level understanding of theoretical particle physics, and for this reason we will not be covering it in more depth in this course.

1.3 The physics of dark energy

In general relativity, the expansion of the universe is described by the Friedman equation,

$$H^2 = \frac{8\pi G}{3c^2}\mathcal{E} - \frac{kc^2}{R_0^2 a^2} + \frac{\Lambda}{3}, \tag{1.5}$$

where H is the Hubble parameter, $\mathcal{E}$ is the energy density in matter and radiation (the latter is negligible at the present time, but dominates in the early universe), k is the curvature (+1, 0 or −1), R_0 is the radius of curvature, and a is the scale factor, defined to be equal to 1 at the present time. The energy density is often expressed in terms of the *density parameter* $\Omega = \mathcal{E}/\mathcal{E}_{\text{crit}}$, where the *critical density* $\mathcal{E}_{\text{crit}}$ is given by

$$\mathcal{E}_{\text{crit}} = \frac{3c^2H^2}{8\pi G}. \tag{1.6}$$

The subscripts r and m distinguish the contributions to the density of radiation and (non-relativistic) matter; the subscript 0 indicates the value of the quantity at the present time.

Dark energy, in the form of the cosmological constant Λ, was first introduced into cosmology by Einstein himself, in 1917. Einstein's intent was to modify the equations of general relativity so as to permit them to describe a static universe, in agreement with the observational data of the time. With Hubble's establishment of the expansion of the universe in 1929–31, this motivation for the introduction of Λ disappeared, and for most of the following 60 years or so it was generally assumed to be zero, despite the lack of any theoretical justification for this. Ironically, given that it was introduced to make the universe static, it was the discovery that the expansion of the universe is actually *accelerating*, rather than slowing down as all $\Lambda = 0$ models predict, that returned the cosmological constant to favour in the late 1990s[27].

The observational evidence for $\Lambda > 0$ comes from astrophysics and cosmology rather than particle astrophysics, and is discussed in PHY306/406. Briefly, the principal points include:

- the Hubble diagram for Type Ia supernovae, showing accelerating expansion in recent times ($z < 0.6$ or so);

- analysis of the anisotropies of the cosmic microwave background, which indicates that the universe is geometrically flat ($\Omega_{\rm tot} = 0$), but that the matter density $\Omega_{\rm m0}$ is only $\sim$0.3;
- simulations of large scale structure, which show good agreement with observations only if $\Lambda > 0$;
- analysis of the X-ray emission from rich clusters of galaxies, which shows a consistent ratio of gas mass to total mass only if $\Omega_\Lambda \sim 0.65$;
- comparison of the age of the universe derived from H_0 with the ages of old objects, e.g. globular clusters (ages derived from stellar evolution fits to the Hertzsprung-Russell diagram) and individual metal-poor stars (radiometrically dated using uranium and thorium).

These independent lines of evidence are all consistent with a cosmological model in which $\Omega_{\rm m0} \simeq 0.3$ and $\Omega_{\Lambda 0} \simeq 0.7$. It is fair to say that the existence of a Λ-like component dominating the present energy density of the universe is well established. What that component actually is, however, is very far from well established—and this topic certainly does fall within the remit of particle astrophysics.

Physically, the standard cosmological constant, with equation of state $P_\Lambda = -\mathcal{E}_\Lambda$, represents the energy density of the vacuum. The idea that this should be non-zero is entirely reasonable in the context of quantum mechanics: according to the Uncertainty Principle, empty space should be full of virtual particle-antiparticle pairs that spontaneously appear and then re-annihilate (after a time short enough that $\Delta E \Delta t < \hbar$), so on average its energy should not be zero. The problem is actually the opposite: calculations of the expected vacuum energy give values that are too large by *many* orders of magnitude (a factor of 10^{120} in the Standard Model, reduced to "only" 10^{60} in supersymmetric models). This is because the momenta of the virtual particles are unknown (since they re-annihilate without being observed), so one has to integrate over all possible values of the momentum and all possible types of particle, giving[27]

$$\mathcal{E}_\Lambda = \frac{1}{2} \sum_{\rm fields} g_i \int\limits_0^\infty \sqrt{k^2 + m^2} \, \frac{{\rm d}^3 k}{(2\pi)^3}, \tag{1.7}$$

where k and m are the momentum and mass of the particle being created and g_i is the number of degrees of freedom of the field (e.g. 2 for a photon, which has two possible polarisation states). As it stands, this integral is infinite: it diverges at the upper limit. We can make it finite by only integrating up to some cut-off factor $k_{\rm max}$: it is then $\propto k_{\rm max}^4$. The justification for such an apparently arbitrary cut-off is normally the appearance of new physics; unfortunately, the cut-off value one would need to impose to get close to the observed value of Λ is about 0.01 eV, whereas the natural cut-off scales for new physics are the Planck mass ($\sim 10^{19}$ GeV) for the Standard Model and around 1 TeV for supersymmetric models.

There is no very obvious escape from this problem. Supersymmetry helps because the g_i factor has a negative sign for bosons and a positive one for fermions, so if it were an exact symmetry the contributions from particle and sparticle would cancel each other out. Unfortunately supersymmetry is clearly *not* an exact symmetry (the masses of SUSY particles are much greater than those of their Standard Model partners, with the possible exception of the stop squark), so the net contribution is $\propto M^4$ where M is the mass scale at which the symmetry is broken, assumed to be of order 1 TeV as stated above (perhaps a bit higher, given that the LHC has so far failed to find SUSY). This has motivated theorists to seek alternative models for this component of the universe—hence the introduction of the less-specific term *dark energy* in place of "cosmological constant". Another possibility is suggested by the anthropic principle: a universe with the natural value of Λ would expand too rapidly for galaxies to form, and so we *must* live in a universe with an anomalously low value of Λ. This argument makes most sense in a "multiverse" model such as that produced by eternal inflation (see page 11): if it is assumed that the value of Λ, while constant for any given mini-universe, varies randomly from one mini-universe to the next, it could be that the mini-universe in which we live has a quite exceptionally small value of Λ (whereas the overwhelming majority of mini-universes have "natural" values of Λ and are uninhabitable). However, as noted earlier, most theorists have an aversion to such anthropic-principle arguments because they are scientifically unproductive; in addition, it is not at all obvious from the argument presented above that Λ *should* behave like a random variable.

If we do not rely on the anthropic principle and instead seek alternative models, the obvious approach, as adopted in inflation (see section 1.2.1) is to postulate a scalar field[27]. From equation (1.1), the equation of state of a scalar field is $P_S = w\mathcal{E}_S$ where

$$w = \frac{-1 + \dot{\phi}^2/2\hbar c^3 V}{1 + \dot{\phi}^2/2\hbar c^3 V}. \tag{1.8}$$

As noted in section 1.2.1, in the case where $\dot{\phi}^2 \ll 2\hbar c^3 V$, $w \simeq -1$: a slowly varying scalar field will look very like a cosmological constant. In general, the value of w will change with time as the field evolves: depending on the shape of $V(\phi)$, models can "freeze" (w evolves towards -1) or "thaw" (w is initially ~ -1 but evolves away from -1) [27]. Such a time-varying w is usually parameterised as $w(a) = w_0 + (1 - a)w_a$, where w_0 and w_a are constants and $a(t)$ is the scale factor (normalised to 1 at the present time). Observational data are beginning to constrain the values of w_0 and w_a (see, e.g., figures 35 and 36 of [11]), but so far the constraints are not very strong.

An attractive feature of some freezing models—so-called "tracker models"—is that the energy density of the scalar field tracks the dominant conventional energy density (radiation or matter) at early times before starting to diverge: this makes the coincidence that we happen to live in the epoch at which both Ω_m and Ω_Λ are comparable in size less improbable than it is in the straightforward vacuum energy model. Attributing dark energy to a scalar field also raises the possibility that the accelerated expansion of the present epoch could be related somehow to the accelerated expansion of inflation, since in this scenario both are driven by scalar fields, albeit with dramatically different energy scales.

The principal issue with scalar-field models of dark energy is, again, the extremely small value of $\mathcal{E}_\Lambda$ at the present time. This requires a very flat potential $V(\phi)$, a very small effective mass, and also an extremely weak coupling of the scalar field to other particles (to avoid introducing unobserved and thus unwanted long-range forces) [27]. It is difficult to incorporate this peculiar field into the wider context of theoretical particle physics. The only known particles that operate at a similar energy scale are neutrinos, and unsurprisingly some theorists have attempted to treat this as a meaningful relationship rather than a coincidence. In neutrino dark energy models (see, e.g., [28]), the scalar field couples to neutrinos, and its energy density is causally related to the neutrino mass (which in these models is generated dynamically and changes with time). The behaviour of neutrino dark energy has a tendency to become unstable when the neutrinos of the cosmic neutrino background become non-relativistic—certainly the case at the present time, for neutrino masses of order 0.05 eV/c^2—but models avoiding this problem can be constructed[28]. As the neutrino mass grows with time in these models, a possible experimental signature would be a conflict between a measured neutrino mass (e.g. a positive result from the KATRIN tritium beta decay experiment) and the upper limit on the sum of neutrino masses derived from the CMB.

In the context of Einstein's field equations, vacuum energy and scalar fields modify the stress-energy tensor, i.e. the matter side of the equation. An alternative approach is to attack the geometric side, i.e. to modify gravity. This approach can be motivated by extra-dimension models, since in many such models gravity (unlike the other forces) also propagates in the extra dimensions and hence is not perfectly described by general relativity. An example discussed in [27] modifies the Friedman equation to

$$H^2 = \frac{8\pi G\mathcal{E}}{3c^2} + \frac{H}{r_i}, \tag{1.9}$$

where r_i is a length scale. The extra term H/r_i causes acceleration at late times when the energy density is small. Unfortunately, such models tend to have unphysical features such as tachyons; Frieman, Turner and Huterer[27] conclude that "it is not clear that a self-consistent model with this dynamical behaviour exists."

In conclusion, the physics of dark energy certainly belongs in the field of particle astrophysics, but so far is proving a rather intractable problem. None of the possible approaches—vacuum energy, dynamically generated dark energy from scalar fields, modified gravity, or simply assuming that we live in a universe which is inhomogeneous on large scales[3]—has yet yielded a good explanation of the observations: so far, the data are all consistent with a simple cosmological constant ($w = -1$ at all times), but there is no theoretical motivation for its small value. Better observational data, more strongly constraining the dark energy equation of state and its possible time variation, should help to decide which avenues of theoretical speculation to pursue.

1.4 High-energy processes in astrophysics

Through most of its history, astronomy has been the study of starlight (reflected starlight, in the case of planets). Starlight is thermal (approximately blackbody) radiation with an effective temperature ranging from about 3000 to 30000 K, corresponding to energies of order 1 eV. The nuclear fusion processes that power starlight take place at temperatures of around 10^7 K for hydrogen burning, going up to 10^8 K for helium and a few times 10^9 K for the short period of heavy-element fusion prior to supernova explosion: these temperatures correspond to nuclear physics energies of 1–100 keV. The *iron peak* in nuclear abundances (see PHY320) is evidence that the elements around iron are

[3] The idea here is that we happen to live in a region which is underdense compared to the rest of the universe; the extra gravitational forces introduced by this can mimic the effect of a cosmological constant. In order for this to be consistent with the highly isotropic nature of the CMB, the underdense region must be very large and the Milky Way must be rather close to the centre of it. This looks decidedly contrived.

made in conditions of nuclear statistical equilibrium, indicating temperatures of order a few MeV (the binding energy of the most stable elements is about 9 MeV per nucleon), but this is still more the domain of nuclear than of particle physics. However, the advent of radio astronomy after the second world war, followed by the rest of the electromagnetic spectrum up to γ rays from the 1960s onwards, provided clear evidence that thermal emission is not the only source of radiation in the cosmos. Further evidence comes in the form of *cosmic rays*, energetic charged particles first unambiguously detected by Victor Hess in 1911. The cosmic ray energy spectrum goes up to extraordinarily high energies ($\sim 10^{20}$ eV or more—that's over a joule of kinetic energy in a single proton!), clearly demonstrating the existence of astrophysical particle accelerators. Unfortunately, as we shall discuss later, the Galactic magnetic field deflects even energetic charged particles to such an extent that the sources of these ultra-high-energy cosmic rays still remain unidentified.

1.4.1 The non-thermal universe

Thermal radiation is described, at least approximately, by the Planck function,

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\,\frac{1}{\exp\left(\frac{h\nu}{k_{\rm B}T}\right) - 1}, \tag{1.10}$$

where ν is frequency, h is Planck's constant and $k_{\rm B}$ is Boltzmann's constant. For low frequencies,

$$\exp\left(\frac{h\nu}{k_{\rm B}T}\right) - 1 \simeq \frac{h\nu}{k_{\rm B}T},$$

giving the *Rayleigh-Jeans approximation*

$$B_\nu(T) \simeq \frac{2\nu^2 k_{\rm B}T}{c^2}. \tag{1.11}$$

If astrophysical radio sources were thermal in nature, one would therefore expect them to have a spectral energy distribution with flux $\propto \nu^2$. In fact, the spectra of radio galaxies usually follow power laws with *negative* spectral indices, $S \propto \nu^\alpha$ where S is the flux and the spectral index $\alpha \sim -1$ (within about a factor of 2). Therefore the radio emission cannot be thermal. It is in fact *synchrotron radiation*, produced by a population of relativistic electrons gyrating in a magnetic field (the name comes from the observation of this radiation in terrestrial particle accelerators, i.e. synchrotrons). *Supernova remnants* such as the Crab Nebula also emit synchrotron radiation at radio frequencies, and therefore must also accelerate electrons to relativistic speeds. It is ironic that the lowest-energy radiation in the electromagnetic spectrum provided the first evidence of high-energy processes in astrophysical sources.

X-ray and γ-ray emission also provides evidence of high-energy processes at work in sources such as supernova remnants and active galaxies. Many sources have spectral energy distributions consistent with *inverse Compton scattering*, where photons gain energy by back-scattering off fast electrons (in contrast to "normal" Compton scattering where X-rays *lose* energy by scattering off stationary or slowly moving electrons). This again requires a population of relativistic electrons in the source. The same sources frequently emit in both the radio and the X-ray/γ-ray regime, as the same population of fast electrons can both emit (radio-frequency) synchrotron photons and back-scatter them to much higher frequencies: this is known as *synchrotron-self-Compton* (SSC) emission. The relative normalisation of the synchrotron radiation and the inverse Compton flux is set by the magnetic field strength (generally not measured independently, but fitted from the flux).

Synchrotron radiation and inverse Compton emission require populations of relativistic electrons (associated, in the first case, with magnetic fields), but do not require fast protons or ions. However, the observation of cosmic ray fluxes extending to extremely high energies unambiguously demonstrates that hadrons are also accelerated by some (unidentified) type(s) of astrophysical source. The presence of high-energy protons in γ-ray sources could be signalled by a different spectral shape: high-energy protons colliding with ambient gas or radiation would be expected to produce large numbers of pions, and the decay $\pi^0 \to \gamma\gamma$ would convert these into a γ-ray signal with a much flatter spectrum than that of inverse Compton scattering. In addition, charged pions would decay through $\pi^+ \to \mu^+\nu_\mu$ ($\pi^- \to \mu^-\overline{\nu}_\mu$): observations of high-energy neutrinos would unambiguously tag a source as accelerating hadrons. Unfortunately, although very-high-energy neutrinos have recently been observed by IceCube[29], no point sources have yet been identified.

In summary, observations outside the optical waveband reveal that several types of astrophysical object are in effect particle accelerators, capable of accelerating electrons, at least, up to very high energies. Cosmic-ray observations supplement this by demonstrating a need for hadron accelerators as well. These observations pose questions to particle astrophysicists: what is the acceleration mechanism (or mechanisms); where does the acceleration take place; what is the origin (or origins) of high-energy cosmic rays; and what do the answers to these questions tell us about the nature of the astrophysical sources?

1.4.2 Detection techniques

So far, we have covered topics relevant to *theoretical* particle astrophysics: inflation and baryogenesis, the physics of dark energy, and the nature and location of particle acceleration in astrophysical sources. However, high-energy particle astrophysics also includes experimental aspects: while radio astronomy and the lower-energy end of X-ray astronomy qualify as conventional astronomy with focusing paraboloid optics (albeit, in the case of X-ray telescopes, with unconventional geometry), very high energy photons and charged particles require technology more usually associated with particle physics experiments.

High-energy photons (γ-rays) do not reflect from materials, so conventional astronomical imaging optics are not possible. Instead, a variety of techniques are used, as listed below.

- *Coded mask telescopes* (see, e.g., [30]) work by placing a patterned mask in front of the instrument. The mask consists of a complex and carefully designed pattern of opaque and transparent sections, such that the shadow it casts on the instrument depends on the direction of the incoming flux. A deconvolution algorithm is used to construct an image of the field being viewed. Although the coded mask technique seems wasteful—you are deliberately blocking off quite a large fraction of your collecting area—it is useful in the hard X-ray/soft γ-ray regime ($\sim$3 keV–20 MeV), where reflecting optics do not work but the incident photon is too soft for a tracking calorimeter (see below). They also have the advantage of a large field of view, and hence make good survey or transient-finding instruments. Examples of coded mask telescopes include the IBIS telescope and SPI spectrometer on INTEGRAL, the Burst Alert Telescope (BAT) on *Swift* and the Wide Field Camera on BeppoSAX[31].

- *Compton imaging* uses Compton scattering to produce an image. If both the scattered particles—the electron and the photon—are detected, and their energies and positions measured, relativistic kinematics can be used to reconstruct the energy and direction of the incoming photon. This technique was used in the COMPTEL instrument[32] on the Compton Gamma Ray Observatory (CGRO) satellite, and is also used in medical imaging. COMPTEL covered an energy range of 0.8–30 MeV with an angular resolution of 1.7–4.4° for individual photons (the source itself could be located with a precision of 5–30 arcmin). The field of view was about one steradian.

- *Pair-conversion tracking calorimeters* are used for higher energy γ-rays, which will readily convert to e^+e^- pairs when passing through material. These are genuine particle physics experiments, much more comprehensible to an LHC physicist than to a conventional astronomer! The ingredients are (1) thin plates of absorber to encourage the γs to convert, interspersed with (2) tracking elements to detect and reconstruct the e^+e^- pair and followed by (3) calorimetry to measure the energy. The first such instrument was EGRET (the Energetic Gamma Ray Experiment Telescope) [33] aboard CGRO. The EGRET pair conversion spectrometer consisted of metal plates alternating with spark chambers, followed by thallium-doped sodium iodide (NaI(Tl)) scintillating crystals for energy measurement. The instrument was covered with a plastic scintillator dome in anticoincidence for background rejection (to veto incoming charged particles). EGRET was sensitive to γ-rays with energies between 20 MeV and 30 GeV.

  The successor to EGRET is the Large Area Telescope (LAT) on board the *Fermi* satellite[34]. The LAT has much more modern particle physics technology: the converter-tracker consists of tungsten absorber interleaved with silicon strip detectors for tracking, and the calorimeter section is thallium-doped caesium iodide scintillating crystals. The anticoincidence detector consists of plastic scintillator tiles. The LAT is sensitive to photons in the energy range 20 MeV–300 GeV, with an energy resolution of order 10% and a single-photon angular resolution ranging from 0.15° above 10 GeV to 3.5° at 100 MeV. The field of view is 2.4 steradians, and point sources can be located to better than $0.5'$.

For energies above 300 GeV, space-based experiments are not practical because the calorimeter needed to contain such high-energy showers would be too heavy, and because high-energy events are rare enough to require larger collecting areas (and thus even heavier calorimeters). Therefore, the very highest energy γs are detected using ground-based instruments.

- *Imaging air Cherenkov telescopes* (IACTs) detect the electromagnetic shower produced when a very-high-energy γ enters the atmosphere. The secondary $e^\pm$ produced in the shower have high enough energies that they are travelling at speeds greater than c/n, where n is the refractive index of air, and therefore generate *Cherenkov radiation* [35] in a narrow cone about the direction of the incoming photon. This light is collected by a parabolic mirror and focused on to a "camera" consisting of an array of small photomultiplier tubes. Examples of IACTs include H.E.S.S. in Namibia, MAGIC in the Canary Islands and VERITAS in the USA. The low energy threshold depends on the size of the telescope, but is typically 30–100 GeV; γs are detected up to energies of many TeV. The main problem is that the Cherenkov emission is very faint, so these telescopes have a relatively poor duty cycle: they can operate only on clear, dark nights.

The detection of cosmic rays presents similar challenges. Cosmic rays are generally protons or heavier nuclei, and therefore the *primary cosmic rays* themselves are not detected at ground level—just the *secondary* cosmic rays such as muons, which are the products of the interactions of primary cosmic rays with the atmosphere.

Early cosmic-ray experiments were flown on high-altitude balloons or rockets. As with γ-ray detection, modern experiments divide into relatively small space-based detectors concentrating on the lower-energy part of the spectrum, and much larger ground-based arrays to detect the rarer ultra-high-energy cosmics.

The orbiting cosmic-ray observatories PAMELA[36] (a satellite) and AMS-02[37] (on the International Space Station) are both magnetic spectrometers, similar to many accelerator-based particle physics experiments. Both instruments have similar aims, namely to study the antimatter component of cosmic rays (positrons, antiprotons and perhaps heavier antinuclei such as antideuterons and antihelium), to conduct indirect searches for dark matter (see below) and to provide precise measurements of the primary cosmic ray flux and spectrum, and its variation over time.

Ground-based cosmic ray experiments, like air Cherenkov telescopes, detect the *extensive air shower* (EAS) produced when a high-energy primary cosmic ray interacts with the Earth's atmosphere. There are two principal approaches.

- *Nitrogen fluorescence* is produced when the secondaries from the interaction (particularly $e^{\pm}$) excite nitrogen molecules in the atmosphere. The de-excitation produces line emission in the near UV (300–400 nm), which is detected using telescopes very similar to the Cherenkov telescopes described above. Unlike Cherenkov radiation, the fluorescence is emitted isotropically, so fluorescence telescopes generally see the shower "side-on" rather than "head-on"; like Cherenkov radiation, it is very faint and therefore detectable only on clear, dark nights.

- *Ground arrays* are, as the name suggests, arrays of small, semi-autonomous detectors designed to sample the fraction of the EAS secondaries that reach the ground. Each small detector triggers independently and sends its time-stamped data to a central facility which combines the data from all detectors to reconstruct the shower. The small detectors need to be simple, robust and cheap to construct (since you want to instrument as much area as possible): the preferred technologies are Cherenkov radiation (using small, self-contained water tanks) or scintillators. Some arrays have also included specialised muon detectors (underground, or underneath the main detectors, so that only muons reach them) to study the particle content of the shower.

  Ground arrays have the advantage of a 24-hour duty cycle, but the disadvantage that in order to cover a large area you must physically distribute detectors over a large area (in contrast to fluorescence telescopes which can detect fluorescence originating a long way from the actual telescope). The largest ground array, the Pierre Auger Observatory[38], is a hybrid instrument combining a very large ground array (1600 water Cherenkov tanks) with a set of fluorescence telescopes arranged to look out over the array, so that on suitable nights both fluorescence and ground sampling data will be available.

Both high-energy γ-rays and cosmic rays are classic particle astrophysics: particle physics technology harnessed to astrophysical applications. It is also worth noting that high-energy particle physics *began* as cosmic-ray physics: the early discoveries such as the positron, the muon, the pion and strange particles were all made in cosmic rays, before terrestrial particle accelerators were developed.

1.5 Neutrinos

Neutrinos are probably the second most abundant particle in the universe, after photons (and possibly axions, if dark matter consists of axions). In view of their weak interactions, nothing is "optically thick" to most neutrinos: for example, the solar neutrinos we detect on Earth have come directly from fusion reactions in the core of the Sun, whereas the photon diffusion time from the core to the photosphere is of the order of 200000 years. In principle, therefore, neutrinos can carry information about processes occurring deep inside astrophysical objects, which cannot possibly be directly observed using photons. Unfortunately, neutrinos are equally reluctant to interact with detectors, so only extremely intense neutrino fluxes provide useful numbers of events in terrestrial detectors. To date, only two astrophysical sources of neutrinos have been identified: the Sun, and Supernova 1987A in the Large Magellanic Cloud.

Astrophysical neutrinos are produced in many contexts, from the early universe to the interiors of main-sequence stars, and span a wide range of energies. Their properties are important in many branches of astrophysics and cosmology.

1.5.1 Neutrinos in cosmology

Like the other fundamental particles, neutrinos are produced in large numbers during the reheating period immediately after inflation. Because of their weak interactions, they decouple from the rest of the matter in the universe at a temperature ~1 MeV, a second or so after the Big Bang, and we expect that both neutrinos and antineutrinos will have survived to the present day.

The Cosmic Neutrino Background (CνB) is very similar to the cosmic microwave background (CMB), except that

- it has a Fermi-Dirac distribution rather than a blackbody distribution;

- it is at a slightly lower temperature (1.95 K rather than 2.725 K), because the photons gain extra energy when electrons and positrons annihilate in the early universe (at $T \sim 0.3$ MeV) whereas the neutrinos, which have already decoupled at that point, do not.

The number of relic neutrinos is predicted to be about 340 per cubic centimetre, split equally among six types (three neutrinos and three antineutrinos). This would be *extremely* difficult to verify experimentally, because the interaction cross-section of such a low energy neutrino is tiny even by weak interaction standards.

In the early universe, the CνB is a significant part of the total energy density of the universe. This has a number of consequences:

- as $H^2 \propto \mathcal{E}$, the neutrinos contribute to the early expansion of the universe, and therefore affect the outcome of big bang nucleosynthesis (faster expansion implies faster cooling, so the neutrons have less time to decay before nucleosynthesis begins, and hence more ^{4}He is made);

- since neutrinos are not massless, they act as *hot dark matter*, which affects the formation of large-scale structure and the anisotropies of the cosmic microwave background.

These effects can be used to place limits on the number of neutrino species and the total mass of all species of neutrinos, $\sum_i m_{\nu_i}$. *Planck*[11] quotes $N_{\text{eff}} = 3.30 \pm 0.27$ for the effective number of neutrino species and $\sum_i m_{\nu_i} < 0.23$ eV/c^2 for the total mass; the latter is a much stronger limit than any currently obtained by direct experiments, but there is some model dependence.

In addition, as discussed in section 1.3, attempts have been made to connect neutrinos with dark energy, on the grounds that they have a similar energy scale.

1.5.2 Solar neutrinos

Certainly the most intensively studied astrophysical neutrinos are those produced by solar fusion reactions. These have fairly low energies, ranging from 400 keV or so for the most numerous pp neutrinos (from $p + p \to d + e^+ + \nu_e$) up to around 15 MeV for the rare ^{8}B neutrinos (from $^8\text{B} \to {}^8\text{Be} + e^+ + \nu_e$). The reactions that produce solar neutrinos are discussed in more detail in PHY320.

Solar neutrinos have the potential to probe the fusion reactions in the Sun's interior—for example, it might be possible to make a direct measurement of the fraction of the Sun's luminosity produced by the CNO reaction cycle, which produces a different set of neutrinos with different energies. However, so far their principal application has been in understanding the physics of neutrinos.

All solar neutrinos are originally produced as ν_e: the Q-values of the reactions are not more than a few MeV, precluding the production of the heavier charged leptons (μ and τ), and therefore of their associated neutrinos. However, experiments which detect only ν_e consistently detect too few, by a factor of 2–3 depending on the energy range to which they are sensitive. This is the so-called *Solar Neutrino Problem*, which remained unresolved for many years. Its resolution in terms of *neutrino oscillations* was finally definitively demonstrated in 2002 by the SNO experiment[39], which used neutrino interactions on heavy water (D_2O) to prove that the *total* neutrino flux was as predicted by theorists, the deficit being due to transformation of ν_e into some other flavour.

1.5.3 Supernova neutrinos

The other confirmed source of astrophysical neutrinos is the Type II core-collapse supernova SN 1987A. A total of 24 neutrinos were observed by three experiments (Kamiokande-II, IMB, and Baksan) about three hours before the optical explosion was detected. This slight time difference is expected, because the first stages of the explosion are opaque to photons, though not (of course) to neutrinos[4]. The number of neutrinos observed, and their energies, were con-

[4] Given that the LMC is about 50 kpc away, this almost-negligible time difference was hard to reconcile with the September 2011 claim by the OPERA experiment that their neutrinos were travelling faster than light—neutrinos travelling at the speed implied by the OPERA results would have arrived four *years* early, not three hours! Much theoretical fudging went into attempting to reconcile these results, but the OPERA measurement was simply wrong.

sistent with the expectation that about 99% of the energy of a core-collapse supernova is emitted in the neutrino burst, with only about 1% going into the visible explosion.

Simulations of core-collapse supernovae suggest that the intense neutrino emission is essential to the physics of the supernova itself. The explosion is initiated when infalling material bounces off the surface of the newly-formed neutron star, creating a shock front: however, in early simulations the shock promptly stalled, causing the rest of the stellar material to fall back on to the neutron star. This produced a black hole and no visible explosion, in contradiction to observations (core-collapse supernovae definitely *do* explode!). Part of the problem was deficiencies in the simulations: supernova ignition seems to be quite asymmetric, so the early models which assumed spherical symmetry (to reduce a 3D problem to a 1D one, with enormous saving in computing power) were not reproducing the physics properly. However, this alone is not enough. It appears that the shock is revived by *neutrino heating*: the density of the material is so high, and the neutrino flux so great, that a significant amount of energy is dumped by the neutrinos into the stalled shock, reinvigorating the explosion.

The number of neutrinos detected from SN 1987A was not large enough to do more than order-of-magnitude calculations (not that this is reflected in the enormous number of theoretical papers on the subject...). However, should a supernova explode in our Galaxy, the number of neutrinos that would be observed by the current generation of detectors would be well into the thousands (a Galactic supernova would be a factor of 5 closer than SN 1987A, and Super-Kamiokande is about an order of magnitude larger than Kamiokande-II). Such a data sample would provide opportunities for both neutrino physics (the initial "neutronisation pulse" of ν_e, generated by the formation of the neutron star, is sharp enough that correlations between arrival time and neutrino energy could be used to set limits on, or perhaps even measure, the neutrino mass) and the astrophysics of supernova explosions (from the time and energy spectra of the subsequent "thermal" neutrinos produced in the early stages of the explosion). Of course, we do not know when the next such event will occur—arguably, given the observations of Tycho's supernova in 1572 and Kepler's in 1604, and the dating of the Cas A explosion to $\sim$1670, we have been unlucky to observe no Galactic supernovae at all in the last 300 years (admittedly, both Tycho's and Kepler's supernovae seem to have been of Type Ia, and would not have produced neutrino bursts). We can but hope.

Of course, neutrinos from *past* core-collapse supernovae still exist, and are still travelling outwards from the original explosion at approximately the speed of light. The flux from such *supernova relic neutrinos* is much lower than the burst from a Galactic supernova, but it is continuous and detectable at all times (if detectable at all). Calculations indicate[40] that there might be a "window" of observability between 20 and 30 MeV: below 20 MeV, the signal is drowned out by solar neutrinos, and above 30 MeV by atmospheric neutrinos from cosmic-ray interactions. Searches by Super-Kamiokande have so far been unsuccessful[41], but the next generation of still larger neutrino detectors might do better. The detection of supernova relic neutrinos could be used to constrain the history of the star formation rate, and would also provide information about neutrino properties (e.g. oscillations, and limits on neutrino lifetimes).

1.5.4 Atmospheric neutrinos

When primary cosmic rays interact in the atmosphere, they produce pions. Charged pions subsequently decay to muons and ν_μ, and the muons then decay by $\mu^- \rightarrow e^- \nu_\mu \overline{\nu}_e$ (or the equivalent for μ^+). Therefore, cosmic ray interactions produce a flux of *atmospheric neutrinos*. At low energies, essentially all of the muons decay, and the atmospheric neutrino flux should consist of ν_μ and ν_e in the ratio 2:1 (ignoring the distinction between neutrinos and antineutrinos); for higher energies, time dilation effects will allow some muons to reach the ground before decaying, and the ν_μ to ν_e ratio should be greater than 2.

In fact, we find that the $\nu_\mu : \nu_e$ ratio depends on the zenith angle of the neutrinos: it is as predicted for neutrinos coming straight down (and therefore travelling about 20 km), but decreases with increasing zenith angle, reaching a minimum for neutrinos coming straight up (and therefore travelling about 12800 km).[42] This is an effect of neutrino oscillations: the ν_μ are oscillating into ν_τ over the longer distances. Atmospheric neutrino measurements provided the first generally accepted evidence for neutrino oscillations[43] (in fact, the solar neutrino problem (see above) had been providing such evidence for two decades, but its reliance on calculations of the solar neutrino flux based on theoretical models made people reluctant to accept it as definitive).

Atmospheric neutrinos qualify as particle astrophysics, since they are secondary products of cosmic rays, but are not generally regarded as such, because the analysis of atmospheric neutrino data provides information about the properties of neutrinos, not about cosmic rays. Their principal significance for neutrino astronomy is as an irreducible background in searches for high-energy neutrinos from astrophysical sources.

1.5.5 High-energy neutrinos

Observations of cosmic rays (see above) provide conclusive proof that some astrophysical sources emit ultra-high-energy protons. Such protons will interact with ambient gas and/or photons in the source to produce pions, and the charged pions will decay into muons and neutrinos. (This is a well-established process, which is responsible for the atmospheric neutrino flux discussed in the preceding section, and also for the production of neutrino beams from terrestrial particle accelerators.) Therefore, all sources of high-energy cosmic rays should also be sources of high-energy neutrinos. As a consequence of the decay kinematics, the neutrino energies will typically be about 5–10% of the proton energies—still very high, given that the proton energies range up to $> 10^{20}$ eV. The great advantage of the neutrinos is that, being uncharged, they will not be deflected by the Galactic magnetic field and will therefore point back to their place of origin. The great disadvantage is that they are weakly interacting and will therefore be very difficult to detect in the first place. It is therefore not at all surprising that point sources of high-energy astrophysical neutrinos have not yet been identified, despite a couple of decades of searching.

Given the small interaction cross-section, the paramount design criterion for "neutrino telescopes" is that they must be as large as possible. As a result, the usual approach is not to build a structure, but instead to instrument a naturally-occurring target medium. So far, the technique of choice is Cherenkov radiation (from the charged lepton produced when a neutrino interacts by W exchange) in natural bodies of water, either liquid (Lake Baikal or the Mediterranean) or solid (the Antarctic icecap). Strings of "optical modules" (consisting of a large photomultiplier tube and its associated electronics, housed in a pressure-resistant glass sphere) are lowered into the water or ice, and the charge and timing information used to reconstruct the Cherenkov cone. A number of neutrino telescopes are currently in operation: the most successful, simply because it has the largest instrumented volume, is the IceCube experiment at the South Pole[44]. IceCube has detected high-energy neutrinos at a rate above the expectation from the atmospheric neutrino background[29], but the number of events to date is small and there is no statistically significant evidence for point sources. This situation will doubtless improve over time.

Other methods of detecting high-energy neutrinos have been proposed, although most are presently still at the stage of R&D or feasibility studies.

- *The Askaryan effect* is a transient radio signal produced when fast particles travel through a dielectric medium (it's a form of Cherenkov radiation). It should in principle be possible to use this effect to detect the electromagnetic shower produced when a very-high-energy neutrino interacts in a radio-transparent medium such as ice or rock (but not liquid water). The ANITA balloon experiment[45], for example, uses the Antarctic icecap as the radiator and is sensitive to neutrinos with energies $> 10^{18}$ eV; to date (after two flights), no significant signal has been observed[46]. Other Askaryan-based searches have used radio telescopes as detectors and the Moon as the radiator.

- *Acoustic detection* of neutrinos relies on the fact that at extreme energies ($\sim 10^{20}$ eV), neutrino interactions are not weak—the W and Z are effectively massless at these energies—so neutrinos will initiate an electromagnetic shower when they penetrate material. In the ocean, the energy dumped by the shower into a narrow cylinder of water will result in a pressure pulse, which can be detected by hydrophones. This has the advantage that the range of sound in water is very large (so a large volume can be instrumented with a small number of detectors) and that hydrophones are off-the-shelf equipment; the disadvantage is that the ocean is a very noisy place, and sophisticated signal processing is required to pick out the characteristic bipolar pulse shape of a neutrino event. Also, the threshold is *very* high, so the expected rates are correspondingly low. Nevertheless, this is such an attractive idea that several feasibility studies have been conducted, including the Sheffield-led ACORNE[47] experiment using a hydrophone array off the west coast of Scotland.

1.6 Dark matter

Dark matter is *the* classic example of particle astrophysics: it is a dominant constituent of the universe and has important effects in cosmology and astrophysics, but both its theoretical explanation and its detection and identification rely on particle physics. However, dark matter is covered in detail in PHY326[5], so I will only summarise the main points here. For a good review article on this material, consult Feng[48].

1.6.1 Astrophysical and cosmological evidence for dark matter

The original astrophysical evidence for dark matter was dynamical: the orbital motions of stars and gas in galaxies, and of galaxies in clusters of galaxies, are too fast to be accounted for by the luminous material. This was first noted by

Fritz Zwicky in 1933 (galaxies in the Coma cluster), and subsequently studied in detail by Vera Rubin and colleagues (rotation curves of spiral galaxies).

This original evidence has now been supplemented through a number of independent routes:

- the temperature profile of the *intracluster medium* (extremely hot, low-density gas) that pervades rich clusters of galaxies, measured using its X-ray emission, shows that the gas mass (which greatly exceeds the mass of the galaxies themselves) accounts for only about one-sixth of the total gravitational mass;
- studies of both weak and strong *gravitational lensing* show that the lensing mass is larger and more widely ditributed than the luminous mass;
- simulations of *large scale structure* cannot reproduce the observed distribution of galaxies without incorporating dark matter;
- analysis of the *power spectrum of the cosmic microwave background* shows that cold dark matter must account for about 25% of the total energy density of the universe.

The astrophysical and cosmological evidence also provides information about the nature of dark matter. The abundances of the light isotopes ^{2}H, ^{4}He and ^{7}Li, which are produced in the early universe, determine the baryon-to-photon ratio η, or equivalently the density of baryonic matter Ω_{b0}, where Ω is the density in units of the critical density. This is found to be $\Omega_{\mathrm{b0}} \simeq 0.04$, which is about 10 times greater than the stellar density (so most baryonic matter is not luminous), but about 6 times less than the matter density inferred from the cosmic microwave background or the gravitational potentials of rich clusters. This implies that the dark matter does not participate in nuclear reactions, and must therefore be *non-baryonic*.

Galaxy redshift surveys and weak lensing surveys, which measure the large-scale distribution of matter, along with the power spectrum of the cosmic microwave background, can be used to infer the dynamical behaviour of dark matter at $z \sim 3000$, when the energy densities of matter and radiation were comparable. If the dark matter were relativistic at that time, i.e. moving with $v \sim c$, it is said to be *hot dark matter*, whereas matter that is non-relativistic at that time is said to be *cold dark matter* (the intermediate case, where the particles are mildly relativistic, is unsurprisingly known as *warm dark matter*). This is important for structure formation: hot dark matter will not be confined in small gravity wells, and will tend to form structure "top-down" (very large structures form first and then fragment into smaller ones), whereas cold dark matter will form structures the size of small galaxies, which then clump together to form larger objects ("bottom-up" structure formation). The presence of small-scale features in the cosmic microwave background, and comparisons of galaxy redshift surveys with simulations of structure formation, unambiguously prefer cold dark matter; in fact, the absence of any significant effects from hot dark matter is what allows an upper limit on neutrino masses to be set using CMB data.

1.6.2 Dark matter candidates

The astrophysical and cosmological evidence outlined above leads to the following requirements for dark matter:

1. it must not absorb or emit light (from the fact that it is not seen to do so)—this implies that it does not interact electromagnetically;

2. it must not be hadronic (from the conflict between the baryonic density as inferred from light elements and the CMB and the total matter density as inferred from gravitational potential measurements and the CMB);

3. it must be non-relativistic at $z \sim 3000$ (from structure formation);

4. it must be stable or very nearly so (from the fact that the density inferred from the CMB, at $z \simeq 1100$, agrees with the density measured locally in galaxy clusters).

This does not match any Standard Model particle. The closest match is the neutrino, which is stable, neutral and weakly interacting, but would be relativistic at radiation-matter equality and is therefore hot dark matter. In addition, combining the mass limit on the electron neutrino from tritium beta decay (2 eV) with the mass differences from oscillations implies that the total neutrino mass cannot be more than about 6 eV/c^2, which is not enough to account for all the dark matter anyway.

It follows that candidate particles for dark matter must represent physics beyond the Standard Model, which in turn implies that the detection and identification of such particles would be a major step forward in particle physics as well as astrophysics and cosmology. Unlike the scalar fields introduced *ad hoc* to explain inflation and dark energy, many dark matter candidates have the great advantage that they were originally postulated in the context of particle physics, and only later turned out to have relevance to the dark matter problem.

An extensive list of dark matter candidates, all with independent motivation beyond solving the dark matter problem, is given in table 1 of [48]. In this section I will discuss the two most favoured options: weakly interacting massive particles (WIMPs) and axions.

Weakly interacting massive particles are, as the name suggests, particles with neutrino-like interactions but much higher masses. They will therefore be moving much more slowly at any given temperature (since $mv^2 \propto T$), and hence will be cold rather than hot dark matter. Cold dark matter with standard weak interactions naturally decouples from ordinary matter at a point in the history of the universe that yields the right sort of relic density to account for dark matter[48], so it appears that WIMPs might account for dark matter with relatively little fine-tuning (the so-called “WIMP miracle”).

Many extensions to the Standard Model predict a WIMP. This is because extensions to the Standard Model must avoid predicting things that clearly do not happen, such as rapid proton decay. The extra particles predicted by beyond-Standard-Model (BSM) theories often do have the potential to mediate proton decay, so the proton is protected by introducing an extra quantum number that prevents this. The usual consequence is that the lightest BSM particle has to be stable, because conservation of the new quantum number prevents it from decaying into non-BSM particles (just as the proton is stable because conservation of baryon number prevents it from decaying into non-baryons). If this stable particle is electrically neutral, it is a potential WIMP candidate (and if it isn’t neutral, the theory in question is ruled out, because a heavy stable *charged* particle would interact with photons and atomic matter, and would be astrophysically obvious).

The most frequently considered extension to the Standard Model is supersymmetry, which predicts that each Standard Model particle has a partner differing in spin by half a unit (so fermions have bosonic partners and vice versa). There are many variants of supersymmetry, but most conserve a new symmetry called R-parity and consequently have a stable lightest supersymmetric particle (LSP). Because none of the supersymmetric particles has yet been discovered, general supersymmetric models have a very large number of free parameters and are difficult to deal with: for this reason, it is usual to make simplifying assumptions that reduce the number of parameters to a manageable level. The most widely used simplified model is the *Constrained Minimal Supersymmetric Standard Model* (CMSSM), also known as *minimal supergravity* (mSUGRA), which has only 4 free parameters and one undetermined sign. In the CMSSM, the lightest supersymmetric particle is normally the lightest *neutralino*, χ_1^0, which is a mixture of the partners of the two neutral Higgs bosons (recall that supersymmetry requires an extra Higgs doublet), the Z, and the photon.

The neutralino is a weakly interacting stable massive particle, as required of a dark matter candidate. Its mass is not known, but is typically assumed to be of order 100–1000 GeV/c^2. It is a Majorana particle, i.e. it is its own antiparticle, so two neutralinos can annihilate via the weak interaction into a fermion-antifermion pair, $\chi_1^0\chi_1^0 \to f\bar{f}$, or a pair of gauge bosons, $\chi_1^0\chi_1^0 \to W^+W^-$. Unfortunately, as the neutralino is a fermion, the $f\bar{f}$ channel is spin suppressed: because the weak interaction is left-handed, the f and $\bar{f}$ want to be left- and right-handed respectively, for a total spin of 1, whereas the Pauli exclusion principle requires that the two annihilating neutralinos, being identical fermions, must have opposite spins, for a total of 0. The result is that the annihilation cross-section is smaller than a typical weak cross-section, which means that the neutralinos decouple earlier and have a higher relic density than they otherwise would. This rather spoils the "WIMP miracle" mentioned above: for most of the CMSSM parameter space, neutralinos would yield too high a relic density to be consistent with observation. However, there are thin slivers of parameter space that are not excluded, and supersymmetry is a highly popular BSM theory among theoretical particle physicists, so the neutralino model for dark matter is very popular despite this disadvantage.

Supersymmetry is not the only extension to the Standard Model that yields a viable WIMP candidate. Feng[48] discusses the predictions of theories with extra spatial dimensions (widely studied because superstring theory, which requires extra dimensions, is regarded as a good candidate for quantum gravity), and mentions several more exotic possibilities. Fortunately, the methods and results of direct detection experiments (see below) do not depend strongly on the assumed nature of the WIMP, since they simply measure the nuclear recoil when a WIMP scatters elastically off a nucleus in the detector. Indirect detection experiments, which look for the products of WIMP annihilation in regions of enhanced WIMP density, are much more model dependent, since WIMPs from different theories will generally have different annihilation branching ratios.

In contrast to WIMPs, *axions* have extremely low masses, in the μeV–meV range. This makes them lighter than at least one neutrino species (the squared mass differences deduced from neutrino oscillations guarantee that the heaviest neutrino species must have a mass of at least 0.05 eV/c^2). It is therefore surprising that axions qualify as cold dark matter—we might expect that, like neutrinos, they would be relativistic at the time when structures start to form. The reason that this is not the case is that axions are not produced in thermal equilibrium, but rather as the result of a phase transition in the early universe (either during or just after inflation).

Axions are *pseudoscalar* particles (spin 0, but negative parity). In BSM physics, the axion field was introduced to solve the *strong CP problem* of the Standard Model—the puzzle of why the strong interaction seems to conserve CP exactly, even though there is a term in the Standard Model Lagrangian that should allow CP violation. The classic observable is the electric dipole moment (EDM) of the neutron, whose "natural" value is of order 10^{-16} e cm. The current best limit is $< 2.9 \times 10^{-26}$ e cm, which requires fine-tuning to a few parts in 10^{10}. Introducing the axion field solves this problem by generating the relevant term in the Standard Model dynamically, with the result that it relaxes to a very small value as the field attains its stable minimum[5].

The axion pseudoscalar field introduces a new energy scale, the *axion decay constant* f_a, which is extremely large ($\sim 10^{12}$ GeV). This is why the phase transition that is assumed to generate axion dark matter takes place so early (it naturally occurs at $T \sim f_a$). The reason that the axion mass is not itself at this level is that it's a *pseudo-Goldstone boson*[49]: Goldstone bosons, which arise from spontaneous symmetry breaking, are massless regardless of the energy scale of the field, and pseudo-Goldstone bosons, though not exactly massless, are generally very light (the pion is much lighter than other non-strange mesons, such as the ρ, because it is a pseudo-Goldstone boson related to chiral symmetry breaking in QCD). The mass of the axion is given by[48]

$$m_a \simeq \frac{\sqrt{m_u m_d}}{m_u + m_d} m_\pi \frac{f_\pi}{f_a}, \tag{1.12}$$

where m_u and m_d are the current masses of the u and d quarks ($\sim$4 and $\sim$8 MeV/c^2 respectively), m_π is the pion mass (135 MeV/c^2) and f_π is the pion decay constant (93 MeV). Putting in the numbers gives

$$m_a \simeq 6\ \mu\text{eV} \frac{f_a}{10^{12}\ \text{GeV}}.$$

Axions couple to photons: there is a term $-g_{a\gamma\gamma}\alpha\mathbf{E}\cdot\mathbf{B}$ in the effective axion Lagrangian which implies an interaction vertex connecting one axion line with two photons. This appears to contradict our earlier requirement that dark matter should not interact electromagnetically, but the coupling is very weak, so this is not in itself a problem. However, it has implications which do place constraints on the axion properties:

1. Axions can decay to two photons. If axions are to be viable dark matter candidates, the lifetime for this decay must be at least equal to the age of the universe. This puts an upper limit of 20 eV/c^2 on the axion mass.

2. Conversely, photons can interact to produce axions. If this occurs inside a star, the axions then escape (like neutrinos), carrying away energy and causing deviations from the predictions of stellar evolution theory. The lack of such deviations, particularly as regards the lifetimes of stars in globular clusters and the length of the neutrino pulse from SN 1987A, forces the axion mass to be <10 meV/c^2 (note: meV not MeV!)[50].

[5] At this point, you may be developing an understandable suspicion that theorists' response to every fine-tuning problem is to introduce a new scalar (or, in this case, pseudoscalar) field and try to arrange that the associated particle is sufficiently difficult to detect that its non-observation does not disprove the model! However, it is worth noting that the Higgs field has exactly these characteristics—it is a scalar field introduced to deal with the problem of non-renormalisable masses for the W and Z—but does in fact seem to exist. The fact that these mechanisms look contrived doesn't necessarily mean that they aren't right.

3. For the same reason, the Sun should be an axion source. Constraints from helioseismology and the solar neutrino flux, which would both be altered by this (as axions carry energy out of the Sun, its core temperature has to be slightly hotter to compensate), place upper limits on the axion mass which are slightly weaker than the limits from SN 1987A.

In some but not all axion models, axions also couple to $e^{\pm}$. This coupling also has an effect on stellar evolution, since stellar interiors are dense plasmas and hence very rich in electrons. This coupling places an upper limit of <10 meV/c^2, similar to the SN 1987A bound, from the rate at which white dwarfs cool and the point at which helium fusion ignites in red giant stars.

Calculations of the relic density of axions from the phase transition are model dependent: the result depends on whether the axion phase transition is assumed to take place before, during or after inflation. The smaller the axion mass, the larger the relic density: the usually quoted lower bound (to avoid producing too much dark matter) is 6 μeV/c^2, though this can be dodged by some fine-tuning of parameters[48]. Axions are also produced thermally, giving a relic density $\Omega_a^{\rm th} \sim 0.22(m_a/80\ {\rm eV}/c^2)$: this is negligible for the axion masses allowed by astrophysical constraints. Thermally-produced axions would be hot dark matter, so the fact that we do not see evidence for hot dark matter places an upper bound on their mass, but this bound is weaker than the astrophysical constraints.

Combining the lower limit from relic density calculations with the upper limits from astrophysical observations, a convenient ballpark estimate for the allowed axion mass range is[48]

$$6\ \mu{\rm eV}/c^2 < m_a < 6\ {\rm meV}/c^2,$$

though the bottom end in particular is rather model dependent. The relic density is roughly proportional to f_a ($\propto f_a^{7/6}$, according to the PDG review[51]), so if axions are to account for all or most of the dark matter their mass needs to be towards the lower end of the allowed range, around 10 μeV/c^2. There appears to be no good theoretical reason to prefer this mass region *a priori*, so explaining the dark matter in terms of axions does require a bit of fine tuning.

Other dark matter candidates include sterile (right-handed) neutrinos, gravitinos, “hidden sector” particles and even axinos (one might feel at this point that the hypothetical supersymmetric partner of a hypothetical particle is one level of hypothesis too many). All of these have some motivation beyond simply explaining dark matter—for example, sterile neutrinos may help to explain the smallness of neutrino masses, as discussed above—and they are discussed in detail by Feng[48]. However, WIMPs and axions remain the favoured candidates among experimental particle physicists.

1.6.3 Detection of dark matter

Methods of detecting dark matter fall into four categories:

1. *direct detection*—the particle interacts in your detector and you observe the consequences of the interaction;

2. *indirect detection*—the particle interacts or annihilates elsewhere, and you detect the products of the interaction;

3. *cosmology*—the particle has some characteristic and identifiable impact on, e.g., light-element abundances or the properties of the cosmic microwave background;

4. *accelerator experiments*—the particle is produced in particle physics experiments and has an identifiable signature.

The different techniques all have individual advantages and disadvantages. Production at accelerator experiments generally provides the most precise information about the properties of the candidate, but does not guarantee that it is actually present in the universe (in particular, a particle with a lifetime comparable to, say, the neutron would appear completely stable to an LHC experiment, but could not possibly account for dark matter). Cosmological constraints are often model dependent, for example assuming that the universe has a flat geometry. Direct detection seems like the most foolproof method, but signals may be faked by background: to date, several direct detection experiments have reported positive signals, but these are not all consistent with each other, and they all contradict upper bounds reported by other experiments. Similarly, purported signals from indirect detection experiments may be due to more conventional astrophysical phenomena. The ideal signal would be confirmed by more than one method—for example, neutralinos could be detected both directly and indirectly, and also produced at the LHC; light sterile neutrinos (warm dark matter) could be detected through their effect on big-bang nucleosynthesis and structure formation, and through anomalies in neutrino oscillation experiments. A good second best to this is a signal with a particularly strong signature: for example, directional WIMP detectors could potentially observe a signal modulated on timescales of a *sidereal*, rather than solar, day, which would strongly disfavour terrestrial background; the mass and other properties of a neutralino candidate detected at the LHC might yield exactly the right relic density when plugged into the appropriate equations.

Detection of WIMPs

On the rare occasions when WIMPs scatter off atomic nuclei, they are massive enough that the recoil of the struck nucleus is significant. This is the basis of all direct-detection WIMP searches. The recoil energy of the nucleus is small—typically tens of keV—so the detectors must be sensitive and well protected from cosmic rays and ambient radioactivity; they are typically located underground, in deep road tunnels or mines.

The nuclear recoil has a number of detectable effects: it may *ionise* neighbouring atoms; it may induce *scintillation* in suitable materials; the increase in energy dumps *heat* into the target, which can be detected in cryogenic experiments cooled to sufficiently low temperatures. It is common for experiments to use more than one of these, in order to distinguish nuclear recoils (signal) from electron recoils (background).

The interaction between the WIMP and the nucleus may be *spin-independent*, where the scattering amplitude does not depend on the relative spins of the WIMP and the nucleon it hits, or *spin-dependent*, where it does. In the former case, when a WIMP strikes a nucleus of atomic mass A, the scattering amplitudes off all the constituent nucleons add coherently, so the event rate scales as A^2. For spin-dependent coupling, in contrast, the cross-section is driven by the net spin of the nucleus, and is not strongly dependent on A. Consequently, direct-detection experiments tend to be much more sensitive to spin-independent interactions, though a careful choice of target material can increase the sensitivity to spin-dependent processes.

Direct detection experiments fall into a number of categories[52].

1. *Cryogenic solid state* detectors consist of high-purity crystals of some suitable material, typically germanium, silicon or calcium tungstate. They read out the heat deposited, along with either the ionisation (CDMS and EDELWEISS, both using germanium, along with silicon for CDMS) or scintillation (CRESST, using Ca WO_4).

2. *Noble liquid* detectors use liquid xenon or liquid argon as the target medium and detector. In both cases, the primary read-out is through scintillation; some detectors also read out ionisation, by using an electric field to drift the electrons into a gaseous region above the liquid. These dual-phase experiments (e.g. XENON, LUX) use the comparison between scintillation and ionisation for background rejection; the single-phase LAr experiments (e.g. DEAP) use the shape of the scintillation pulse. One inconvenient feature of both liquids is that the scintillation is in the far UV (178 nm for LXe, 128 nm for LAr), which means that it must be read out using special-purpose photomultiplier tubes (the glass windows of standard PMTs are not transparent at these wavelengths). Argon has a naturally-occurring radioactive isotope, ^{39}Ar, whose signals can be rejected using pulse shape information, but which might cause rate problems for large detectors; as ^{39}Ar is created by cosmic-ray bombardment, a possible solution is to use argon from underground sources, which has been shielded from cosmics and is therefore lower in ^{39}Ar.

3. *Scintillating crystal* detectors, principally NaI, are a long-established class of dark matter experiment. The scintillation pulse shape can be used to discriminate against background. The most famous, or possibly infamous, experiment of this class is DAMA/LIBRA, which claims a positive dark matter signal based on annual modulation (the Earth's orbital motion causes a small variation (±7%) in the net speed of the Earth around the Galactic centre; as the WIMP "gas" comprising the dark halo of the Galaxy has no net rotation, this variation in net orbital speed creates a variation in the resulting "WIMP headwind", and hence in the rate of events). This signal is in conflict with the upper limits reported by cryogenic crystal and noble-liquid experiments, which makes it difficult to interpret as dark matter, but no convincing explanation in terms of background has been put forward either.

4. *Superheated liquid* detectors observe the bubbles created when a WIMP interaction dumps enough extra energy into the liquid to induce a phase transition. Electron recoils dump their energy over a longer distance, because the recoiling electron travels further, and do not create bubbles. This is basically the same idea as that traditional staple of particle physics, the bubble chamber, and indeed some of these experiments do take the form of bubble chambers; others use superheated droplets. These experiments tend to have lighter target nuclei and be sensitive to lower-mass WIMPs; they have the advantage that many suitable liquids incorporate fluorine, which is a target of choice for spin-dependent interactions.

5. *Directional detectors* are a somewhat different category of direct-detection experiment, aiming to establish a definitive WIMP signal by looking for

> the *diurnal* modulation in the direction of the WIMP headwind caused by the tilt of the Earth's axis relative to the direction of the Sun's orbital motion. Unlike the small annual modulation, this is a large effect (the direction in a detector at mid-northern latitudes changes by about 90°), and should be unambiguous if observed. The pioneering experiment in this field is DRIFT[53], which has strong involvement from Sheffield. The main problem with directional detectors is that they have a gaseous target (so that the recoiling nucleus will travel far enough for its direction to be measured), and hence a low target mass, which limits their sensitivity at present.

The best direct-detection limits to date come from the LUX liquid xenon experiment[54]. Together with the earlier results from the XENON experiment, which include a special analysis for lighter WIMPs[55], these appear to rule out a number of claimed signals for low-mass WIMPs from other experiments (DAMA/LIBRA, CoGeNT, CDMS-II Si, and CRESST), which in fact are not very consistent among themselves either. The inconsistency of results reported by different experiments suggests either an incorrect physical picture, or (probably more likely) some uncontrolled systematic errors.

Indirect searches for WIMPs have more model dependence than direct searches, because theoretical calculations of branching ratios are needed to convert observations of presumed decay products into WIMP limits. In general, indirect searches are predicated on the assumption that the WIMP is a SUSY neutralino (see above), and that in regions of higher than normal WIMP number density, $\chi_1^0\chi_1^0$ annihilations will occur and produce detectable secondaries.

"Regions of higher than normal WIMP number density" are generally assumed to include the Sun, the Galactic centre and dwarf galaxies. WIMPs passing through the Sun can become gravitationally captured by the Sun if they scatter off a nucleus in the solar interior. Initially, the resulting orbit will probably be highly elliptical, but repeated scatters as the WIMP passes through the Sun on each perihelion passage will eventually bring the WIMP into thermal equilibrium with the solar interior, and it will settle down close to the centre of the Sun (as it is more massive than the Sun's hydrogen, its equilibrium velocity will be smaller, so it will fall into the central core). Therefore, over the 4.6 Gyr of the Sun's existence, it should now have collected a central clump of WIMPs, at high enough concentration for annihilations to be taking place; in most models, the WIMP capture cross-section is such that the WIMP concentration in the Sun has reached equilibrium, with the rate of new captures balanced by the loss due to annihilation.

Obviously, most of the products of annihilation are promptly absorbed by the dense material of the solar core. The only secondaries that are likely to escape are neutrinos. Because of the spin suppression mentioned above, the cross-section for neutralinos to annihilate directly into $\nu\overline{\nu}$ is negligible, but fortunately neutrinos are generated as decay products of annihilations into W^+W^-, ZZ, $\tau^+\tau^-$, $b\bar{b}$ and, if kinematically allowed, $t\bar{t}$. The neutrinos from the first three channels are "harder" (have higher average energy) than those from the quark decays, because the former are two-body decays (such as $W^+ \to \mu^+\nu_\mu$) and the latter three-body (such as $b \to c\mu^-\overline{\nu}_\mu$); typically, the neutrinos from two-body decays will have energies of the order of half the neutralino mass, and those from three-body decays about a quarter.

Such neutrinos could be detected by neutrino telescopes such as IceCube. They have low energies by IceCube standards, a few tens to a few hundreds of GeV depending on the neutralino mass, but the atmospheric neutrino background can be significantly reduced by requiring that the direction of the incoming neutrino is consistent with the position of the Sun. Because the Sun is largely made of hydrogen, which—being a single proton—has spin $\frac{1}{2}$, and no A^2 enhancement, indirect searches for neutralino annihilation in the Sun are most sensitive to spin-dependent interactions, where they compete favourably with direct searches[56]; they are not competitive for spin-independent cross-sections.

The supermassive black hole in the Galactic centre is another possible concentrator of WIMPs. In this case, the range of products that could escape and be detected is much greater, and includes γ-rays and antiparticles such as positrons (which have much lower astrophysical backgrounds than the corresponding particles). These have the advantage of much larger interaction cross-sections than neutrinos, making it much easier to obtain a statistically significant sample; the disadvantage is that the putative source is much less well understood (the central core of the Galaxy is a complicated region, containing much more than just a black hole). Charged particles have the additional disadvantage that, as discussed above in the context of cosmic rays, their directions are effectively scrambled by the Galaxy's magnetic field: if there is in fact an excess of positrons coming from the Galactic centre, they will not *hit your detector* from the direction of the Galactic centre. This makes eliminating backgrounds much more difficult.

Candidate dark matter signals include a rising fraction of positrons in primary cosmic rays at energies above 7 GeV, seen by PAMELA, *Fermi*–LAT and AMS-02[57]. This signal is unquestionably real, but not unquestionably due to dark matter: there are alternative explanations involving conventional astrophysical sources such as pulsars (some of which are known sources of TeV γ-rays, and must therefore accelerate electrons to very high energies, as discussed later). In fact, interpreting the signal as dark matter is problematic[48], so the conventional explanations are probably more likely at this point.

Another potential signal is the observation of γ-rays from the Galactic centre by *Fermi*–LAT: this is viewed by Daylan et al.[58] as providing "a compelling case for annihilating dark matter", but Gómez-Vargas et al.[59] use the same *Fermi*–LAT signal to construct upper limits on the annihilation cross-section, and argue that "one may interpret these results as implying that vanilla WIMP models and contracted DM profiles are incompatible with the *Fermi* data." The *Fermi*–LAT Collaboration themselves have not analysed their Galactic centre data in terms of dark matter (though their view may possibly be deduced from the fact that [59] is listed on the *Fermi*–LAT publications page, whereas [58] is not); they do, however, present an upper limit on dark matter annihilation based on observations of the γ-ray flux from dwarf galaxies[60].

In summary, it is fair to conclude that neither direct not indirect searches have yet resulted in a definite signal for dark matter. Such positive hints as have been reported are controversial, either because of the possibility of alternative explanations or because of inconsistency with other experiments. There remains, however, the possibility of identifying WIMP candidates through their production in accelerator-based experiments.

The width of the Z^0 as measured at LEP constrains the number of light neutrino species to 2.984 ± 0.008[61]. This limit restricts WIMPs that couple to the Z to masses above 45 GeV/c^2; however, it is possible that a relatively light WIMP might evade this bound by not coupling to the Z—this is true for the lightest neutralino in some regions of SUSY parameter space.

WIMPs with masses below 1 TeV/c^2 or so could be produced by pp collisions at the LHC. The WIMP itself, being stable and weakly interacting, would not be detected, but its presence could be inferred from missing transverse energy and momentum. No positive signals have been reported from any of the LHC experiments (for this or any other SUSY particle); however, as with the LEP bound, exclusions derived from this are generally only valid in certain regions of parameter space. For example, Calibbi et al.[62] use the relic density and accelerator limits to constrain the MSSM parameter space for light neutralinos ($M_{\chi_1^0} < 30$ GeV/c^2) and then investigate how this parameter space is further constrained by LHC searches. They find, firstly, that the claimed direct-detection signals for light WIMPs cannot be accommodated within the MSSM, because the large WIMP-nucleon scattering cross-section implies a region of parameter space excluded by other measurements, and secondly that the allowed MSSM region is strongly constrained by existing searches and will be closed completely with a small increase in sensitivity. However, they caution that larger neutralino masses "would open the possibility of satisfying the relic density constraints with compressed spectra that can, at the same time, (i) evade the LEP searches for light sfermions, (ii) be insensitive to constraints from Z-pole observables, (iii) be very hard to be tested at the LHC." This is a somewhat disheartening conclusion, although they do not specify whether such an "invisible" SUSY sector would also be able to evade detection by direct or indirect WIMP searches.

Detection of axions

The detection of axions relies on the $a\gamma\gamma$ coupling referred to above. The usual technique is to persuade the axions to convert to photons by interacting with a magnetic field, which is essentially a source of virtual photons: $a\gamma^* \to \gamma$. This process, which is more usually used to produce π^0s, is known as the *Primakoff effect.*

The ADMX experiment[63] consists of a high-Q resonant cavity installed in an 8 tesla superconducting solenoidal magnet. Tuning the cavity to a specific frequency enhances the probability that axions of the appropriate mass will convert to microwave photons: this is then detected as a very small increase in the microwave noise from the cavity. ADMX is currently the only experiment that is sensitive to axions in the mass range where they would make a significant contribution to the dark matter (see above). The problem with this technique is that any given resonance frequency picks out only one value of the axion mass, with a very small bandwidth defined by the Q of the cavity, so the experiment has to be run in a scanning mode where one collects enough statistics to rule out axions at mass m, retunes the cavity to mass $m + \delta m$, collects more statistics, retunes, and so on. This is a *very* slow process, such that years of scanning covered only a small part of the interesting mass range. ADMX is currently undergoing an upgrade that will reduce the temperature of the cavity and therefore the thermal noise, improving the signal to noise ratio and allowing much faster scanning, so the remaining mass range should be covered considerably more quickly.

Another possibility is to make use of the fact that the Sun should be a strong axion source, and point an "axion telescope" at the Sun. The CAST experiment[64] recycled a prototype LHC magnet—a 9 tesla, 10 m long, dual bore dipole—instrumented with X-ray detectors. CAST is sensitive to higher-mass axions, which could not contribute significantly to dark matter. Its sensitivity is comparable to astrophysical bounds, though with very different systematic errors. In general, solar axion telescopes do not probe the right axion mass region for dark matter, though of course the axion can solve the strong CP problem whether or not it also solves the dark matter problem.

An alternative approach, conceptually equivalent to detecting WIMPs in LHC data, is the "light shining through walls" method: conversion of a photon to an axion would allow the photon to "tunnel" through an opaque barrier. Obviously this is a very improbable occurrence, so an extremely intense light source (usually a powerful laser) is required. The technique is sensitive to any light particle that couples to photons, but no current or planned experiment has the sensitivity needed to detect axion dark matter. An example of this type of detector is the ALPS-II experiment at DESY[65].

1.7 Summary

This chapter has presented a quick survey of the core disciplines of particle astrophysics. It should have become apparent that particle astrophysics has a very broad scope, from highly theoretical topics such as the mechanics of inflation and the physics of dark energy to the engineering challenges posed by neutrino telescopes and tonne-scale dark matter detectors. Some areas—solar and atmospheric neutrinos, γ-ray astronomy—are established and productive, and have already made major contributions to particle physics and astrophysics respectively, while others, such as dark matter searches, are based on solid observational and theoretical foundations but have yet to bear fruit; a few, such as scalar field models of dark energy, are frankly speculative.

This field is far too large to cover in a single course, and some aspects of it, particularly dark matter and solar neutrinos, are covered elsewhere. However, the three areas of cosmic rays, γ rays and high energy neutrinos are closely related, both in production mechanisms and detector technology, and form a coherent subfield that we might refer to as *high energy particle astrophysics*. In the remainder of this course, we shall focus on this area, covering the evidence for populations of fast particles in astrophysical objects, the means by which particles might be accelerated to such energies, the nature of the sources, and the methods by which high-energy particles are detected and studied in terrestrial experiments.

1.8 Questions and Problems

1. In a universe where the energy density is dominated by matter (as was the case in our universe until relatively recently), the expansion is described by $a(t) = (t/t_0)^{2/3}$, where a is the scale factor and t_0 is the present time. The *horizon distance*, i.e. the furthest distance that you can possibly see, is given by $d_{\text{hor}}(t) = 3ct$.

   (a) Calculate the horizon distance, d_1, at $t = 380\,000$ years.

   (b) Between $t = 380\,000$ years and $t = t_0$, all distances have been stretched by a factor $1/a(t)$. Assuming that $t_0 = 14 \times 10^9$ years, calculate the *present* value of d_1. Call this d_1'.

   (c) Calculate the horizon distance now, i.e. at $t = t_0$. Call this d_2.

(d) What is the angle subtended by distance d_1' at distance d_2? Comment on the implication of this for the temperature of the cosmic microwave background.

[This is the *horizon problem* of the classic Big Bang model, which is solved by introducing inflation.]

2. Using equation (1.6), calculate the critical density, in units of GeV m^{-3}, for $H_0 = 70.0$ km s^{-1} Mpc^{-1}. (Note that 1 Mpc $= 3.09 \times 10^{19}$ km.) Hence calculate the average number density of dark matter particles, assuming that $\Omega_{\mathrm{DM}} = 0.25$, for the case where the dark matter is (i) a WIMP of mass 100 GeV/c^2, (ii) an axion of mass 10 μeV/c^2.

3. Assuming that the refractive index of air at an altitude of 10 km is 1.000095, calculate the minimum energy an electron would need to have to produce Cherenkov radiation.

4. The rotation curves of spiral galaxies are approximately flat, $V =$ constant, at large distances from the centre. Show that this implies that $\rho(r) \propto 1/r^2$, where $\rho(r)$ is the density a distance r from the centre, stating any assumptions that you make.

5. If the Sun is 8.5 kpc from the Galactic centre and is orbiting with a speed of 220 km s^{-1}, what is the local density of dark matter in GeV m^{-3}? State any assumptions that you make.

6. The temperature in the core of the Sun is approximately 15×10^6 K. Estimate the average speed of a captured WIMP of mass 100 GeV/c^2, if it has relaxed into thermal equilibrium.

7. The average Galactic magnetic field near the Sun is about 6 μG (0.6 nT). If this field were uniform (it isn't!), how far would a proton with kinetic energy 1 TeV need to travel for its direction to change by 90°?

Chapter 2

Astrophysical Accelerators: The Observational Evidence

2.1 Introduction

Over the past century or so, we have amassed a great deal of observational evidence for the presence of highly relativistic particles in some astrophysical objects. In this chapter, we will review this evidence and its implications. Where the measurements are not made by conventional astrophysical means, we will also discuss the techniques used for collection and analysis of the data, since any limitations or systematic errors imposed by these may impact on our understanding of the astrophysics.

As in dark matter searches, the evidence for fast particles in astrophysical sources may be divided into "direct" and "indirect". In this case, direct evidence consists of observations of actual relativistic particles, i.e. cosmic rays, while indirect evidence comprises observations of secondary products of fast particles, i.e. everything else. Observations of high-energy γ-rays and neutrinos count as indirect evidence, because electrically neutral particles cannot be accelerated directly: they must be produced secondarily, as a result of either collisions between particles or interactions between particles and magnetic fields.

Direct and indirect detection are regarded as "complementary" in dark matter searches because they have different sensitivities (e.g., mainly spin-dependent for neutrino-based indirect detection compared to mainly spin-independent for direct detection) and different sources of systematic error. However, in principle either direct or indirect detection could suffice for a discovery of dark matter, if the detection were sufficiently compelling (this is not the case for the various claimed "signals" currently extant). The situation for high-energy particle astrophysics is somewhat different, in that it is *not* possible to understand the characteristics of astrophysical accelerators from any one individual technique: it is absolutely essential to synthesise data from multiple sources. Cosmic-ray data establish the existence of protons and heavier ions accelerated to extreme energies, but provide little information about the sites of such acceleration because of deflection by Galactic magnetic fields. In contrast, data from the electromagnetic spectrum (from radio to γ-rays) identify sources, and establish the presence of a population of energetic electrons, but in most cases do not provide evidence either for or against the presence of energetic baryons. Neutrino data could in principle provide both source identification and proof that baryons are being accelerated, but the limited statistics and angular resolution available (see below) imply that source identification will probably have to be done in combination with electromagnetic data.

As summarised in the previous chapter, the principal sources of information on high-energy particle astrophysics are:

1. *cosmic rays*, which demonstrate the existence of astrophysical particle accelerators capable of accelerating protons to energies in excess of 10^{20} eV;

2. *radio emission*, which is clearly non-thermal in origin and requires the presence in the source of both relativistic electrons and a significant magnetic field;

3. *high-energy photon emission* (X-rays and soft γ-rays), which is highly correlated with radio emission and implies similar source properties;

4. *high-energy γ-rays*, which require higher-energy electrons than radio and X-ray emission (in the case where the spectrum is consistent with inverse Compton) and may in some cases imply acceleration of baryons (if the spectrum requires π^0 decay instead of or in addition to inverse Compton);

5. *high-energy neutrinos*, which (at least in the context of the Standard Model) must be produced via secondary interactions of high-energy baryons—there is no plausible way to produce detectable fluxes by purely leptonic interactions—and are therefore diagnostic of acceleration of baryons, if they can be localised to a point source.

These are in roughly chronological order: cosmic rays were discovered around 1912, radio emission in the 1930s, X-ray emission in the early 1970s[66] and high-energy γ-ray emission in the 1990s. High-energy neutrino emission, although long predicted, has only just been observed[29], and we have yet to identify any point sources of high-energy neutrinos.

2.2 Cosmic rays

2.2.1 A brief history

Cosmic rays, in the sense of some form of radiation hitting Earth from above, were established (after some earlier indications) by Viktor Hess[67] in 1912. By taking electroscopes with him on a number of balloon flights, he was able to demonstrate that the amount of ionising radiation increased with altitude, indicating that the source was extraterrestrial rather than terrestrial in nature. One of the balloon flights was undertaken during a solar eclipse, and the lack of any decrease in rate caused Hess to conclude that the radiation did not originate from the Sun.

The term “cosmic rays” came into use in the mid-1920s: the first paper in the ADSABS[68] database using that exact term dates from January 1926[69]. However, the physical nature of this cosmic radiation took rather longer to become clear. Many early researchers, particularly Robert Millikan[70], thought that cosmic rays were electromagnetic, i.e. ultra-high-energy γ-rays—probably because measurements showed them to be extremely penetrating, in contrast to the behaviour of charged particles in the laboratory. With hindsight, we can see that the penetrating component of cosmic rays is composed of muons, but muons were not discovered until 1937, by which time the particulate nature of cosmic rays had been established.

The “smoking gun” demonstrating that primary cosmic rays are in fact charged particles was provided in 1932 by Arthur Compton of Compton scattering fame[71]. Compton’s measurements showed that the flux of cosmic rays varies with latitude, as is “to be expected if the rays consist of electrically charged particles which are deflected by the Earth’s magnetic field”[71], and is certainly not consistent with their being high-energy photons. A year later, two experiments[72, 73] showed that the arrival directions of cosmic rays show an east–west asymmetry, also due to deflection by the Earth’s magnetic field, and indicating that the primary cosmic rays are predominantly positively charged. By 1939, Johnson and Barry[74] were able to argue that the “hard” component of primary cosmic radiation consisted “probably of protons or some other more massive positive ion.”

Although we shall be concerned with cosmic rays as particle astrophysics, we should note their immense contribution to the early history of particle physics. The positron[75], the muon[76, 77], the pion[78] and strange particles[79] were all discovered in cosmic rays; it was only with the discovery of the antiproton at the Bevatron[80], which was designed specifically for the purpose, that accelerator-based experiments took over from cosmic-ray observations as the drivers of progress in particle physics.

2.2.2 Detection of cosmic rays

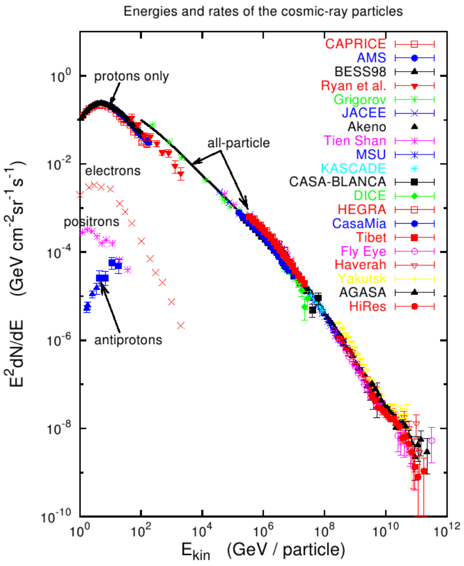


Figure 2.1: The cosmic-ray energy spectrum from 1 GeV to 10^9 GeV, from [81]. Note the large number of different experiments that have contributed to this spectrum.

The energy spectrum of cosmic rays spans an enormous range, from ∼100 eV/nucleon to over 10^{20} eV/nucleon. Clearly, a single detector, or even a single detector technology, is not going to cover the whole of this range, and consequently our understanding of the properties of cosmic rays relies on combining results from many different experiments, as shown in figure 2.1[1].

At the low energy/high flux end of this spectrum, the data come from balloon-borne experiments such as CREAM[82] and space-based platforms such as ACE[83] (at very low energies) and PAMELA[36] (at higher energies). However, at energies above about 10^5 GeV, the cosmic-ray flux is of the order of a few particles per square metre per day, decreasing to a few particles per square kilometre per year at the very highest energies, and this is clearly not practical for experiments that have to satisfy the mass and size constraints imposed by balloon and

[1] Note that this spectrum has been scaled by E^2. It is common practice in displaying the energy spectrum of cosmic rays to scale dN/dE by some power of E, in order to make subtle features of the spectrum more apparent.

rocket launches. Consequently, the highest-energy cosmic rays must be studied using ground-based techniques.

Basic concepts

Ideally, we would like to measure the following properties of primary cosmic rays:

- the *energy spectrum*;
- the *particle composition*;
- any dependence of flux on *time* or *direction*.

Although the Galactic magnetic field makes it almost impossible to identify the sources of cosmic rays on the basis of the direction from which they strike the detector, such information *may* be useful at the very highest energies (where the amount of deflection is relatively small), and may also provide useful data on the magnetic fields close to the solar system. For lower energy cosmic rays, which may originate from the Sun or be significantly affected by solar magnetic activity, time dependence (especially correlation with the solar activity cycle) is important information, which is not only interesting in itself but also has an impact on the planning of space missions and the study of possible effects on the Earth's climate[84].

The equation of motion of a particle of charge Ze and rest mass m in a uniform magnetic field $\mathbf{B}$ is given by

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\gamma_v m\mathbf{v}) = Ze(\mathbf{v} \times \mathbf{B}), \tag{2.1}$$

where $\mathbf{p}$ is the particle's momentum, $\mathbf{v}$ is its velocity and γ_v is the usual relativistic γ factor. This equation motivates the definition of a variable called **rigidity**,

$$R = \frac{pc}{Ze}. \tag{2.2}$$

The importance of rigidity is that ions of equal rigidity will respond in the same way to a given magnetic field. The first implication of this is that all the cosmic rays that reach us from a given point source will have equal rigidity (but quite different energies). The second implication is that the maximum energy for particles from a given astrophysical source will depend on the particle species, with heavier ions having higher maximum emergies. This is because in order for a particle to be accelerated to high energies it must be confined within the acceleration region, and the only way to do this is by a magnetic field. This will only be possible up to a certain critical rigidity (whose value will depend on the size of the source and the strength of the magnetic field); therefore, for a given source type, the cut-off energy will increase $\propto Z$. This has an interesting consequence: if a feature of the cosmic-ray spectrum, for example a change in the slope, is associated with a change in source type, then the feature should be correlated with a shift in composition towards heavier species.

Energy measurement

Balloon-borne and space-based cosmic-ray experiments are similar in design to accelerator-based particle physics experiments, and use similar techniques to extract information. Given that primary cosmic rays are charged particles, there are two basic methods of measuring their energies: magnetic spectrometers, which determine the momentum of the incident particle by measuring the curvature of its track in a known magnetic field, and calorimeters, which measure the energy deposited when the incident particle passes through matter.

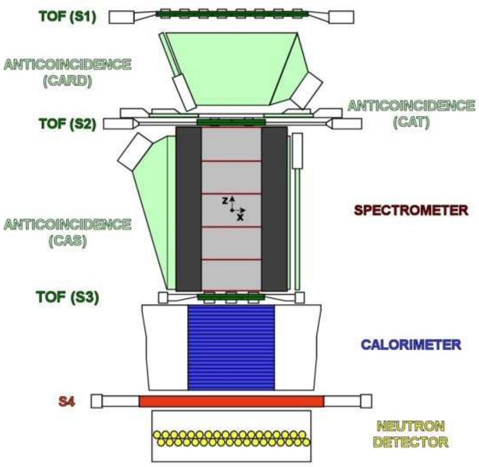


Figure 2.2: Schematic of the PAMELA instrument[36]. PAMELA uses a magnetic spectrometer to measure the rigidity and the sign of the electric charge, and a calorimeter to distinguish electrons and positrons from antiprotons and protons, respectively, with the same momenta. The anticoincidence counters veto particles that came in from the side instead of traversing the whole detector, and the time-of-flight (TOF) system measures velocity, which provides particle identification for those particles with speeds measurably less than c.

A typical example of a magnetic spectrometer is the PAMELA satellite[36] (see figure 2.2). The magnetic spectrometer is based on a permanent magnet with a field of 0.43 T, instrumented with silicon-strip detectors. It is capable of measuring rigidities up to 800 GV if all 6 silicon-strip planes are hit, and 500 GV if 5 planes are hit. The electromagnetic calorimeter beneath the spectrometer is composed of tungsten plates interleaved with silicon-strip detectors, and can measure electron energies with a precision of 5.5% from 10 to 300 GeV; it also separates electrons (which shower) from protons of the same rigidity (which don't). Using a permanent magnet has the advantage that no power supplies or cryogenic systems are needed, which increases the lifetime of the experiment (the lifetimes of cryogenically cooled space-based experiments are generally limited by their supply of coolant).

The AMS-02 magnetic spectrometer[37] is similar in concept to PAMELA, but uses a superconducting electromagnet in preference to a permanent magnet, and has more sophisticated particle identification (see below). As it is mounted on the International Space Station, AMS-02 is easily accessible for resupply of liquid helium, so lifetime considerations are less of an issue in this case.

An example of calorimetric energy measurement is the CREAM balloon-borne experiment[82, 85] (see figure 2.3). The principal components of this experiment are the two transition radiation detectors (TRDs), the silicon charge detector (SCD) and the calorimeter.

Transition radiation[86] is emitted when a charged particle passes from one medium to another (hence the name). The key property of transition radiation for high-energy physics and particle astrophysics is that, for highly relativistic particles with $v \simeq c$, the energy emitted in transition radiation is proportional to the γ-factor of the particle. Since $\gamma = E/mc^2$, if the particle's mass is known, its energy can therefore be deduced. In the case of CREAM, the mass of the incident particle is inferred from its charge, which is measured by the silicon charge detector (SCD).

CREAM also measures particle energy using a sampling calorimeter made up of tungsten sheets interleaving with scintillating fibres. As the calorimeter is quite thin, it is preceded by two graphite targets intended to initiate hadronic showers.

The remaining components of CREAM are there principally as vetos to improve the performance of the instrument. Non-relativistic particles behave differently from ultra-relativistic particles as regards transition radiation, so a Cherenkov detector is installed to veto these (only relativistic particles will produce Cherenkov radiation). Secondary particles from the hadronic shower that are back-scattered through the detector may compromise charge measurements, so the Top Charge Detector (TCD) is installed to tag such back-scatters by their later arrival (they will hit the TCD at least 3 ns after the primary particle). In contrast to PAMELA and AMS-02, which are very similar to accelerator-based particle physics experiments, CREAM looks quite unfamiliar to particle physicists, but the individual detectors all have analogues in particle physics (ATLAS, for example, uses TRDs).

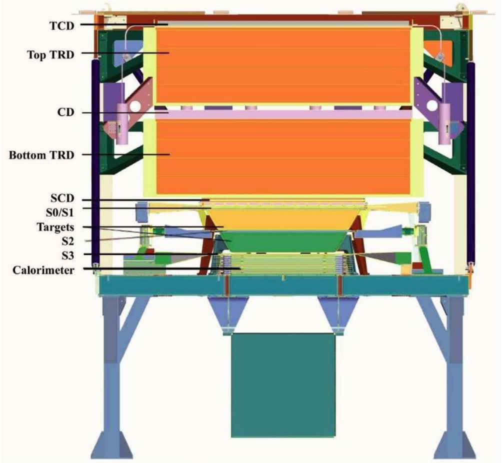


Figure 2.3: Schematic of the CREAM-III instrument[85]. CREAM uses transition radiation detectors (TRDs) and a calorimeter for energy measurement, coupled with a charge detector (SCD) to measure the charge of the particle and thus infer its mass. The top charge detector (TCD) is there to identify and veto back-scattered particles, which otherwise introduce noise into the TRD measurement.

Non-magnetic detectors have a weight advantage—magnets tend to be heavy—and are often more compact, since a magnetic tracking detector needs to have some lever arm to measure the curvature of the track. However, magnetic spectrometers have the key advantage that they can determine the *sign* of the electric charge. This is essential for distinguishing matter (electrons, protons) from antimatter (positrons, antiprotons); observations of antimatter are important in indirect searches for dark matter, and valuable in their own right since the matter-antimatter asymmetry is one of the great unsolved problems of cosmology (see section 1.2.2).

Both balloon-based and space-based experiments measure the primary particle directly, and they do so in an energy regime which overlaps with terrestrial accelerators. Therefore, most of these detectors have been tested and calibrated using test-beams at accelerator complexes such as CERN. In contrast, the ground-based air shower detectors on which we rely for data on the highest-energy cosmic rays do not detect the primary particle, and their detection techniques are not easily calibrated with a test-beam. It is, therefore, quite understandable that results from air-shower arrays show disagreements which can be interpreted in terms of differences in absolute normalisation (see figure 2.4).

As discussed in section 1.4.2, the detection of extensive air showers relies on a number of complementary technologies. There are two fundamentally different approaches: observe the air shower as it develops in the atmosphere, using nitrogen fluorescence or Cherenkov radiation, or sample the shower once it reaches the ground. The largest air-shower detectors, the Pierre Auger Observatory[38]

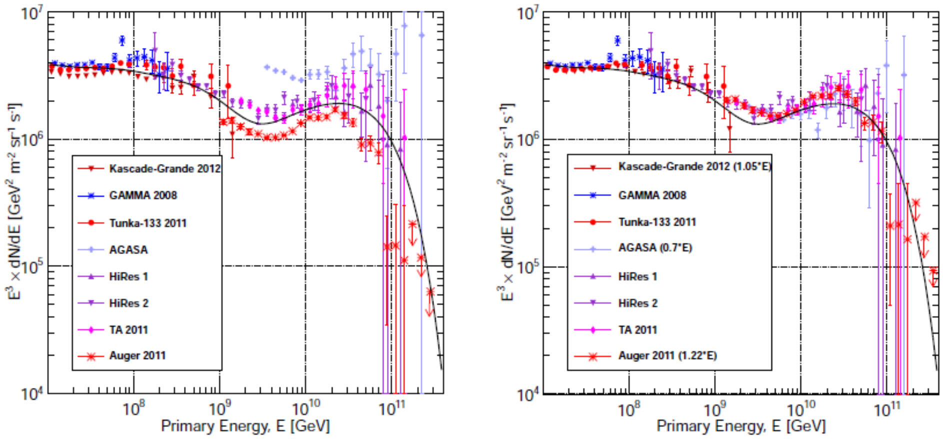


Figure 2.4: The high energy tail of the cosmic-ray spectrum, measured by large-area air shower experiments. Left, data as published; right, data with normalisation of energy scale adjusted as shown. Note that scaling the energy spectrum by E^n, as here ($n = 3$ in this case), exaggerates the disagreement caused by normalisation differences. Figure from Gaisser, Stanev and Tilav[87].

in the southern hemisphere and the Telescope Array[88] in the northern hemisphere, are both "hybrid" detectors which combine a ground array with fluorescence telescopes.

The key observables for Cherenkov and fluorescence detectors are the shower size, i.e. the number of $e^\pm$ produced, N_e (both techniques are sensitive primarily to $e^\pm$), and the location of the shower maximum, usually denoted $X_{\max}$. Essentially, the shower size gives the shower energy, and the depth of shower maximum is sensitive to primary particle mass. The depth is normally expressed in units of g cm^{-2}, column density of air.

Cherenkov radiation is emitted when a charged particle is travelling faster than c/n, where n is the refractive index of the medium. The light travels in a forward cone of angle $\cos^{-1}(1/n\beta)$, as shown in figure 2.5; for extensive air showers $\beta = 1$ to high precision. The refractive index of air depends on temperature, pressure and water vapour content[89]; a value of 1.0002 or so would be reasonable for an altitude of 5 km. This would correspond to a cone of opening angle $\sim 1°$ and a threshold γ-factor of 50, which means a minimum electron energy of 25 MeV (the Cherenkov light will come mainly from secondary $e^\pm$ in the air shower). Cosmic ray air shower experiments using Cherenkov light, e.g. the Tunka array[90] in Siberia, simply consist of an array of upward-facing photomultiplier tubes. This contrasts with the imaging air Cherenkov telescopes (see below and page 23 above) used for high-energy γ-ray astronomy, which use focusing optics to reconstruct the incoming shower direction accurately.

The amount of Cherenkov radiation generated increases as the energy of the air shower increases. The Tunka Cherenkov array found[90] that

$$E = CQ_{175}^{g}, \tag{2.3}$$

where E is the energy of the shower, Q_{175} is the Cherenkov light flux density 175 m from the core of the shower and the index g varies from 0.95 for protons to 0.91 for iron nuclei; as the identity of the incoming particle is not known *a priori*, they used $g = 0.93$ in their energy reconstruction algorithm. The Tunka array studies cosmic rays in the energy range 10^{16}–10^{18} eV and claims an energy resolution of about 15%.

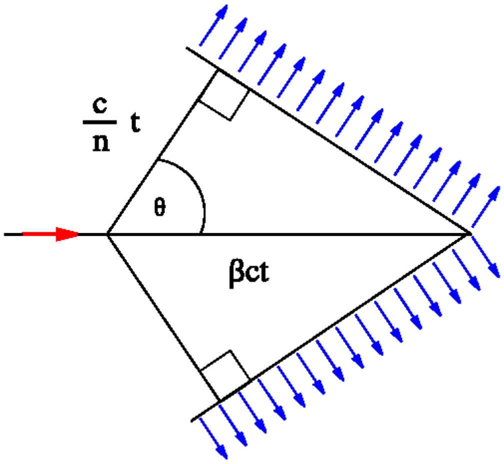


Figure 2.5: Geometry of Cherenkov radiation. The particle travels a distance βct (where $\beta = v/c$) in time t, whereas light travels a distance ct/n in the same time. The result is a coherent wavefront travelling outwards at an angle θ to the trajectory of the particle, where $\cos\theta = 1/n\beta \simeq 1/n$ for highly relativistic particles with $v \simeq c$. Image from Wikimedia Commons.

The disadvantage of Cherenkov radiation in air is the narrow emission cone, which means that the cosmic ray has to be pointing more or less directly at the detector array in order to be seen. Most cosmic-ray experiments which detect the shower in the atmosphere therefore opt to look for nitrogen fluorescence, which is emitted isotropically and hence can be detected at much greater distances from the core of the shower.

Nitrogen fluorescence occurs when $e^{\pm}$ from the air shower excite nitrogen molecules in the atmosphere. The molecules subsequently de-excite, producing a number of discrete emission lines in the near UV (300–400 nm). The total fluorescence yield is proportional to the number of particles in the shower, and therefore to the shower energy; unfortunately, the yield depends on a number of atmospheric parameters such as temperature, pressure and water vapour content, which introduces a substantial systematic error (typically ∼20%) on the energy calibration. The disagreements between different experiments shown in figure 2.4 can be at least partly ascribed to this problem[91].

Nitrogen fluorescence is detected using focusing optics, with a curved mirror directing the light to a focal plane where it is collected by photomultiplier tubes. Because the air shower is not a point source, the focusing does not have to be of particularly high quality: the collecting mirror is often segmented, or spherical rather than parabolic, and the number of PMTs in the focal plane can be quite small. As with Cherenkov light, the aim is to reconstruct the direction of the shower, the total light yield (and hence the shower energy), and the depth of the shower maximum.

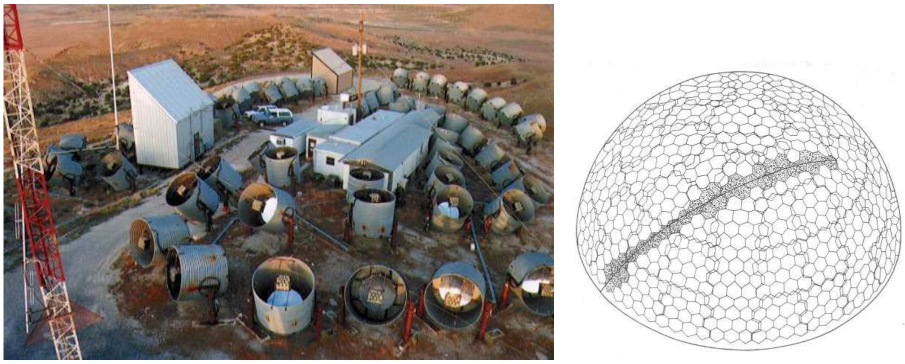


Figure 2.6: The Fly's Eye. Left: a photograph of the experiment; right, an event display. The way that the PMT pixels map on to the sky gave rise to the name of the experiment. Images from [92].

The classic fluorescence detector was the University of Utah's Fly's Eye[92], (see figure 2.6) which operated from 1981 to 1993. The Fly's Eye consisted of

67 modules each with a 1.6 m focusing mirror and a focal-plane "camera" of 12 or 14 photomultiplier tubes. The modules were oriented such that the PMTs together created an 880-pixel map of the sky; in 1986 a second detector, using identical technology but with fewer modules, was constructed 3.4 km from the first, so that the showers could be viewed stereoscopically. This greatly improved the reconstruction of the showers. When the CASA-MIA[93] ground array, comprising the Chicago Air Shower Array of 1089 scintillation detectors and the MIchigan Antiarray of 1024 underground muon detectors, was constructed around the second Fly's Eye in 1992, the modern hybrid experiment concept was born.

The Fly's Eye was designed as a stand-alone detector: its telescopes are oriented outward from a central site. In the Pierre Auger Observatory[38] in Argentina and the Telescope Array[88] in Utah, which were designed as hybrid experiments from the start, the fluorescence telescopes are situated on the edges of the ground array and look inwards, over the array. The layout of the Pierre Auger Observatory and its fluorescence detector stations is shown in figure 2.7.

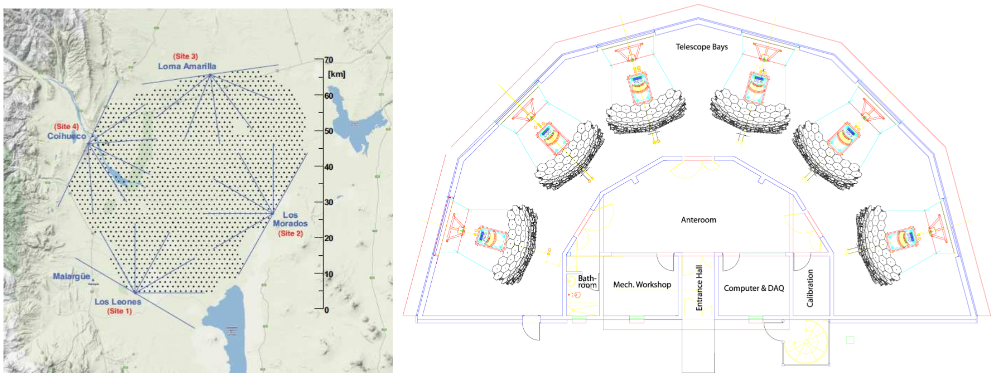


Figure 2.7: Left: layout of the Pierre Auger Observatory[94]. The dots are the stations of the ground array, and the lines show the orientation of the fluorescence telescopes. Each fluorescence detector station contains six individual fluorescence telescopes, as shown in the schematic on the right[95].

Both Cherenkov radiation and fluorescence are sensitive to the electromagnetic component of the shower. For hadronic primaries (protons and heavier nuclei), this is not the total energy: decays of $\pi^{\pm}$ ensure that some of the energy is carried off by muons and neutrinos. This is a comparatively small fraction of the total energy, but it does depend on the identity of the primary, ranging from about 5% for protons up to 15% for iron nuclei (and, of course, zero for γ-rays)[96]. Using a "standard" correction of, say, 10% will therefore produce a small systematic bias, overestimating the energies of protons and γ-rays and underestimating those of heavier nuclei. Alternatively, the correction can be tuned to the $X_{\max}$ value, which is sensitive to the nature of the primary (see below). In hybrid arrays, especially those with muon identification, the non-electromagnetic content of the shower can be estimated directly.

Ground arrays sample the shower when it reaches the ground. Even more than Cherenkov detectors, they have to be physically hit by the shower: clearly a shower that does not hit the array will not produce a signal. The individual components of the ground array need to be robust, relatively simple to construct, and cheap: the Telescope Array has over 500 ground stations, and the Pierre Auger Observatory 1600. There are two suitable technologies: water Cherenkov detectors, consisting of a sealed tank of clean water viewed by one or more photomultiplier tubes (as used by Auger), and plastic scintillator slabs, sealed in a light-tight and weatherproof box and again read out by photomultiplier tubes, either directly or via wavelength-shifting fibre (used by the TA). In both cases, individual stations are powered by solar panels (with battery packs for night-time operation) and equipped with GPS receivers for accurate time-stamping of data. Some ground arrays also have dedicated muon detectors, placed underground as with CASA-MIA[93] or shielded by absorbing material as in KASCADE[97].

The energy of the shower is usually estimated from the shower density at a certain distance from the shower core; the specific distance varies from array to array, and is chosen to minimise the effect of statistical fluctations in shower development and the dependence on the mass of the primary. The Telescope Array uses $S(800)$, the density of shower particles at a distance of 800 m from the shower core[98]; the larger and more sparsely sampled Pierre Auger Observatory uses $S(1000)$ [99].

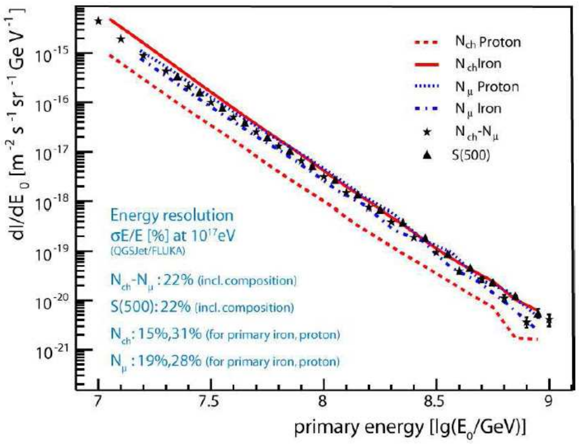


Figure 2.8: Comparison of the cosmic ray energy spectrum derived by KASCADE-Grande using different methods[100]. The energy resolutions are given at bottom left. Only the N_{ch} method assuming incident protons produces a significantly different result.

Other energy estimators can also be used. The smaller KASCADE ground array[97], which had both plastic scintillators and dedicated muon detectors, compared four different methods[100]: N_{ch}, the total number of charged particles (estimated from the signal in the scintillators); N_{μ}, the total number of muons (from the muon detectors); a combination of both N_{ch} and N_{μ} (which should be able to account for the effect of the primary mass); and $S(500)$, the shower density at 500 m from the core. The methods were generally in good agreement, except for N_{ch} calibrated assuming protons, which is significantly lower. At face value, this suggests that the composition of cosmic rays at these energies is rather heavy; however, there is significant dependence on the Monte Carlo simulation used to calibrate the methods.

Hybrid detectors can cross-check energy calibrations between the surface array and the fluorescence telescopes, and can attempt to calibrate the fluorescence telescopes directly with reference light sources and a nitrogen laser, which can be used to excite a known amount of nitrogen fluorescence. However, all of these have some systematic model dependence: the Pierre Auger Observatory is thorough and diligent in its calibration[101], but its results as presented in the left panel of figure 2.4 seem clearly anomalous compared to lower-energy experiments (unlike those of the Telescope Array and its predecessor HiRes[102], which match the lower energy results well). Auger and the TA are addressing the mismatch between their energy calibrations by a series of joint analyses and common calibrations[103].

Particle identification

Identifying the nature of primary cosmic rays is important in several contexts:

- the *elemental* composition of cosmic rays sheds light on nucleosynthesis, especially of the rare elements lithium, beryllium and boron, and on galactic chemical evolution (see also PHY320);

- the *isotopic* composition can provide information on the circumstances under which cosmic rays are originally produced (e.g., the presence or absence of long-lived radioactive isotopes can be used to infer the time delay between supernovae and emission of cosmic rays from supernova remnants);

- the presence of *antimatter* might provide indirect evidence for dark matter (see section 1.6.3), and potentially for the large-scale existence of antimatter elsewhere in the universe (positrons and antiprotons, and perhaps even antideuterons, could be produced locally by interactions of high-energy particles, as occurs in terrestrial accelerators, but the existence of antihelium would strongly imply anti-stars).

In balloon-borne and space-based cosmic ray experiments, the principal means of particle identification is measurement of the charge Z from the amount of ionisation produced in an appropriate detector (both silicon detectors and scintillators are sensitive to the charge of the incoming particle). This technique will separate *elements* (for example, the CREAM balloon-borne experiment separates elements up to nickel ($Z = 28$) with misidentification rates no worse than 2–3%[104]), but not *isotopes*, and will also not distinguish between the various particles of charge ±1 (electrons, positrons, protons, antiprotons).

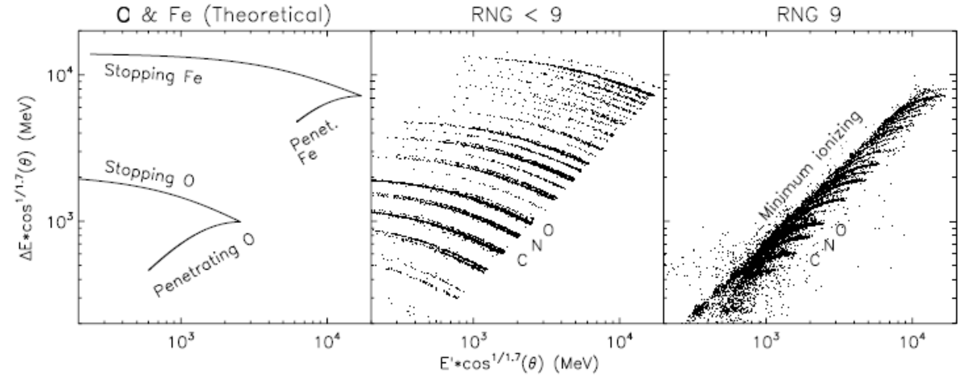


Figure 2.9: Separation of elements and isotopes using the CRIS instrument on ACE. The middle plot shows ions that stop in the stack; the right-hand one shows penetrating particles. On the left, theoretical behaviour for oxygen and iron. Separate isotopic bands can be seen in the middle plot, e.g. for oxygen. The $\cos^{1/1.7}\theta$ factor correects for the angle of incidence of the track. Figure from [106].

The CRIS (Cosmic Ray Isotope Spectrometer) instrument[105] aboard the ACE (Advanced Composition Explorer) spacecraft[83] also uses silicon detectors, but is designed primarily to study lower energy ions which stop in the 9-layer silicon stack. By comparing ΔE, the energy deposited in the first half of the stack, with E', the energy deposited in the second half, stopping ions can be separated to the level of individual isotopes, while penetrating particles can be separated at the level of elements (see figure 2.9[106]).

Charge measurement can be supplemented by various more sophisticated particle identification techniques. Most of these rely on sensitivity to the *speed* of the particle, which can be combined with information on its momentum or energy to determine its mass. Obviously, such methods work best for relatively low-energy particles: as a relativistic particle's energy increases, its speed rapidly becomes very close indeed to c. A proton with an energy of 10^{17} eV has a speed which is just 53 nm s^{-1} less than c—clearly not a detectable difference. However, at energies up to a few GeV, there are several techniques that can be (and have been) used.

1. *Cherenkov radiation* can be used for particle identification in two distinct ways. Most simply, Cherenkov radiation will only be produced if $v > c/n$, where v is the speed of the particle and n is the refractive index of the medium: thus, in air at ground level ($n \simeq 1.0003$), a 25 MeV electron will produce Cherenkov radiation, but a proton with kinetic energy 25 MeV will not. In accelerator experiments, where the energy of incoming particles is known, *threshold Cherenkov* counters can be designed with carefully tuned refractive indices, such that one species of particle—say, pions—will radiate while another species—kaons, perhaps—will not. This technique is not useful in cosmic-ray experiments, because the incoming energy is not well known; however, it can be used to veto non-relativistic particles, as in the CREAM transition radiation detectors discussed above.

   The second method, *ring imaging Cherenkov detectors* or RICH, uses the fact that the half-angle of the Cherenkov cone is given by $\theta = \cos^{-1}(1/n\beta)$ where $\beta = v/c$ (see figure 2.5). Typically, the particle passes through a thin Cherenkov radiator, so that a well-defined cone of light is produced, and this cone is then imaged as a ring on a detector surface behind the radiator. The refractive index of the radiator is accurately known, as is the geometry of the system, so the radius of the reconstructed ring gives θ and hence β.

2. Detectors with precise timing can be used to measure the *time of flight* (TOF) of the particle across a well-defined distance. The speed is calculated simply from distance/time.

3. Particles lose energy as they travel through matter. The mean energy loss is given[107] by the **Bethe formula**

   $$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = \frac{N_A e^4}{8\pi\epsilon_0^2 m_e c^2} \frac{Z}{A} \frac{z^2}{\beta^2} \left[\ln\left(\frac{2m_e c^2 \beta^2 \gamma^2 W_{\max}}{I^2} \right) - 2\beta^2 - \delta(\beta\gamma) \right], \tag{2.4}$$

   where Z and A are the atomic number and mass number of the absorbing material, z and β are the charge (in units of e) and speed (in units of c) of the incident particle, γ is its Lorentz factor, N_A is Avogadro's number, m_e is the electron mass, $W_{\max}$ is the maximum possible energy transfer in a single collision, I is the mean excitation energy, and $\delta(\beta\gamma)$ is a correction factor depending on the density of the medium. The authoritative *Review of Particle Properties*[107] remarks that "few concepts in high-energy physics are as misused as $\langle \mathrm{d}E/\mathrm{d}x \rangle$" (because the mean is skewed by rare events that deposit a lot of energy, so that the *most probable* energy loss is a great deal less than the mean); nevertheless, measures of $\mathrm{d}E/\mathrm{d}x$ (usually truncated means, excluding very low and very high values) are valid and valuable methods of particle identification.

The PAMELA satellite uses both dE/dx and TOF for particle identification. Figure 2.10 shows the dE/dx measurement as a function of rigidity (cp/Ze). Although this measurement was only used to separate $Z = 1$ from $Z = 2$ in [108], the bands for the different isotopes can be clearly seen. The TOF results, which were used to separate the isotopes, are shown in figure 2.11.

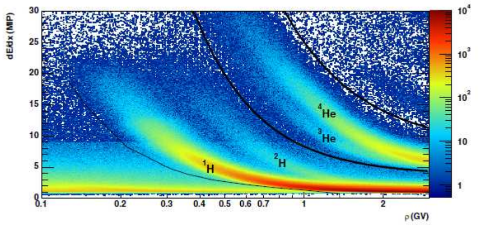


Figure 2.10: dE/dx measured by the PAMELA silicon tracking system [108], plotted against rigidity (denoted by ρ rather than R). Note the separation of deuterium from hydrogen, and of the two helium isotopes.

AMS-02 has a RICH system in addition to dE/dx and TOF, but results presented to date do not seem to have used it. It is intended to provide isotope separation up to $A \sim 15-20$ and $1 < p/A < 12$ GeV/c[109]. The CAPRICE balloon-borne experiment[110], which flew in 1993, was the first high-altitude cosmic-ray experiment to incorporate a ring-imaging Cherenkov detector in addition to TOF and dE/dx (from scintillator); it successfully separated deuterons from protons in the rigidity range 1–5 GV[111].

As with energy measurement, particle identification in air-shower experiments is necessarily much more indirect than it is for balloon-borne and space-based experiments. In Cherenkov and fluorescence detectors, the key observable is the depth of shower maximum, $X_{\max}$, and its variance: heavier nuclei shower earlier (they have smaller $X_{\max}$) and have smaller variance than protons. The data are compared to simulations, which can be tuned using LHC data; different models do disagree, at the level of a few percent, but the difference between protons and iron nuclei is about a factor of 5 bigger than this systematic.

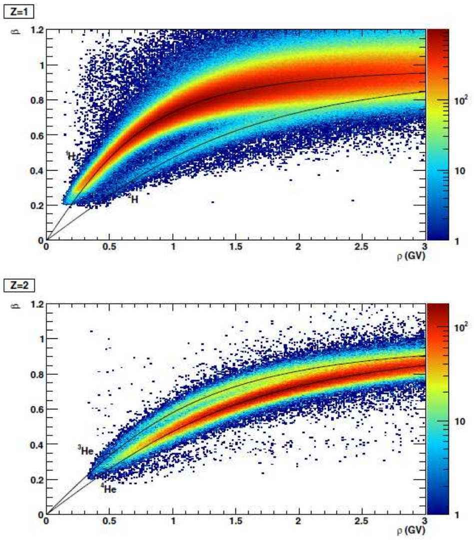


Figure 2.11: PAMELA time-of-flight measurements[108].

Ground arrays with muon detection, such as KASCADE[97] and the IceTop air shower array[112] at the South Pole (which operates in conjunction with IceCube[44]) can assess composition based on the number of muons relative to the overall shower size (heavier nuclei produce more pions, and therefore more muons, for a given shower size).

As with the energy measurements, there are disagreements between experiments, which are not currently understood but are presumably the result of systematic differences in calibration. Of particular note is a significant disagreement between Auger on the one hand, and HiRes and the Telescope Array on the other, as to the composition of the very highest energy cosmic rays, with HiRes and the TA both concluding that the composition is proton-like up to the highest energies, while Auger sees a move toward heavier primaries. This issue is discussed further in the next section.

Direction of incoming particle

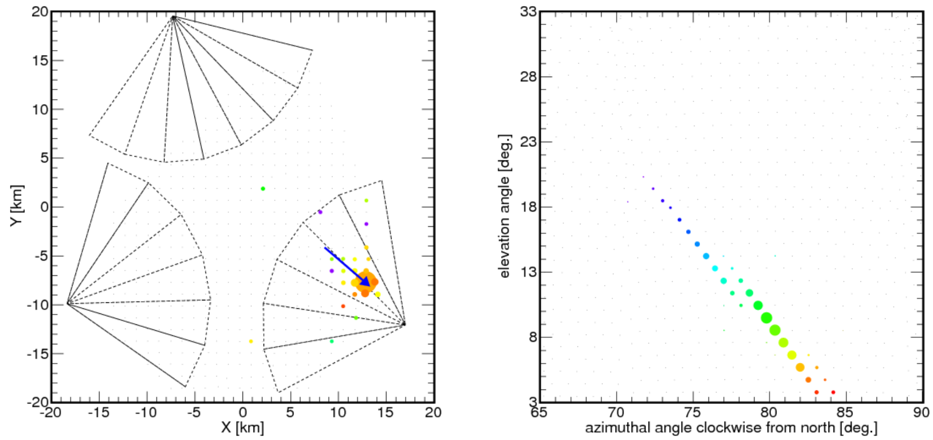


Figure 2.12: A hybrid event display from the Telescope Array[113]. Left, map of the array showing the hit surface detectors. The colour of the circle (from purple to red) indicates relative timing, and the size of the circle indicates the number of photoelectrons. The arrow shows the reconstructed direction of the shower. Right, data from the fluorescence telescopes; each dot represents a hit PMT, and the colour and size are as before.

Determining the direction of the incoming particle is straightforward in experiments that see the primary particle directly: anything that reconstructs a track or shower axis (magnetic spectrometers, transition radiation detectors, tracking calorimeters) will provide directional information. The only case that requires more discussion is that of ground arrays.

The critical information required to reconstruct direction in ground arrays is the *relative time* of each hit ground station or PMT. For fluorescence detectors, two coordinates (θ and ϕ, or elevation and azimuth) are provided by the orientation of the hit PMT; the arrival time gives the third dimension. For ground stations, the shower front is a slightly convex surface perpendicular to the shower direction: the lower edge of the front will hit first, producing an elliptical pattern of hit stations with a time gradient along the long axis. Figure 2.12 shows a hybrid event display from the Telescope Array[113], in which these features can be seen.

2.2.3 Observed properties

As noted on page 46, the interesting properties of cosmic rays are their energy spectrum, their chemical composition, and any variation with time or position.

As shown in figure 2.13, an updated version of the famous "Swordy plot"[114], the cosmic ray energy spectrum is essentially a negative power law with an overall spectral index of about –2.8. It has three main features: a levelling off at energies below a few GeV, a change of slope—the "knee"—at about 5×10^{15} eV, and a rather less well-defined change of slope—the "ankle"—at about 5×10^{18} eV. The last data points are at a few times 10^{20} eV: we need to understand whether this is an artefact of statistics, or a genuine cutoff.

The low-energy behaviour is reasonably well understood: this is not really an inherent feature of cosmic rays, but a consequence of their interaction with the solar magnetic field. As a result, the spectrum below ~5 GeV varies with solar activity: disagreements between experiments in this energy range are not to be taken seriously unless the experiments in question are known to have been taking data at the same time. From the viewpoint of high-energy particle astrophysics, the more interesting features are the knee, the ankle, and the high-energy cutoff.

The existence of the knee is well established by a large number of independent experiments, although its exact location is uncertain because different experiments have somewhat different energy calibrations (see, e.g., [87], figure 2) and there is no good criterion for determining which is “best”. An important feature of the data is that the knee is associated with a clear shift towards heavier composition, as shown in figure 2.14. As noted on page 46, this is exactly the behaviour that would be expected if the slope break at the knee is caused by a change in source population, with the lower energy source cutting off for different nuclei at energies $\propto Z$ because of the effects of magnetic fields.

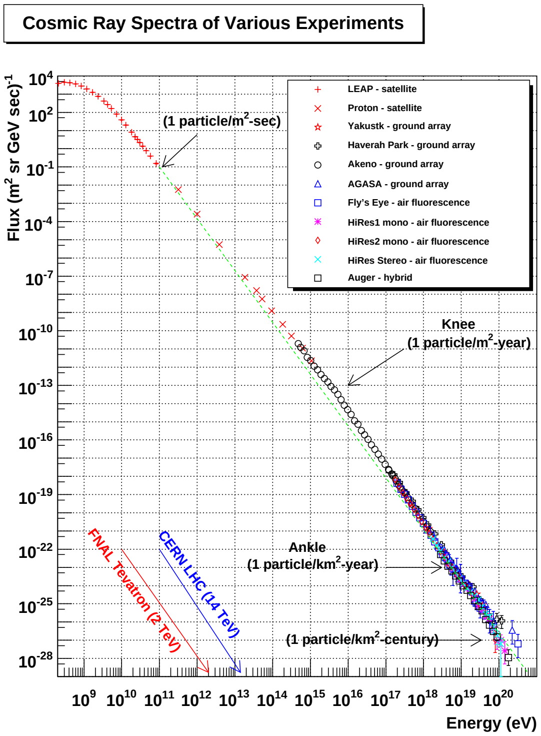


Figure 2.13: The energy spectrum of cosmic rays, originally produced by Simon Swordy[114]; this version updated by W Hanlon[115].

The gyroradius of a particle in a magnetic field is given by

$$r_g = \frac{p}{zeB} = \frac{R}{Bc}, \tag{2.5}$$

where ze is the particle’s charge, R is its rigidity, and B is the applied magnetic field. The magnetic field of the Milky Way is complicated, but its strength is of the order of 0.1 nT. This gives a gyroradius of the order of 5 pc for the knee—protons of this energy are magnetically confined in the Milky Way, and the dominant sources both below and above the knee must be Galactic. However, the gyroradius for the ankle is around 5 kpc. This is close enough to the size of the Milky Way that the ankle is generally believed to indicate a shift from predominantly Galactic to predominantly extragalactic sources.

When the overall power law is taken out by multiplying the flux by E^3, as in figure 2.4, the resulting spectrum at high energies has three distinct features: a change of slope at just over 10^{17} eV, a pronounced dip at 5×10^{18} eV (this is what is seen in the unscaled plot as the ankle), and a sharp cutoff at energies above about 2×10^{19} eV.

An attractive explanation for the high-energy cutoff is the so-called *GZK limit*[116]. A proton of sufficiently high energy can interact with a photon of the cosmic microwave background to produce the $\Delta(1232)$ resonance, which will then decay via the strong interaction to a nucleon and a pion:

$$p + \gamma \rightarrow \Delta \rightarrow p + \pi^0 (n + \pi^+). \tag{2.6}$$

For the production of the Δ, we have, in the most favourable case where the proton and the photon collide head-on,

$$E_\Delta = E_p + E_\gamma;$$
$$p_\Delta = p_p - E_\gamma$$

(taking $c = 1$ as usual in particle physics, so that $p_\gamma = E_\gamma$). If we square these equations and subtract them, we get

$$M^2 = m^2 + 2E_\gamma(E_p + p_p),$$

where M is the mass of the Δ and m is the mass of the proton. The proton is highly relativistic, $m_p \ll E_p$, so $p_p \simeq E_p$ and we have

$$E_p = \frac{M^2 - m^2}{4E_\gamma}. \tag{2.7}$$

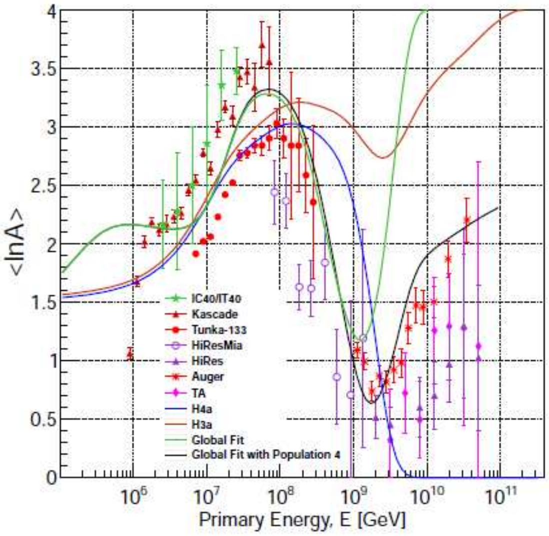


Figure 2.14: The composition of primary cosmic rays as a function of energy[87]. Note the strong peak just above the "knee" energy, and the suggestion, driven mostly by Auger data, of another peak at the high energy endpoint.

The average energy of a blackbody photon is $\overline{E} = 2.7k_BT$. If we put this into equation (2.7), we get a threshold energy for this reaction of 2.5×10^{20} eV. In fact, the threshold energy will be lower than this, because microwave background photons are extremely numerous so the protons will be able to interact with the high-energy tail of the blackbody distribution. The observed cutoff of $\sim 5 \times 10^{19}$ eV corresponds to photons of about five times the mean energy, which seems entirely reasonable.

The kinematics of the Δ decay give (assuming the $p\pi^0$ mode, and taking the case where the proton and pion momenta are parallel or antiparallel to the original proton momentum)

$$M^2 = m^2 + m_\pi^2 + 2E'_pE_\pi + 2p'_pp_\pi,$$

where E'_p and p'_p are the energy and momentum of the "new" proton. Assuming that the proton and the pion are relativistic, but not neglecting their masses altogether, we can write

$$p = \sqrt{E^2 - m^2} \simeq E\left(1 - \frac{m^2}{2E^2}\right)$$

which gives

$$M^2 - m^2 - m_\pi^2 = m^2\frac{E_\pi}{E'_p} + m_\pi^2\frac{E'_p}{E_\pi}.$$

If we write $Q = M^2 - m^2 - m_\pi^2$ and $x = E_\pi/E'_p$, we can express this as a quadratic,

$$m^2x^2 - Qx + m_\pi^2 = 0.$$

The two solutions to this are the minimum and maximum pion energies. Taking the minimum gives $x \simeq 0.03$: about 3% of the original proton energy is transferred to the pion. If the original proton energy was greater than the threshold, this process will repeat until the energy of the produced proton is too low.

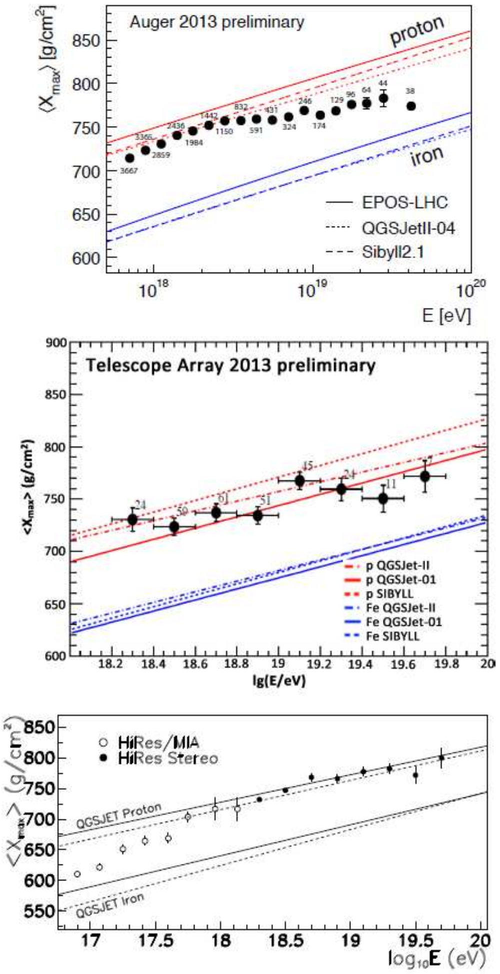


Figure 2.15: The mean depth of shower maximum in the highest-energy region[118], as measured by Auger (top) and the Telescope Array (middle). Note that the two plots cannot be directly superimposed, because the event selections are different for the two experiments and so are the models shown. Bottom, older data from HiRes[119].

To see how this is likely to affect the cosmic ray spectrum, we need to calculate the mean free path of a cosmic-ray proton,

$$\ell = \frac{1}{\sigma_{\gamma p} n_\gamma}, \qquad (2.8)$$

where $\sigma_{\gamma p}$ is the photon-proton cross-section and n_γ is the number density of photons.

The total cross-section for pion photoproduction at the $\Delta(1232)$ resonance is about 300 μb $= 3 \times 10^{-32}$ m^2[117], and the number density of CMB photons is

$$\begin{aligned} n = \frac{\mathcal{E}_{BB}}{\overline{E}} &= \frac{4\sigma T^4}{c}\,\frac{1}{2.7k_BT} \\ &= 4 \times 10^8\,\mathrm{m}^{-3}, \end{aligned}$$

where $\mathcal{E}_{BB}$ is the energy density of blackbody radiation and $\overline{E}$ is the mean photon energy. This gives a mean free path of about 3 Mpc. For protons close to the threshold energy, the mean free path will be substantially more than this, because they can only interact with the high-energy tail of the distribution, but it will still only be of the order of tens of Mpc—a ballpark figure of 50–100 Mpc is usually quoted. This is *quite small* by cosmological standards: the Coma cluster of galaxies, which is the *nearest* rich regular cluster, is about 100 Mpc away. It follows that *cosmic-ray protons with energies significantly in excess of* 5×10^{19} *eV must be produced in the local universe.*

The high-energy cutoff shown in figure 2.4 seems consistent with expectations from GZK. However, the increase in mean particle mass suggested by the data in figure 2.14 could be interpreted as indicating that the cause of the cutoff is not, in fact, the GZK limit, but simply that the source has reached its maximum energy. Therefore, it is important to have a clear picture of the composition of cosmic rays at the highest observed energies.

Unfortunately, at present the picture at these energies is far from clear: there is a long-standing disagreement between the Pierre Auger Observatory and the Utah experiments (Telescope Array and its predecessor HiRes), as shown in figure 2.15. The data from the TA and HiRes are consistent with a light composition to the highest energies, whereas the Auger data clearly indicate a shift to heavier composition. It should be noted that the TA statistics are substantially lower than Auger's (it's a smaller array, and has been operating for a much shorter time), so that an admixture of heavier nuclei as favoured by Auger is not *very* strongly disfavoured by the TA; the HiRes data are more difficult to compare, because the simulation is an older version (note that the QGSJet simulations in the upper plots do move the "proton" expectation closer to "iron" than other models).

HiRes and the TA are closely related experiments involving many of the same people, so they cannot be said to "outvote" Auger: if there were a systematic error in the HiRes analysis, it would probably have been inherited by the TA. It is fair to conclude, as do Gaisser, Stanev and Tilav[87], that "there is not at present a satisfactory understanding of the highest energy cosmic rays." The fact that the Auger and TA Collaborations are working together to resolve the discrepancies, as shown by joint contributions to conferences[103], is therefore a very welcome development.

At lower energies, the elemental and isotopic composition can be determined in more detail, using the particle identification techniques discussed above. Compared to the solar system, the elemental abundances in lower-energy (definitely Galactic) cosmic rays show clear systematic differences (see figure 2.16). Hydrogen and helium are under-represented by about a factor of 10 compared to silicon, while the light elements lithium, beryllium and boron, the elements with odd Z and the elements just below iron ($20 < Z < 26$) are all over-represented—the light elements by a factor of up to 10^6. As discussed in PHY320, this is caused by *spallation*, where interactions break off small pieces of heavier nuclei. This mechanism is believed to be the source of almost all the cosmic abundance of Li, Be and B (only ^{7}Li is made by another process, in the early universe), as these nuclei readily convert to ^{4}He in stellar interiors.

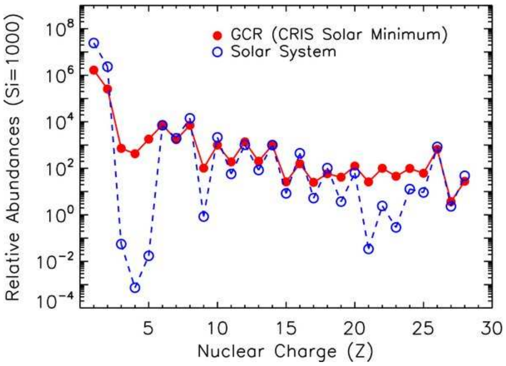


Figure 2.16: The abundance of the light elements (up to $Z = 28$, nickel), in Galactic cosmic rays (red) and the solar system (blue), normalised (as is standard in such plots) to silicon. The CR data are from CRIS[120], except for the points at $Z \leq 4$ which are from BESS[121]. The solar system data are from Lodders[122].

Understanding the composition of cosmic rays and its dependence on variables such as energy per nucleon or rigidity is a complicated problem (see [124] section 2): the number density of a particular particle species as a function of momentum will depend on its production rate (as a primary particle, a spallation product, a product of radioactive decay, or some combination of these), its diffusion rate through the Galaxy (which depends on the Galactic magnetic field and on Galactic winds), the probability that its momentum is changed by scattering off turbulent magnetic fields (as described in the next chapter), and the rate at which this species is destroyed by spallation or radioactive decay. These in turn depend on interaction cross-sections, the density and composition of the interstellar medium, and the structure of the Galaxy's magnetic field. The general form of the propagation equation for a given species is[124]

$$\begin{aligned}\frac{\partial\psi(\mathbf{r},p,t)}{\partial t} &= q(\mathbf{r},p,t) + \nabla\cdot(D_{xx}\nabla\psi - \mathbf{V}\psi) + \frac{\partial}{\partial p}p^2 D_{pp}\frac{\partial}{\partial p}\frac{1}{p^2}\psi \\ &\quad - \frac{\partial}{\partial p}\left[p\psi - \frac{p}{3}(\nabla\cdot\mathbf{V})\psi\right] - \frac{1}{\tau_f}\psi - \frac{1}{\tau_r}\psi,\end{aligned} \tag{2.9}$$

where ψ is the number density per unit of total momentum p at position $\mathbf{r}$ and time t, q is the rate of production from all sources (primary, spallation and decay), D_{xx} is the spatial diffusion coefficient (which depends on $\mathbf{r}$, the rigidity R, and the velocity $\mathbf{v}$, and is determined by the gas dynamics and magnetic field of the Galaxy), $\mathbf{V}$ is the convection velocity (relating to Galactic winds), D_{pp} is the coefficient of diffusion in momentum space, and describes changes in particle momentum caused by scattering off magnetic turbulence etc., τ_f is the timescale for destruction of this species by fragmentation, and τ_r is the timescale for destruction by radioactive decay.

This is clearly not a very tractable expression from an analytical perspective, but there are publicly available computer codes such as `GALPROP`[125] which solve it numerically. There are also simplified models, particularly the "leaky box"[124], which is widely used to produce theoretical distributions for comparison with observational data. In the leaky box model, the sources are distributed uniformly throughout the box, and the terms in equation (2.9) which result in loss of particles are subsumed into a "leakage" from the box with characteristic escape time τ_{esc}.

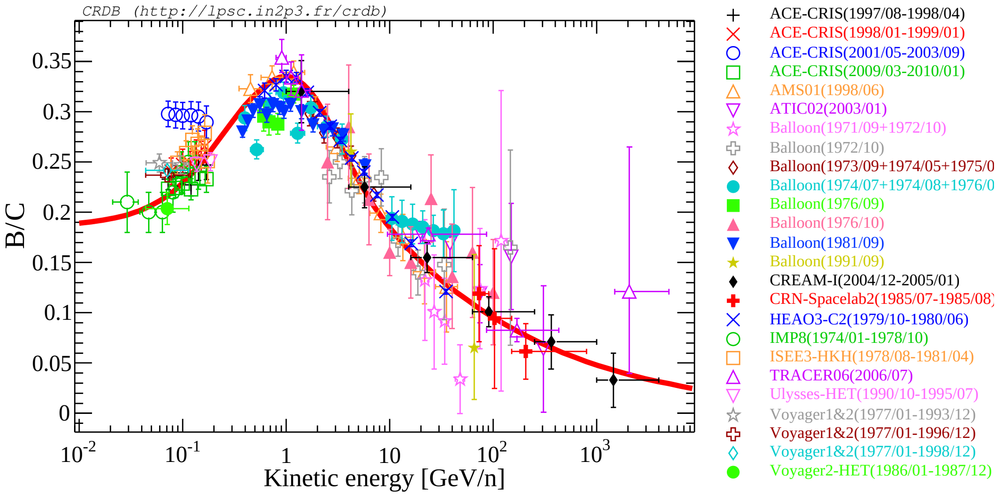


Figure 2.17: Boron to carbon abundance ratio as a function of kinetic energy per nucleon, from a range of balloon-borne and space-based experiments. Note the effect of solar activity: the open blue circles (CRIS 2001–2003) represent solar maximum, while the other low energy data are at solar minimum. Data from the Cosmic Ray Database[123]. Solid line: expectation from `GALPROP`[125] with solar modulation set to 300 MV.

One test of propagation models is a comparison between *secondary* elements or isotopes (produced by spallation) and *primary* elements or isotopes (produced in the source). The usual benchmark is the boron to carbon ratio (see figure 2.17), because boron is almost purely secondary (see above and figure

2.16) while carbon is primary, and both of them are light enough to be identified reliably by many experiments. These data agree well with each other—the apparent outlier at low energies is caused by a difference in solar activity—and are well described by leaky box models or by `GALPROP` as shown.

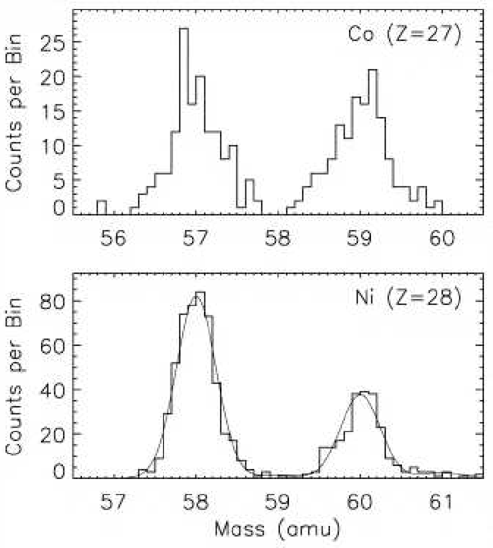


Figure 2.18: The isotopic abundance of Co (top) and Ni (bottom), as measured by CRIS[126]. Note the absence of atomic mass 59 for nickel, despite the fact that models indicate that a large fraction of the atomic mass 59 material synthesised in supernovae is nickel-59.

Radioactive nuclides can act as "clocks" to investigate the timescales on which things happen. Of particular interest in this respect are a few isotopes (A, Z) for which $M_n(A, Z-1) < M_n(A, Z) < M_n(A, Z-1) + m_e$, where M_n is the *nuclear* mass, $M_n = M_A - Zm_e$ where M_A is the tabulated atomic mass. Such isotopes can decay by electron capture, ${}^{A}_{Z}\mathrm{X} + e^- \rightarrow {}^{A}_{Z-1}\mathrm{X}' + \nu_e$, but *not* by β^+ decay, ${}^{A}_{Z}\mathrm{X} \rightarrow {}^{A}_{Z-1}\mathrm{X}' + e^+ + \nu_e$. Therefore, *if they have been stripped of all electrons, they are completely stable*—they can only decay if at least the innermost electron shell is present (in principle, they could decay by capturing a free electron, but the relative velocities of the electron and the nucleus would have to be extremely low in order to achieve capture rather than scattering, so this is usually a small effect). The most interesting primary isotopes of this kind are ^{59}Ni (half-life 76000 years) and ^{57}Co (0.74 years), both of which are iron-peak elements and therefore expected to be synthesised in quantity in supernovae. The absence of ^{59}Ni in cosmic rays[126] (see figure 2.18) indicates that the nickel has time to cool, attach electrons and—in the case of ^{59}Ni—decay by electron capture to ^{59}Co before being accelerated and ejected as cosmic rays. This is particularly interesting as young supernova remnants, e.g. Tycho's, are a favoured candidate site for acceleration of Galactic cosmic rays; as these SNRs are young compared to the half-life of ^{59}Ni, it suggests that they must be accelerating pre-existing interstellar material rather than their own ejecta.

We would expect secondary isotopes that are only unstable against electron capture to be present in cosmic rays: if they are produced by spallation, they have been at high energies for their entire existence. However, isotopic separation at these comparatively high masses is only possible for very low-energy cosmic rays, where the possibility of electron reattachment in flight is not negligible. Figure 2.19[127] shows that the unstable isotopes are indeed present (note that the electron-capture half-lives of ^{49}V and ^{51}Cr are only 330 days and 27.7 days respectively, so the electron-capture decay must be very much suppressed for these isotopes to be detectable at all), but the abundances of their daughter nuclides are enhanced at lower energies, as expected if some electron reattachment and subsequent electron-capture decay is taking place. The degree of enhancement is in qualitative agreement with the leaky-box predictions shown, but there is some detailed disagreement: the abundance of ^{49}Ti is systematically low, and of ^{51}V systematically high, compared to the expectation.

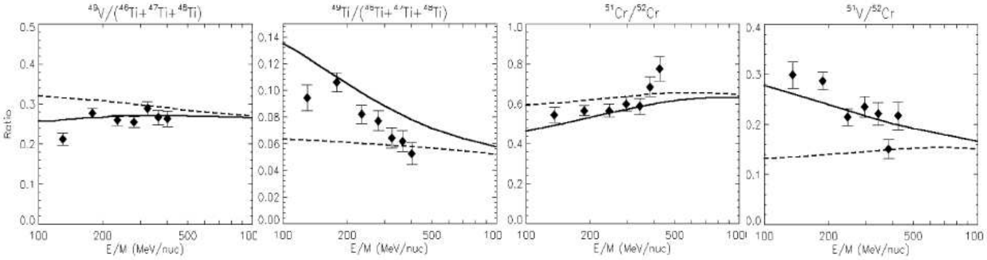


Figure 2.19: The abundance of the electron-capture isotopes ^{49}V (left) and ^{51}Cr (right), and their stable daughters ^{49}Ti (left) and ^{51}V (right), as measured by CRIS[127]. All abundances are presented as ratios with nearby stable isotopes. The dashed lines are the expectations from a leaky-box model without electron reattachment, and the solid lines from a model with reattachment. Measurements taken at solar minimum. Figure from [127].

Because the disagreement is not consistent, it is not well understood: systematically high values would point to significant reacceleration of secondary nuclides (i.e. some of the nuclides in question have previously spent some time at lower energies where electron reattachment is more likely), while systematically low values might be due to a miscalculation of the effect of solar activity (so-called *solar modulation* tends to cause incoming cosmics to lose energy, so the original energies would be higher, and hence the probability of electron reattachment lower), but one would expect either of these effects to occur for *all* species, not just some of them.

Individual antiparticles (positrons and antiprotons) can be produced as secondaries in high-energy interactions: any high-energy photon can convert into an e^+e^- pair in the vicinity of another charged particle, and π^0s "Dalitz decay" into γe^+e^- about 1% of the time. Antiprotons require higher-energy interactions as a consequence of their greater mass: the minimum proton energy required to create a $\bar{p}p$ pair in a collision with a stationary nucleus is 6.2 GeV. Antideuterons could also be produced as secondaries—they have been observed in terrestrial ep[128] and pp collisions—but have not so far been observed in cosmic ray experiments (an upper limit of 1.9×10^{-4} (m^2 s sr GeV/n)$^{-1}$ (95% CL) for the energy range 0.17–1.15 GeV/nucleon has been reported by BESS[129]). Although it has been claimed[130] that "discovery of a single anti-helium nucleus in the cosmic ray flux would definitely point toward the existence of stars and even of entire galaxies made of anti-matter", antihelium nuclei are produced in Pb–Pb collisions at the LHC[131] (see figure 2.20), so a very low level of antinuclei could perhaps be consistent with secondary production. In any case, no such detection has been made to date.

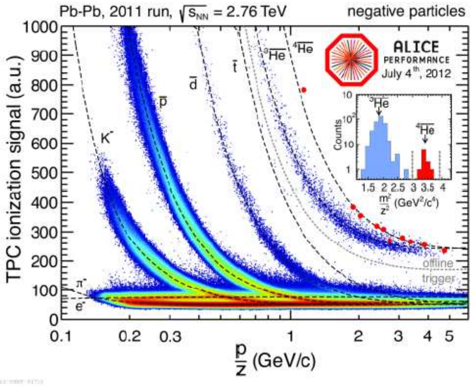


Figure 2.20: Antinuclei identified by dE/dx in heavy ion collisions at LHC[131]. Large numbers of antideuterons and $\overline{^3\text{He}}$ are seen, along with a few probable antitritons and ten identified $\overline{^4\text{He}}$ (red dots).

Antiprotons and positrons, in contrast, have been detected by many ex-

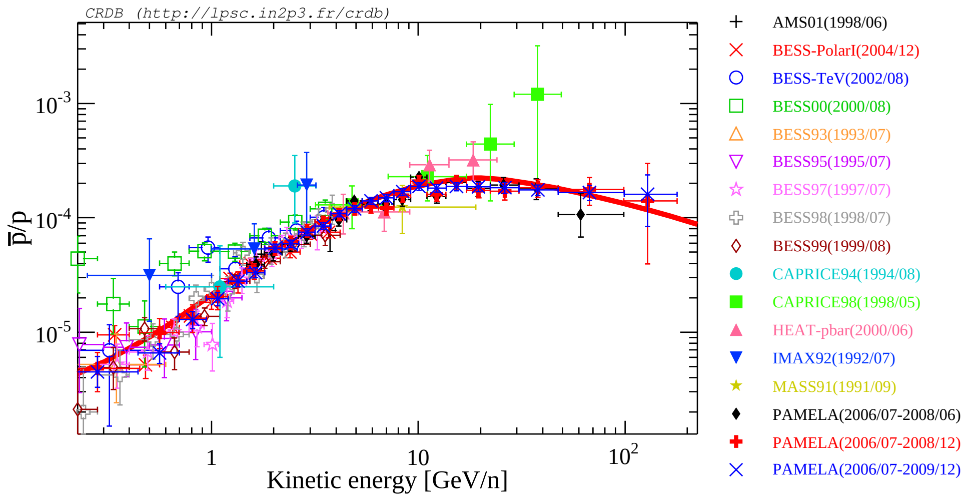


Figure 2.21: Ratio of antiprotons to protons as a function of kinetic energy, from a range of balloon-borne and space-based experiments. Data from the Cosmic Ray Database[123]. Solid line: expectation from GALPROP[125] with solar modulation set to 500 MV and model parameters corresponding to model DC of Moskalenko et al.[132]

periments. The antiproton flux can be well described by models of secondary production, as shown in figure 2.21; there appears to be no need to assume production of antiprotons by the sources.

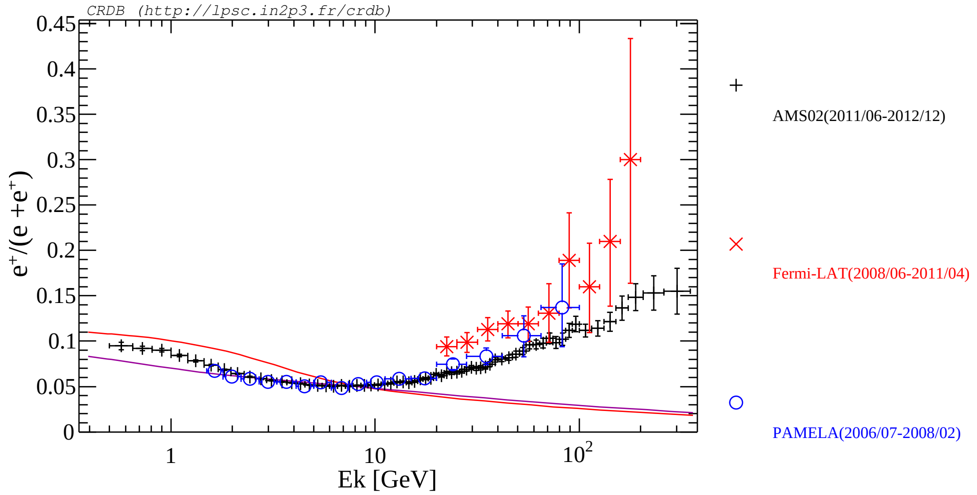


Figure 2.22: Ratio of e^+ to $(e^+ + e^-)$ as a function of kinetic energy, from AMS-02, *Fermi*-LAT and PAMELA. Data from the Cosmic Ray Database[123]. The solid lines are the expectations from GALPROP[125] with solar modulation set to 500 MV and default parameters (red) or parameters from model DC of Moskalenko et al.[132] (magenta).

The situation for positrons is somewhat different, as shown in figure 2.22: the rising positron fraction at high energies is not expected from simple secondary production. Various exotic explanations for this “positron excess” have been proposed[133], but it appears that conventional astrophysical sources, particularly supernova remnants and pulsar wind nebulae (supernova remnants powered by young pulsars, like the Crab Nebula), can account for the spectrum without appealing to new physics[134].

Given that the origin of cosmic rays is still unclear, variations of their intensity with time or direction are potentially interesting. The directions of the highest-energy cosmic rays may not have been completely scrambled by Galactic magnetic fields, and therefore may be correlated with their (presumably extragalactic) sources; even lower-energy cosmic rays may preserve some information if some of their sources are very local (for example, one of the models of the positron flux discussed by Di Mauro et al.[134] ascribes most of it to a single source, the γ-ray pulsar Geminga[135], which is only about 250 pc away). For such local sources, time variation might also be informative: at low energies, the propagation time of charged cosmic rays is much larger than for light (because of magnetic field deflections), but if any component of the higher-energy cosmic ray flux were dominated by one or a small number of nearby *variable* source(s), one might expect to see correlated variations in the received flux.

In fact, although time variation is clearly observed in low-energy cosmic rays, its source is well understood: it is highly correlated with solar activity, and is a consequence of the effect of the Sun's magnetic field on the local cosmic ray flux (solar modulation). An example is shown in figure 2.23[136], where the proton flux measured by PAMELA is shown for the years leading up to the solar minimum in 2009. As a consequence of this effect, data from different experiments are not directly comparable at energies below about 5 GeV, unless the experiments are known to have been taking data at the same time, or at least at the same phase of the solar cycle (so, 2009 would be comparable with 1998 but not with 2003).

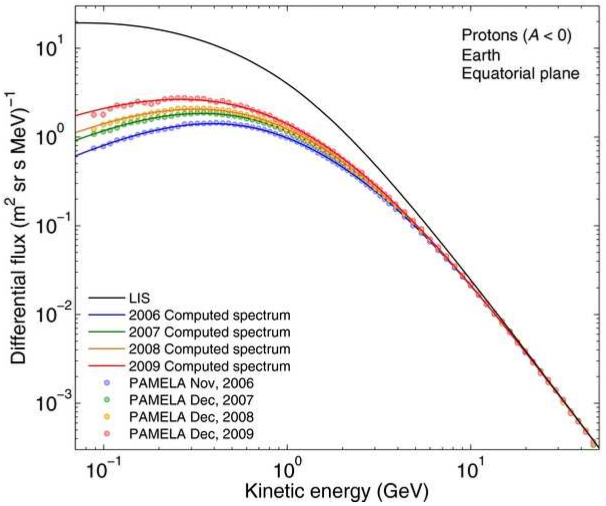


Figure 2.23: Low-energy cosmic-ray proton flux as a function of time, measured by PAMELA[136]. The black line shows the modelled "local interstellar spectrum" in the absence of solar modulation, and the coloured lines show different levels of solar modulation according to solar activity.

As noted above, the cosmic ray flux at the Earth is nearly isotropic, as a result of the effects of the Galactic magnetic field. This isotropy is modified at low energies by the effects of local magnetic fields: we saw above that the variation in low-energy cosmic-ray flux with geomagnetic latitude, and the east-west asymmetry, were instrumental in identifying primary cosmic rays as predominantly positively charged particles.

Measurement of anisotropies at higher energies is complicated by experiment acceptance: any experiment not located at one of the poles will have different effective live time for different parts of the sky, and any seasonal variation in efficiency of data taking (e.g. effect of weather on Cherenkov and fluorescence detectors) can also introduce systematic biases. There are various data-driven methods of constructing "reference maps" which preserve these biases while removing any real structure; see, e.g., section 3.1 of [139].

The gyroradius of a 10^6 GeV proton in a 1 μG (0.1 nT) magnetic field is about 1 pc, so we should expect that the directions of cosmic rays below about 10^7 GeV or so are thoroughly scrambled by the Galactic magnetic field even if the sources are quite local. Nevertheless, significant anisotropies at about the 0.1% level are observed at both large and medium angular scales.

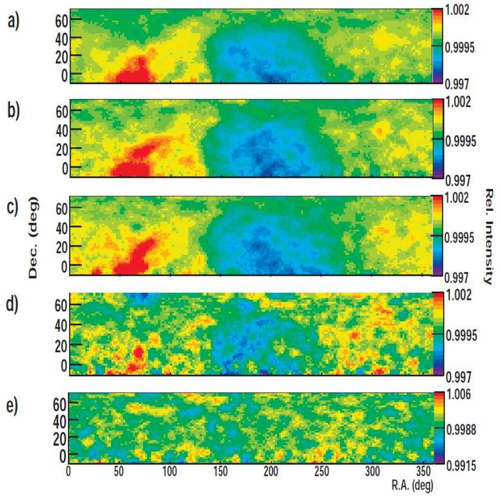


Figure 2.24: Large-scale anisotropy in arrival directions of cosmic rays, as measured by the Tibet Air Shower Arrays[137, 138]. From top to bottom, the panels correspond to CR energies of (a) 4 TeV, (b) 6.2 TeV, (c) 12 TeV, (d) 50 TeV and (e) 300 TeV; it can be seen that the large-scale anisotropy vanishes almost completely at energies above 100 TeV or so. Figure from [138].

Owing to the Earth's orbital velocity of about 30 km s^{-1}, ground-based cosmic-ray detectors should see a seasonal variation in intensity given by[139]

$$\frac{\Delta I}{\langle I \rangle} = (\alpha + 2)\frac{v_{\oplus}}{c}\cos\theta, \qquad (2.10)$$

where I is the intensity of cosmic rays, $\alpha = 2.7$ is the power law index of the cosmic ray energy spectrum (see figure 2.13), $v_{\oplus}$ is the Earth's orbital velocity and θ is the angle between the direction of the incoming cosmic ray and the velocity of the Earth. This corresponds to an anisotropy at the level of 4.7×10^{-4}, but only in a coordinate system in which the direction of the Sun is fixed—in sidereal coordinate systems, whether equatorial or Galactic, it should average out over the course of a year. This "solar dipole" is observed by IceCube[139] at a level consistent with expectation.

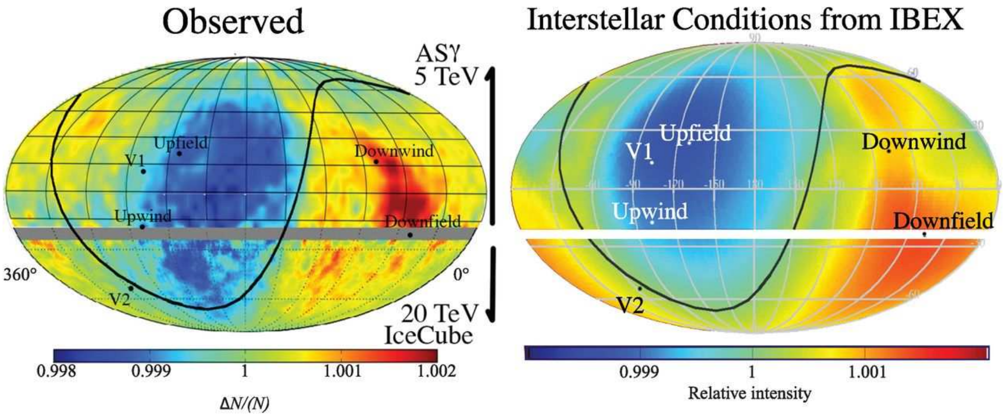


Figure 2.25: A comparison of the measured large-scale anisotropy from Tibet–ASγ[138] and IceCube[139] with a model based on measurements by IBEX[140].

The large-scale sidereal anisotropy was first measured with precision by the Tibet–ASγ ground array[137] (see figure 2.24). This is *not* a pure dipole anisotropy (such as is seen in the CMB as a result of the motion of the solar system relative to the CMB rest frame): the excess and the deficit clearly are not 180° apart. Furthermore, the effect is energy-dependent: it is much weaker in panel (d), corresponding to typical energies of 50 TeV, and has essentially disappeared in panel (e), 300 TeV. This is what we would expect if the anisotropy were driven by magnetic fields, and indeed a study[140] by the IBEX Collaboration[141] shows that the structure is consistent with predictions from measurements of the interaction between the heliosphere and the local interstellar magnetic field (see figure 2.25).

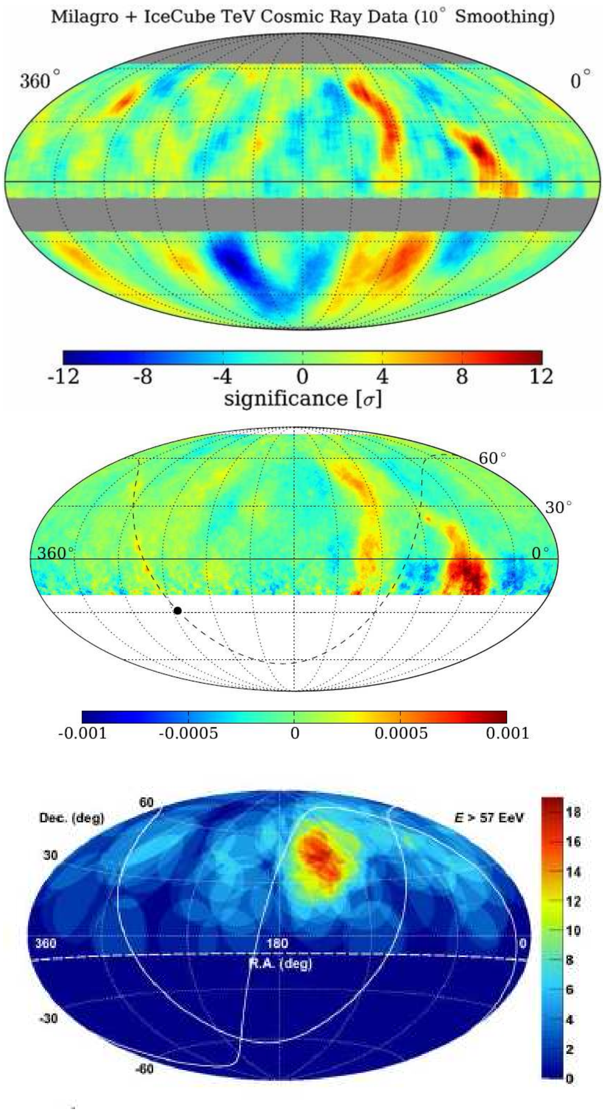


Figure 2.26: Examples of medium-scale anisotropy in cosmic-ray arrival directions. Top, MILAGRO[142] (northern hemisphere) and IceCube[139] (southern hemisphere); the compilation plot is from [139]. Middle, northern hemisphere from ARGO-YBJ[143]. Dipole and quadrupole terms have been removed from these maps to highlight the structures at smaller angular scales. Bottom, northern hemisphere from the Telescope Array[144], at much higher energies.

As well as this large-scale anisotropy, which is well explained by the effects of the local environment, experiments also report medium-scale anisotropies on angular scales of 10 or 20°. Some examples are shown in figure 2.26. The upper two plots are for cosmic ray energies of order 10 TeV; the lower plot, from Telescope Array[144], is for ultra-high-energy cosmic rays with energies above 57 EeV (5.7×10^{19} eV).

The structures in the upper plots are clearly real, and not experimental artefacts: ARGO-YBJ (middle) and MILAGRO (top) show the same features, and there appears to be some continuity between the northern and southern hemispheres, particularly in the region between 120 and 150° where the ridge of higher flux seems to continue across the gap (the lower significance for MILAGRO and IceCube close to the equator may be an artefact of reduced effective live time close to the limit of each experiment's acceptance; in the ARGO-YBJ plot, which nearly fills in the gap between MILAGRO and IceCube, the structure looks continuous). Although the "hotspot" in the TA map looks suspiciously close to the top end of this structure, this must surely be a coincidence: the gyroradii of 10 TeV and 60 EeV protons are radically different (0.01 pc and 60 kpc respectively), so similar arrival directions cannot be held to imply similar source directions.

The cause of these medium-scale anisotropies is not currently well understood. The small gyroradius implies that arrival directions should be completely uncorrelated with source directions, so the fact that one of the TeV-energy "hotspots" is close to the Vela pulsar is presumably a coincidence, unless very exotic phenomena are involved (see, e.g., [145]). Nevertheless, if the flux is dominated by a small number of nearby sources, anisotropies in the source distribution could lead to anisotropies in the cosmic ray flux despite the effects of magnetic fields; on the other hand, the anisotropy could reflect the structure of the Galactic magnetic field itself. Theoretical calculations of the propagation of cosmic rays from nearby supernovae generally *overestimate* the expected anisotropy, often by one or two orders of magnitude[124, 146], so the problem appears to be understanding why the effect is small, rather than understanding why it exists at all.

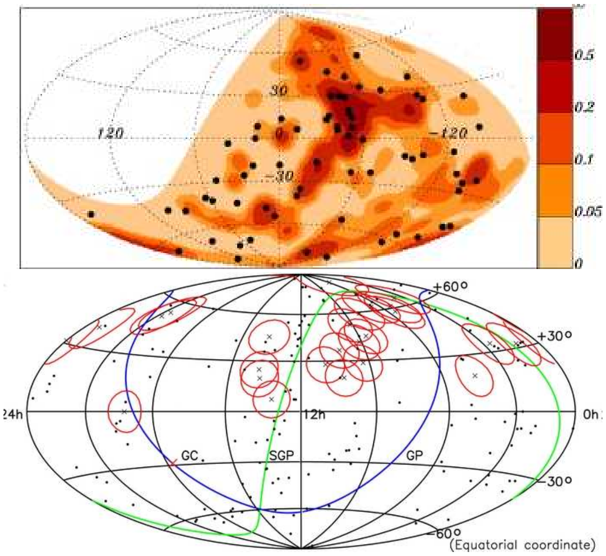


Figure 2.27: Arrival directions of ultra-high-energy cosmic rays. Top, events detected by the Pierre Auger Observatory[147], overlaid on the AGN map from the *Swift*–BAT catalogue of hard X-ray sources[148]. Bottom, events detected by the Telescope Array[149]: the crosses surrounded by red circles are the TA events, while the black dots are the *Swift*–BAT sources.

At very high energies, such as those in the TA plot in figure 2.26, the gyroradius is large enough that it *is* reasonable to expect some correlation between incoming direction and source direction, at least if the incoming cosmic rays are protons; if there is a large component of heavier species, as suggested by the results from the Pierre Auger Observatory (see figure 2.15), the situation is less clear, as the gyroradius is inversely proportional to Z.

Figure 2.27 shows the arrival directions of ultra-high-energy cosmic rays observed by the Pierre Auger Observatory[147] and the Telescope Array[149]. The Auger data are overlaid on a density map representing the AGN of the *Swift*–BAT 58-month catalogue of hard X-ray sources[148], which might reasonably be considered as potential sources; the TA map shows the same sources as dots. Analyses of the correlation between the Auger arrival directions and AGN catalogues suggest some excess over the expectation from isotropy, but the level of correlation has actually decreased as statistics have increased: the latest published value for the VCV AGN catalogue[150] is $(33 \pm 5)\%$[151], compared to $\left(38^{+7}_{-6}\right)\%$ in [147] and $\left(69^{+13}_{-11}\right)\%$ reported in 2007[152]. The most recent value is still more than 2σ above the expectation of 21% from complete isotropy, but seems to require a fairly large isotropic component. The Telescope Array[149] compare their data with a number of appropriate catalogues, finding no significant correlation once the effect of scanning over multiple catalogues is taken into account. Their best correlation ($P = 0.01$) is of events above 57 EeV with the *Swift*–BAT catalogue as shown in figure 2.27, but the total sample is only 25 events (compared to 69 above 55 EeV for Auger). Note that these energies are around the GZK cutoff, so the experiments compare only with nearby AGN, imposing a redshift limit of $z_{\max} \sim 0.018$ (for the TA analysis, this value is optimised separately for each catalogue; Auger uses 0.018 for the VCV catalogue, and weights objects in the *Swift*–BAT catalogue according to the expected GZK attenuation).

Although there are some interesting features in these maps, none is yet statistically significant. Auger observes an excess of events close to the direction of Centaurus A, at 3.8 ± 0.1 Mpc[153] the closest AGN, but this is significant at only 2σ or so (4% of samples drawn from an isotropic distribution had similar or greater overdensities[147]); the TA's "hotspot" (most visible in figure 2.26) has a larger significance of 3.6σ including "look-elsewhere" effects (i.e., the probability of finding such a concentration of events *anywhere* on the sky corresponds to 3.6σ), but is rather broad and does not coincide with a known astrophysical object[144].

To summarise, there is a wealth of information about the flux, energy spectrum, elemental and (to a lesser extent) isotopic composition, time variation, and spatial anisotropy of cosmic rays, acquired from a range of balloon-borne, space-based and ground-based experiments. However, the lack of useful directional information is a serious barrier to full understanding: as discussed below, even the general assumption that Galactic cosmic rays below the "knee" are accelerated in supernova remnants is not a conclusively proven fact. Observations of TeV γ-rays, and especially of high-energy neutrinos (both discussed later in this chapter), preserve directional information and are likely to prove crucial in improving this situation.

2.3 Radio emission

2.3.1 The radio sky

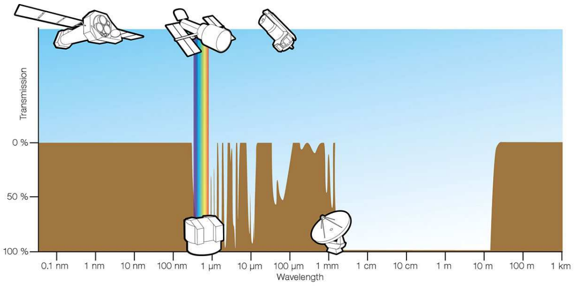


Figure 2.28: Transparency of the atmosphere as a function of wavelength, showing the optical and radio windows. Figure from ESO[154].

For most of history, astronomical observation was carried out using optical wavelengths, for the simple reason that the imaging instrument was generally the human eye. From an evolutionary point of view, the optical waveband is unique: electromagnetic radiation with wavelengths between about 300 and 1000 nm is both copiously produced by the Sun and not significantly absorbed by the Earth's atmosphere. It is therefore not surprising that most visual systems evolved on Earth are sensitive in this region (many insects and birds can see in the near-UV range of 300–400 nm; some freshwater fish can see into the near IR). The UV edge of atmospheric transmission is quite sharp (see figure 2.28); the IR edge is limited by absorption bands from water vapour, with reasonable transmission in some limited wavebands (especially in dry air, hence the tendency for ground-based infrared telescopes to be situated in exceptionally dry places like the South Pole and the Atacama Desert).

However, the optical window is not the only wavelength range at which the atmosphere is transparent: the *radio window*, extending from about 1 mm to 10 m, is much wider and less weather-sensitive (but useless for vision, for two reasons: the Sun is not very luminous at these wavelengths, and the diffraction-limited resolution is much poorer—so it's always dark, and the image produced by a reasonable-sized "eye" is very blurred). Since the 1950s, radio astronomy has become an important branch of ground-based observational astronomy.

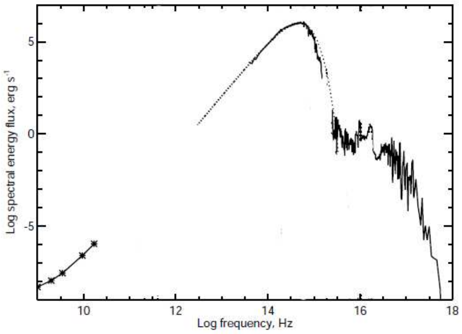


Figure 2.29: The spectral energy distribution, νf_ν, of the quiet Sun, adapted from [159]. The lowest frequencies here correspond to wavelengths of around 30 cm. The UV data are from Warren [160]; the visible/NIR spectrum is the Wehrli Standard Extraterrestrial Solar Irradiance Spectrum[161]; the microwave points come from the Nobeyama Radio Observatory[162]. The dotted line is a blackbody curve for 5778 K.

The radio sky is very different from the visible night sky. Stars are not bright radio sources; the dominant source of radio emission is the disc of the Milky Way, first detected by Jansky[155] and Reber[156, 157] in the 1930s. Superimposed on this are individual sources, but they are not stellar: as long ago as 1954, Baade and Minkowski[158] established that the (then few) identified radio sources were associated with supernova remnants, peculiar galaxies (strong sources), and normal spiral galaxies (faint sources).

The reason for this dramatic difference is that thermal (near-blackbody) sources completely dominate in the visual, but are not very luminous at radio wavelengths, as can be seen from figure 2.29. Although there is some thermal emission at radio wavelengths—the cosmic microwave background is the most obvious example—non-thermal sources contribute a much larger proportion of the total flux. Even the radio emission of the Sun is usually dominated by non-thermal (or, at least, non-blackbody) contributions: the spectral energy distribution shown in figure 2.29 is for solar minimum. Figure 2.30 shows the solar microwave flux and sunspot number from 1951 to 2013; it can be seen that the flux below 4 GHz more than doubles at solar maximum.

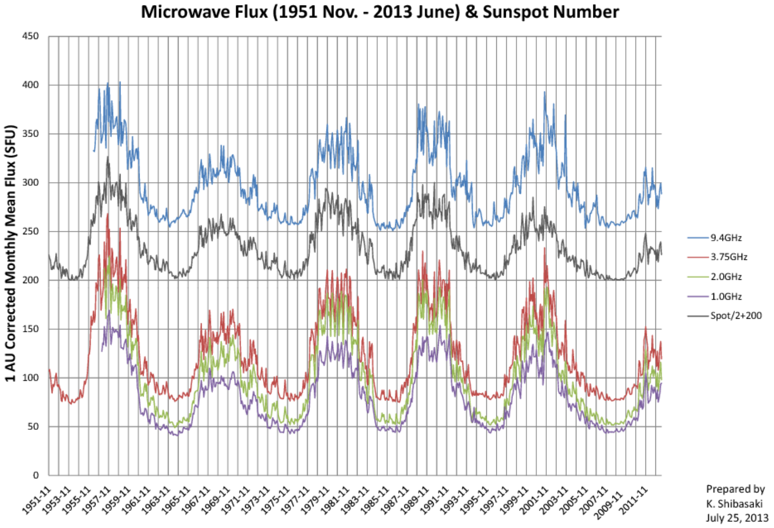


Figure 2.30: Microwave flux from the Sun at various frequencies, compared to sunspot number (note that for ease of comparison, what is plotted for sunspot number $N_{\rm spot}$ is actually $200 + \frac{1}{2}N_{\rm spot}$). Figure by K Shibasaki of the Nobeyama Radio Observatory[163].

It is this dominance of non-thermal sources that makes radio astronomy important in particle astrophysics. The detection of astrophysical radio sources is not inherently particle astrophysics: the techniques owe much to wartime radar, but little to particle physics. However, several of the emission mechanisms discussed below require the presence of relativistic electrons in the source, and these electrons and their acceleration to relativistic energies certainly do lie in the domain of particle astrophysics. Therefore, in this section we shall neglect the many complexities of radio astronomy techniques and technology[164], and instead focus on emission mechanisms and diagnostics.

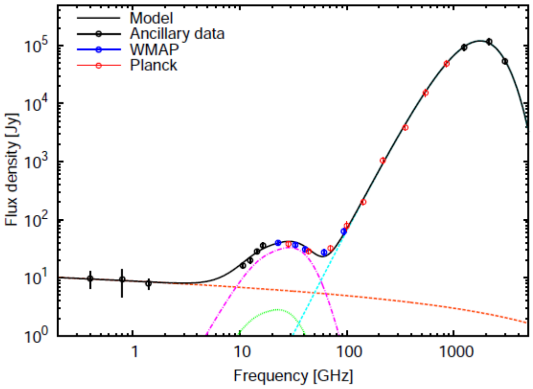


Figure 2.31: Spectrum of AME-G160.26–18.62 in the Perseus molecular cloud[165], using data from WMAP and *Planck* combined with radio surveys at the low-frequency end and DIRBE far-infrared data at high frequencies. The data are very well described by a combination of thermal emission from dust (cyan), free-free emission (orange) and a two-component spinning dust model (high-density molecular gas, magenta, and low-density atomic gas, green).

2.3.2 Radio emission mechanisms

The principal emission mechanisms at radio/microwave wavelengths are listed below. They are arranged approximately in increasing order of wavelength: most produce a continuous spectrum, but the spectra have different frequency dependence, so different mechanisms dominate at different frequencies.

1. *Thermal emission from dust.*

   The plane of our Galaxy contains a great deal of warm dust at temperatures of around 10–30 K. The thermal radiation from this material lies primarily in the far-infrared or submillimetre bands, which are accessed mostly by satellite (see figure 2.28). It is important to astronomers studying star and planet formation, but has little significance in particle astrophysics, so will not be discussed further.

2. *Thermal emission from the early universe (the cosmic microwave background).*

   The CMB peaks at longer wavelengths than the radiation from warm dust, because the temperature is lower. In wavelength units, the peak of the distribution is at just over 1 mm, so—as the name suggests—most of the radiation is in the microwave region, though it does extend into the far infrared. The properties of the CMB, particularly the power spectrum of the small variations in temperature, are important for dark matter and dark energy (see sections 1.3 and 1.6), but not for the high-energy particle astrophysics we are considering in this course; they are covered in the last section of PHY306/406[7].

3. *Spinning dust.*

   "Anomalous microwave emission"[165] is an additional component of diffuse microwave emission that occurs in the frequency range 10–60 GHz (wavelengths of 5–30 mm), and is therefore important as a foreground for CMB studies. It is known to be spatially correlated with thermal emission from dust, and is believed to be emitted by very small (i.e. the size of large molecules), rapidly-spinning dust grains made primarily of organic compounds (probably polycyclic aromatic hydrocarbons, basically compounds containing multiple benzene rings). The grains spin as a result of collisions with other grains, and pick up charge as a consequence of collisions with photons (causing emission of electrons by the photoelectric effect), free electrons, and ions: the combination of non-zero charge and non-zero spin causes them to emit electromagnetic radiation. This model describes the observed emission well, as can be seen in figure 2.31.

Besides being a significant foreground for CMB analyses, spinning dust emission is interesting for studies of the interstellar medium and potentially also for astrobiology (since polycyclic aromatic hydrocarbons represent an important class of prebiotic organic chemistry), but has little relevance to particle astrophysics.

4. *Line emission from gas.*

   The most famous radio spectral line is the 21 cm line of neutral hydrogen, caused by the electron spin flipping from parallel to the proton's spin to antiparallel, but many important molecules also have spectral lines in the radio, microwave or submillimetre wavebands[166]. These lines are obviously important for studying the chemistry of molecular clouds; in addition, the 21 cm line is essential for mapping neutral atomic hydrogen and measuring the rotation curves of spiral galaxies, CO is the standard tracer for molecular gas (the dominant constituent of molecular gas, H_2, produces few lines because it is a symmetrical molecule) and OH (hydroxyl) and water lines are frequently found as natural masers (microwave lasers), which can be used, for example, to determine the distances of other galaxies. However, as with dust emission, radio spectral lines are essentially low-temperature phenomena and thus not of importance in particle astrophysics.

5. *Bremsstrahlung, or free-free emission.*

   Bremsstrahlung (from the German *bremsen*, to brake, and *Strahlung*, radiation: thus, "braking radiation") is radiation emitted when an electron loses energy in an interaction with an ion (in principle, in the interaction of any two charged particles, but electron-ion collisions are by far the most effective in producing radiation). In astrophysics, bremsstrahlung[2] generally occurs in ionised gases, and is also known as *free-free emission* because the electron is not bound to the ion either before or after the interaction.

   Most astrophysical bremsstrahlung involves fast-moving but non-relativistic electrons in hot plasma (*thermal bremsstrahlung*). Relativistic electrons interacting with ions also emit bremsstrahlung, but the photon energy scales with the electron energy (see below); relativistic bremsstrahlung is therefore more likely to be observed in γ-rays than at radio wavelengths.

[2] In German, this would have a capital B, since all nouns are capitalised in German. I am following the Oxford English Dictionary in accepting that it has been imported into English and thus no longer needs a capital. End of grammar nit-pick!

6. *Synchrotron radiation.*

   Synchrotron radiation is radiation emitted by highly relativistic electrons gyrating in a magnetic field, and is the dominant source of astrophysical radio emission at low frequencies. It is also produced at terrestrial particle accelerators[167], where it was first observed[168] and from which it gets its name. *Cyclotron radiation*, which is the same phenomenon, but produced by only mildly relativistic electrons, is much less common, but is also observed in some astrophysical sources[169, 170].

   Synchrotron radiation is diagnostic of the presence of relativistic electrons and is therefore an important observational tool for particle astrophysics. As we shall see below, synchrotron radiation at radio wavelengths is often associated with X-ray and γ-ray emission caused by inverse Compton scattering of the same electron population.

   In the literal sense, synchrotron radiation is also "bremsstrahlung", in that it is emission of electromagnetic radiation by an accelerated charged particle. However, conventionally, the term bremsstrahlung is restricted to emission as a consequence of particle-particle interactions, as opposed to cases where the interaction is between a particle and an ambient magnetic field.

2.3.3 Electromagnetic radiation from an accelerated charge

The key ingredient of both bremsstrahlung and synchrotron radiation is the radiation emitted by an accelerated charged particle. The derivation that follows is taken from Malcolm Longair's *High Energy Astrophysics*[171], section 6.2; he credits it to JJ Thomson.

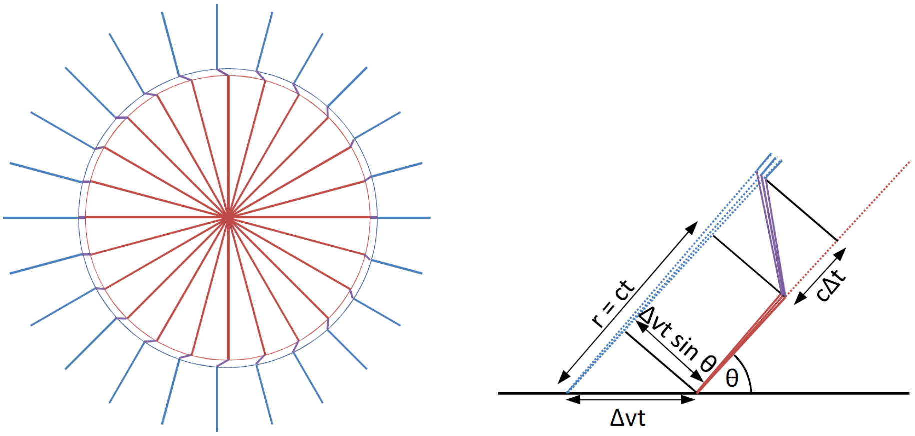


Figure 2.32: Radiation from an accelerated charge. Left, the field lines a time t after a small acceleration $\Delta v/\Delta t$. Right, a close-up of the "kink" in the field lines at distance $r = ct$ from the charge. Based on figure 6.1 of [171].

At time $t = 0$ a charge Q is stationary in reference frame S. It then accelerates by an amount Δv in a time Δt. After a time t, the field lines from the particle must have a kink in them, as shown in figure 2.32: at distances further than $r = ct$, the field lines do not "know" that the charge has moved. This kink introduces an azimuthal component of the electric field which moves outward as t increases, so it represents a pulse of electromagnetic radiation emitted by the particle. (As the change in velocity $\Delta v \ll c$, we can assume that the field lines at $r < ct$ and $r > c(t + \Delta t)$ are purely radial. In principle they aren't, because of aberration, but this effect is extremely small for small Δv.)

The strength of the azimuthal component of the electric field in the kink is given by

$$\frac{E_\theta}{E_r} = \frac{\Delta v t \sin\theta}{c\Delta t} \tag{2.11}$$

and the radial field is given by Coulomb's law,

$$E_r = \frac{Q}{4\pi\epsilon_0 r^2} = \frac{Q}{4\pi\epsilon_0 c^2 t^2}.$$

Therefore

$$E_\theta = \frac{Q\sin\theta}{4\pi\epsilon_0 c^2 r}\frac{\Delta v}{\Delta t} = \frac{Q\ddot{\mathbf{r}}\sin\theta}{4\pi\epsilon_0 c^2 r}, \tag{2.12}$$

where $\ddot{\mathbf{r}} = \Delta v/\Delta t$ is the acceleration of the particle.

The energy carried by this electromagnetic pulse is given by the *Poynting vector* (with dimensions of power per unit area)

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}, \tag{2.13}$$

where μ_0 is the permeability of free space; recall that $\mu_0\epsilon_0 = 1/c^2$. For an electromagnetic wave in free space $E/B = c$ and $\mathbf{E}$ is perpendicular to $\mathbf{B}$, so

$$S = \frac{E^2}{c\mu_0} = c\epsilon_0 E^2.$$

The power $P(\theta)\mathrm{d}\Omega$ radiated by our charge through solid angle $\mathrm{d}\Omega$ at angle θ to the direction of the acceleration is

$$P(\theta)\mathrm{d}\Omega = \frac{Q^2|\ddot{\mathbf{r}}|^2\sin^2\theta}{16\pi^2\epsilon_0 c^3 r^2}r^2\mathrm{d}\Omega = \frac{Q^2|\ddot{\mathbf{r}}|^2\sin^2\theta}{16\pi^2\epsilon_0 c^3}\mathrm{d}\Omega. \tag{2.14}$$

To get the total power, we integrate over solid angle:

$$P_{\mathrm{rad}} = \frac{Q^2|\ddot{\mathbf{r}}|^2}{16\pi^2\epsilon_0 c^3}\,2\pi\int_{-1}^{+1}(1-\cos^2\theta)\mathrm{d}(\cos\theta) = \frac{Q^2|\ddot{\mathbf{r}}|^2}{6\pi\epsilon_0 c^3}. \tag{2.15}$$

This expression is called *Larmor's formula.* It is true in any frame, because $\mathrm{d}E/\mathrm{d}t$ is Lorentz invariant, although the angular distribution will change dramatically in different reference frames. The acceleration $\ddot{\mathbf{r}}$ is the *proper acceleration*, i.e. the acceleration measured in the instantaneous rest frame of the particle.

From equation (2.14) we note that, considered in its rest frame, the particle radiates as a dipole: the electric field is $\propto \sin\theta$ and the power $\propto \sin^2\theta$. The radiation is zero in the direction of the acceleration and maximal perpendicular to this. Also, as the kink in the electric field lines is always parallel to the acceleration vector, the radiation is *polarised.* This is a useful diagnostic tool: both synchrotron radiation and cyclotron radiation have characteristic polarisation signatures (in contrast, bremsstrahlung is typically unpolarised, because the electrons approach the ions at random orientations; although each individual encounter produces polarised light, there is no preferred direction and hence no net polarisation).

To express equation (2.15) in terms of the measured acceleration in the lab frame, we introduce the acceleration four-vector

$$\begin{aligned} a_\mu = \gamma \frac{\partial v_\mu}{\partial t} &= \gamma \frac{\partial}{\partial t}(\gamma c; \gamma \mathbf{v}) \\ &= \left(\gamma^4 \frac{\mathbf{v} \cdot \mathbf{a}}{c}; \gamma^2 \mathbf{a} + \gamma^4 \left(\frac{\mathbf{v} \cdot \mathbf{a}}{c^2} \right) \mathbf{v} \right), \end{aligned} \tag{2.16}$$

where v_μ is the four-velocity $\gamma(c; \mathbf{v})$, $\mathbf{v}$ is the ordinary vector velocity, and $\mathbf{a}$ is the ordinary vector acceleration. The acceleration four-vector in the instantaneous rest frame of the particle (the proper acceleration four-vector) is $(0; \ddot{\mathbf{r}})$ where $\ddot{\mathbf{r}}$ is the proper acceleration as it appears in equation (2.15). Since the magnitude of the four-vector is a Lorentz invariant, we must have

$$|\ddot{\mathbf{r}}|^2 = \left(\gamma^2 \mathbf{a} + \gamma^4 \left(\frac{\mathbf{v} \cdot \mathbf{a}}{c^2} \right) \mathbf{v} \right)^2 - \left(\gamma^4 \frac{\mathbf{v} \cdot \mathbf{a}}{c} \right)^2,$$

which simplifies to

$$|\ddot{\mathbf{r}}|^2 = \gamma^4 \left(|\mathbf{a}|^2 + \gamma^2 \left(\frac{\mathbf{v} \cdot \mathbf{a}}{c} \right)^2 \right). \tag{2.17}$$

It is often useful to split this into components parallel to and perpendicular to the velocity, $a_\parallel$ and $a_\perp$. Since $\mathbf{v} \cdot \mathbf{a} = v a_\parallel$, this gives

$$|\ddot{\mathbf{r}}|^2 = \gamma^4 \left(a_\perp^2 + \gamma^2 a_\parallel^2 \right). \tag{2.18}$$

The spectrum of the radiation from an accelerated charge is obtained by taking the Fourier transform of the acceleration:

$$\ddot{R}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega t} \ddot{r}(t) \, \mathrm{d}t.$$

By Parseval's Theorem[172],

$$\int_{-\infty}^{\infty} |\ddot{r}(t)|^2 \, \mathrm{d}t = \int_{-\infty}^{\infty} |\ddot{R}(\omega)|^2 \, \mathrm{d}\omega =$$

Also, for a real function,

$$\int_{0}^{\infty} |\ddot{R}(\omega)|^2 \, \mathrm{d}\omega = \int_{-\infty}^{0} |\ddot{R}(\omega)|^2 \, \mathrm{d}\omega,$$

so the total emitted radiation is given by

$$\int_{-\infty}^{\infty} P_{\mathrm{rad}}(t) \, \mathrm{d}t = \int_{-\infty}^{\infty} \frac{Q^2 |\ddot{\mathbf{r}}(t)|^2}{6\pi\epsilon_0 c^3} \, \mathrm{d}t = \int_{-\infty}^{\infty} \frac{Q^2 |\ddot{R}(\omega)|^2}{6\pi\epsilon_0 c^3} \, \mathrm{d}\omega = 2 \int_{0}^{\infty} \frac{Q^2 |\ddot{R}(\omega)|^2}{6\pi\epsilon_0 c^3} \, \mathrm{d}\omega.$$

Therefore, the spectrum of the radiation is given by

$$I(\omega) \, \mathrm{d}\omega = \frac{Q^2 |\ddot{R}(\omega)|^2}{3\pi\epsilon_0 c^3} \, \mathrm{d}\omega. \tag{2.19}$$

$I(\omega)$ is the total energy per unit bandwidth emitted over the entire time taken for the interaction (formally, the whole of time from $-\infty$ to ∞, but of course the emission is really restricted to the period for which the charge is being accelerated).

2.3.4 Bremsstrahlung

Figure 2.33 shows the path of an electron past an ion of charge Ze. The essential parameter in this is the impact parameter b, defined as the distance of closest approach of the electron to the ion, but often taken to be approximately equal to the perpendicular distance between the initial trajectory of the electron and the ion (labelled b' on figure). As can be seen from the bottom panel of figure 2.33, most of the acceleration takes place over a short period $\tau \simeq 2b/\gamma v$.

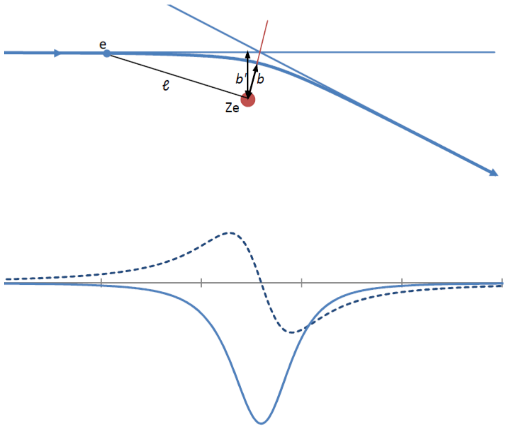


Figure 2.33: Bremsstrahlung. Top, path of electron past an ion of charge Ze. Bottom, acceleration of the electron parallel to (dashed line) and perpendicular to (solid line) its velocity, for the nonrelativistic case.

We can simplify the calculation considerably by assuming that the encounter is sufficiently distant (i.e. b is sufficiently large) that (1) $b' \simeq b$ (the direction of the electron is not much altered) and (2) the *speed* of the electron is unchanged (only its direction changes). A more precise treatment can be found in sections 6.3 to 6.6 of Longair[171].

Defining $t = 0$ to be the time of closest approach, the acceleration of the electron at time t is given by

$$\begin{aligned} a_{\|}(t) &= \frac{Ze^2}{4\pi\epsilon_0 m_e}\frac{\gamma v t}{(b^2+(\gamma v t)^2)^{3/2}}; \\ a_{\perp}(t) &= \frac{Ze^2}{4\pi\epsilon_0 m_e}\frac{\gamma b}{(b^2+(\gamma v t)^2)^{3/2}}, \end{aligned} \tag{2.20}$$

where v is the speed of the electron and m_e is its mass (the γ factor in the $a_\perp$ equation comes from the transformation of the electric field, see Longair[171] section 5.3).

To get the emitted spectrum, we take the Fourier transform of these:

$$\begin{aligned} A_{\|}(\omega) &= \frac{1}{\sqrt{2\pi}}\frac{Ze^2}{4\pi\epsilon_0 m_e}\int\limits_{-\infty}^{\infty}\frac{\gamma v t}{(b^2+(\gamma v t)^2)^{3/2}}e^{i\omega t}\mathrm{d}t; \\ A_{\perp}(\omega) &= \frac{1}{\sqrt{2\pi}}\frac{Ze^2}{4\pi\epsilon_0 m_e}\int\limits_{-\infty}^{\infty}\frac{\gamma b}{(b^2+(\gamma v t)^2)^{3/2}}e^{i\omega t}\mathrm{d}t. \end{aligned} \tag{2.21}$$

These look like thoroughly unappealing integrals. However, for radio emission, where the frequencies involved are low ($\omega\tau \ll 1$), we can assume that radiation resulting from parallel acceleration is negligible (averaged over the wavelength of a radio wave, the bipolar acceleration pulse shown by the dashed line in figure 2.33 will sum to zero), and that the perpendicular acceleration pulse looks like a delta function. The Fourier transform of a delta function in the time domain is flat across all frequencies, so

$$A_\perp(\omega) \simeq \frac{1}{\sqrt{2\pi}}\Delta v_\perp,$$

where $\Delta v_\perp$ is the area under the $a_\perp$ "delta function". Assuming that v and γ do not change, this is given by

$$\begin{aligned}\Delta v_\perp &= \frac{Ze^2}{4\pi\epsilon_0 m_e}\int_{-\infty}^{\infty}\frac{\gamma b\,\mathrm{d}t}{(b^2+(\gamma v t)^2)^{3/2}}\\ &= \frac{Ze^2\gamma}{4\pi\epsilon_0 m_e b^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{(1+(\gamma v t/b)^2)^{3/2}} = \frac{2Ze^2}{4\pi\epsilon_0 m_e v b},\end{aligned} \tag{2.22}$$

so we have

$$I(\omega) = \frac{e^2}{3\pi\epsilon_0 c^3}|A(\omega)|^2 = \frac{Z^2e^6}{24\pi^4\epsilon_0^3c^3m_e^2v^2b^2}, \tag{2.23}$$

i.e. the spectrum of bremsstrahlung is flat at low frequencies. In fact, to a very good approximation the spectrum is flat up to frequencies of order $\omega \sim \gamma v/b$ and then falls off exponentially.

The above derivation was for a single electron with a fixed velocity v and impact parameter b. The next step is to extend this to deal with a population of electrons of fixed velocity interacting with a population of stationary ions. The rate of collisions with impact parameter between b and $b + \mathrm{d}b$ is

$$n_e n_i \gamma v 2\pi b\,\mathrm{d}b,$$

where n_e and n_i are the number densities of electrons and ions respectively (to see where the $2\pi b\,\mathrm{d}b$ comes from, think of a target, e.g. for archery: a ring of width $\mathrm{d}b$ at distance b from the centre has area $2\pi b\,\mathrm{d}b$). Multiplying $I(\omega)$ by this and integrating over b gives

$$P(\omega, v) = \frac{Z^2e^6n_en_i\gamma}{12\pi^3\epsilon_0^3c^3m_e^2v}\int_{b_{\min}}^{b_{\max}}\frac{\mathrm{d}b}{b} = \frac{Z^2e^6n_en_i\gamma}{12\pi^3\epsilon_0^3c^3m_e^2v}\ln\frac{b_{\max}}{b_{\min}}. \tag{2.24}$$

The limits $b_{\max}$ and $b_{\min}$ have to be estimated from the physics of the situation: fortunately, since they are inside a log, the estimates do not need to be particularly precise. The more obvious one is $b_{\max}$: we said above that the emission falls off exponentially above $\omega \sim \gamma v/b$, so we should only integrate out to $b_{\max} = \gamma v/\omega$. At high velocities, the uncertainty principle gives us $b_{\min} = \hbar/(2m_e v)$. (One might expect this to be $\hbar/(2\gamma m_e v)$. However, we need the radiation to be coherent across the size of the electron, say Δx: therefore $\Delta t = b_{\min}/\gamma v$ must be equal to $\Delta x/v$. If we then use the uncertainty principle, $\Delta x\Delta p \geq \hbar/2$, with $\Delta p = \gamma m_e v$, the result follows.)

At low velocities, there is a classical limit: for our assumption of a straight line trajectory to be even approximately right, we should have $\Delta v_\perp \leq v$, which implies $b_{\min} = Ze^2/(2\pi\epsilon_0 m_e v^2)$ from equation (2.22). Longair[171] gives a lower value, $b_{\min} = Ze^2/(8\pi\epsilon_0 m_e v^2)$, based on the maximum possible momentum transfer, which is hardly consistent with a straight line—but the difference this makes to the log is small anyway. Which of these lower limits, the classical or the quantum, is appropriate depends on the astrophysical situation: for radio emission from an HII region at $\sim 10^4$ K, the classical limit is relevant, whereas for the intracluster medium of a rich cluster of galaxies at $\sim 10^8$ K, the quantum limit would apply. The boundary value of v is about $Zc/137$, where the factor $1/137$ is the fine structure constant α.

Equation (2.24) is essentially identical to the result of the full quantum-mechanical treatment[173, 171] except for a somewhat different form of the logarithm: for a non-relativistic electron with kinetic energy $E = \frac{1}{2}m_e v^2$, the exact result is

$$\begin{aligned} P(\omega, v) &= \frac{8}{3} Z^2 \alpha \hbar r_e^2 \frac{m_e c^2}{E} v n_i \ln \frac{1 + \sqrt{1 - \hbar\omega/E}}{1 - \sqrt{1 - \hbar\omega/E}} \\ &= \frac{Z^2 e^6 n_i}{12\pi^3 \epsilon_0^3 c^3 m_e^2 v} \ln \frac{1 + \sqrt{1 - \hbar\omega/E}}{1 - \sqrt{1 - \hbar\omega/E}}, \end{aligned} \tag{2.25}$$

where $\alpha = e^2/(4\pi\epsilon_0 \hbar c)$ and r_e is the classical radius of the electron, $r_e = e^2/(4\pi\epsilon_0 m_e c^2)$. The part in front of the logarithm is identical to equation (2.24) apart from the lack of the factors n_e (this is for a single electron) and γ (the electron is non-relativistic). For low photon energies, $\hbar\omega \ll E$, the logarithm reduces to $\ln(4E/\hbar\omega)$, which is what we get from equation (2.24) if we use the quantum expression for b_{min}.

To determine the rate of energy loss by bremsstrahlung, we need to integrate equation (2.24) or (2.25) over frequency. As the spectrum is flat up to a maximum frequency ω_{max} and then cuts off rapidly, to a good approximation we can just multiply the power per unit bandwidth by ω_{max}. As an order of magnitude estimate, this maximum frequency is given by $\hbar\omega_{\text{max}} = \frac{1}{2}m_e v^2$: the electron cannot lose more than its entire initial kinetic energy. Using this value, the rate of energy loss for a non-relativistic electron is

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right)_{\text{brem}} = \frac{Z^2 e^6 n_i v}{24\pi^3 \epsilon_0^3 c^3 m_e \hbar} \ln \frac{b_{\text{max}}}{b_{\text{min}}}, \tag{2.26}$$

i.e. the rate of energy loss is proportional to $Z^2 n_i v$, or to $Z^2 n_i \sqrt{E}$ where E is the electron kinetic energy.

Relativistic electrons, in contrast, are likely to be colliding with neutral atoms of interstellar gas rather than being part of a hot plasma. Therefore, the maximum impact parameter b_{max} is set by the screening effect of the atomic electrons: the electron will not “see” the nuclear charge unless its trajectory takes it inside the electron cloud. Longair[171] uses the semi-classical Thomas-Fermi model of the atom[174, 175, 176] to justify an estimate of

$$b_{\text{max}} = 1.4 r_0 Z^{-1/3},$$

where r_0 is the Bohr radius of the hydrogen atom, $r_0 = 4\pi\epsilon_0\hbar^2/(m_e c^2)$. The value of b_{min} can be taken to be the quantum limit $b_{\text{min}} = \hbar/(2m_e v)$ derived earlier. In addition, we need to transform $P(\omega, v)$ from equation (2.24) from the instantaneous rest frame of the electron into the lab frame; this just requires us to divide by γ (to take into account the fact that the bandwidth, $\Delta\omega$, increases by a factor γ owing to time dilation, and one therefore has to divide by γ to get back to unit bandwidth; recall that the total rate of energy radiation, $\mathrm{d}E/\mathrm{d}t$, is an invariant).

The power radiated per unit bandwidth is independent of frequency up to a cutoff value ω_{max} given by $\omega_{\text{max}} = (\gamma - 1)m_e c^2$; as in the non-relativistic case, this is estimated by equating the energy of the photon to the total kinetic energy E of the electron. Multiplying equation (2.24) by ω_{max} and inserting the limits on b gives, for a single relativistic electron,

$$P(E) = \frac{Z^2 e^6 n_i E}{12\pi^3 \epsilon_0^3 m_e^2 c^4 \hbar} \ln \frac{192}{Z^{1/3}},$$

where we have taken $c^3 v \simeq c^4$ since this is a relativistic electron. This again compares well with the result of the full quantum-mechanical treatment[173, 171]

$$P(E) = \frac{Z(Z+1.3)e^6 n_i E}{16\pi^3 \epsilon_0^3 m_e^2 c^4 \hbar} \ln\left(\frac{183}{Z^{1/3}} + \frac{1}{8}\right); \tag{2.27}$$

apart from the slight difference in the logarithm, the main change is that Z^2 is replaced by $Z(Z+1.3)$. This comes from including the interaction of the incoming electron with the electron cloud of the atom, as well as with its nucleus. Note that in this case the radiated power is proportional to E, instead of to $\sqrt{E}$ as in the non-relativistic case.

For astrophysical applications, these formulae have to be integrated over the velocity distribution of the relevant electron population. In the non-relativistic case, the relevant distribution is usually thermal, i.e. Maxwellian:

$$n_e(v)\,\mathrm{d}v = 4\pi n_e \left(\frac{m_e}{2\pi kT}\right)^{3/2} v^2 \exp\left(-\frac{m_e v^2}{2kT}\right)\mathrm{d}v. \tag{2.28}$$

This is a non-trivial integral; the usual approach is to take an order-of-magnitude estimate $\frac{1}{2}m_e v^2 = \frac{3}{2}kT$ and subsume the details of the integration into the **Gaunt factor** $g(\omega, T)$, which is numerically of order unity. This gives, at low frequencies,

$$P(\omega) = \frac{Z^2 e^6 n_e n_i}{12\sqrt{3}\pi^3 \epsilon_0^3 c^3 m_e^2} \left(\frac{m_e}{kT}\right)^{1/2} g(\omega, T), \tag{2.29}$$

where the $\sqrt{3}$ has appeared because it's in the conventional definition of the Gaunt factor. All the frequency dependence is inside $g(\omega, T)$, and there is a high-frequency cutoff $\propto \exp(-\hbar\omega/kT)$, so integrating $P(\omega)$ over ω amounts to multiplying the coefficient by $\omega_{\max} = kT/\hbar$ and averaging g over ω. This leads to an expression for the radiated power[171]

$$-\frac{\mathrm{d}E}{\mathrm{d}t} \propto Z^2 n_i n_e T^{1/2} \bar{g}. \tag{2.30}$$

Longair[171] quotes the numerical results for the full calculation, expressed in terms of frequency ν rather than angular frequency ω. The spectral emissivity is

$$j_\nu = 6.8 \times 10^{-51} Z^2 T^{-1/2} n_i n_e g(\nu, T) \exp\left(-\frac{h\nu}{kT}\right) \text{ W m}^{-3} \text{ Hz}^{-1}, \tag{2.31}$$

where

$$g(\nu, T) = \begin{cases} \dfrac{\sqrt{3}}{2\pi}\left[\ln\left(\dfrac{128\epsilon_0^2 k^3 T^3}{m_e e^4 \nu^2 Z^2}\right) - \gamma_{\mathrm{E}}^{1/2}\right] & \text{(radio);} \\[2ex] \dfrac{\sqrt{3}}{\pi}\ln\left(\dfrac{kT}{h\nu}\right) & \text{(X-rays),} \end{cases} \tag{2.32}$$

where $\gamma_{\mathrm{E}} = 0.5772...$ is Euler's constant (I have added the subscript E to distinguish it from the relativistic γ factor).

The total energy loss rate is given by

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = 1.435 \times 10^{-40} Z^2 n_i n_e T^{1/2} \bar{g} \text{ W m}^{-3}, \tag{2.33}$$

where $\bar{g} \simeq 1.2$.

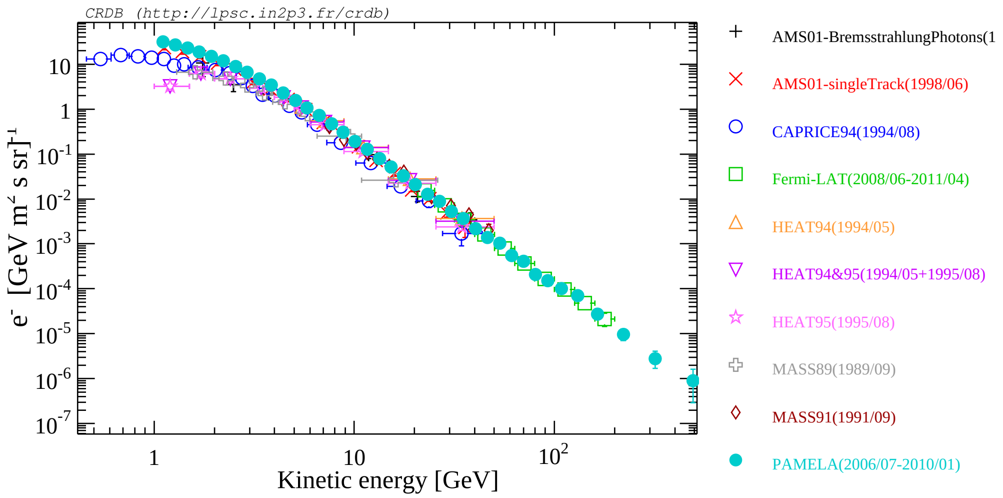


Figure 2.34: Energy spectrum of cosmic ray electrons. Data from balloon and space-based experiments since 1980, extracted from the Cosmic Ray Database[123]. Note the power law spectrum above ∼3 GeV; below this, the flux is affected by solar modulation.

For relativistic bremsstrahlung, the situation is usually simpler, as such electrons are not so likely to be found in a thermal distribution. Cosmic ray electrons have a power-law spectrum, as shown in figure 2.34; for a power law electron spectrum with a given spectral index α, energy loss by relativistic bremsstrahlung results in a photon spectrum, $\mathrm{d}N_\gamma/\mathrm{d}E_\gamma$, with the same spectral index.

Bremsstrahlung is the dominant mode of energy loss for high-energy electrons (energy loss for low-energy electrons is dominated by ionisation losses, see Longair[171] chapter 5). The critical energy at which bremsstrahlung starts to dominate over ionisation depends on the material: it is 340 MeV for hydrogen[171], and less than this for other materials as a result of the Z^2 dependence of bremsstrahlung. Since the energy loss rate is proportional to the electron energy, as shown in equation (2.27), the electron energy declines exponentially as it travels through matter. This can be expressed in terms of a *radiation length* X_0, which is the distance over which the electron energy drops by a factor of e, or by a column density $\xi_0 = \rho X_0$; for bremsstrahlung energy loss by relativistic electrons, the values of ξ_0 for hydrogen and air are[171] 580 kg m^{-2} and 365 kg m^{-2} respectively. As the total column density of air in the Earth's atmosphere is about 10^4 kg m^{-2}, this shows that cosmic-ray electrons initiate electromagnetic showers very high up in the atmosphere.

2.3.5 Synchrotron radiation

Synchrotron radiation is emitted by particles moving in a magnetic field. The equation of motion for a particle of charge Ze in a uniform static magnetic field $\mathbf{B}$ is

$$\frac{\mathrm{d}(\gamma m_0 \mathbf{v})}{\mathrm{d}t} = Ze(\mathbf{v} \times \mathbf{B}).$$

Since $\gamma = (1 - v^2/c^2)^{-1/2}$, the left-hand side of this is

$$\frac{\mathrm{d}(\gamma m_0 \mathbf{v})}{\mathrm{d}t} = \gamma m_0 \mathbf{a} + \gamma^3 \frac{\mathbf{v} \cdot \mathbf{a}}{c} m_0 \mathbf{v},$$

where $\mathbf{a} = \mathrm{d}\mathbf{v}/\mathrm{d}t$, but the second term must be zero because the cross product on the right-hand side guarantees that $\mathbf{a}$ is perpendicular to $\mathbf{v}$. If we define $v_\parallel$

and $v_\perp$ as the components of the velocity parallel to and perpendicular to the magnetic field, then $v_\parallel$ = constant and we have

$$\gamma m_0 a_\perp = Zev_\perp B, \tag{2.34}$$

where the direction of $a_\perp$ is perpendicular to $v_\perp$ and to $\mathbf{B}$. The particle moves in a spiral path with constant pitch angle $\theta = \tan^{-1}(v_\perp/v_\parallel)$ and radius r_g (the *gyroradius*) given by

$$\frac{v_\perp^2}{r_g} = \frac{v^2 \sin^2\theta}{r_g} = \frac{ZevB\sin\theta}{\gamma m_0},$$

i.e.

$$r_g = \frac{\gamma m_0 v \sin\theta}{ZeB}. \tag{2.35}$$

The corresponding *angular gyrofrequency* ω_r is

$$\omega_r = \frac{v_\perp}{r_g} = \frac{ZeB}{\gamma m_0}; \tag{2.36}$$

note that this is independent of the pitch angle. The *gyrofrequency*, ν_r, is $\omega_r/2\pi$ as usual.

The total energy loss from synchrotron radiation is found by combining equations (2.15), (2.18) and (2.34):

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) = \frac{Z^2e^2}{6\pi\epsilon_0 c^3}\gamma^4 a_\perp^2 = \frac{Z^4e^4B^2}{6\pi\epsilon_0 c}\frac{v^2}{c^2}\frac{\gamma^2}{m_0^2}\sin^2\theta. \tag{2.37}$$

Since $\gamma = E/m_0c^2$, for a particle of a given energy the power radiated is proportional to $1/m_0^4$. This explains why synchrotron radiation from astrophysical sources is totally dominated by electrons (and also why it is more difficult to construct high-energy electron accelerators than proton accelerators).

The quantity

$$\sigma_\mathrm{T} = \frac{e^4}{6\pi\epsilon_0^2 c^4 m_e^2} \tag{2.38}$$

is the Thomson cross-section for electron scattering, and $B^2/2\mu_0 = \frac{1}{2}\epsilon_0 c^2 B^2$ is the energy density of a magnetic field, U_mag. Therefore, assuming we are dealing with electrons so that $Z = -1$ and $m_0 = m_e$, equation (2.37) can be written in the form

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) = 2c\sigma_\mathrm{T}U_\mathrm{mag}\beta^2\gamma^2\sin^2\theta, \tag{2.39}$$

where $\beta = v/c \simeq 1$ for relativistic electrons. This form is useful because of its similarity with the equivalent equation for inverse Compton scattering, see below.

Assuming that the distribution of pitch angles is isotropic in $\cos\theta$, $p(\theta)\,\mathrm{d}\theta = \frac{1}{2}\mathrm{d}(\cos\theta)$ (the $\frac{1}{2}$ is there for normalisation), we can average equation (2.39) over pitch angle to get

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) = 2c\sigma_\mathrm{T}U_\mathrm{mag}\beta^2\gamma^2\frac{1}{2}\int_{-1}^{+1}(1-\cos^2\theta)\mathrm{d}(\cos\theta) = \frac{4}{3}c\sigma_\mathrm{T}U_\mathrm{mag}\beta^2\gamma^2. \tag{2.40}$$

As noted in section 2.3.3, radiation from an accelerated charge is polarised. In bremsstrahlung, this does not result in net polarisation because the electrons encounter ions at random angles: the individual interactions produce polarised light, but there is no preferred direction overall. In contrast, the magnetic field responsible for synchrotron emission are ordered over quite large distances, so synchrotron radiation typically has a high degree of polarisation.

Cyclotron radiation

Cyclotron radiation is produced by non-relativistic or mildly relativistic electrons gyrating in magnetic fields. For a non-relativistic electron, $\gamma \simeq 1$ in equation (2.39) and the radiated power is

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) = \frac{2\sigma_{\mathrm{T}}}{c} U_{\mathrm{mag}} v^2 \sin^2\theta,$$

emitted at the *cyclotron frequency* $\nu_g = eB/(2\pi m_e)$ (the gyrofrequency ν_r for the case $\gamma = 1$). The polarisation depends on the viewing angle: if the magnetic field is oriented perpendicular to the line of sight, the radiation will be linearly polarised; if the field is along the line of sight, the radiation will be circularly polarised; intermediate angles give elliptical polarisation.

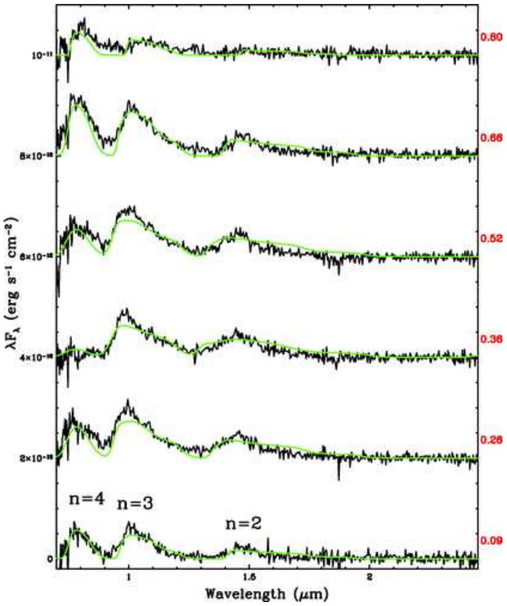


Figure 2.35: Cyclotron resonance lines in the near infra-red, observed in the AM Herculis type binary system EQ Ceti[177]. The model (green lines) assumes a magnetic field of 34 MG (3.4 kT). The red numbers at the side are the phase of the binary orbit. The lines are interpreted as $l = 2, 3, 4$ (labelled n on figure). Figure from [177].

If the electron is mildly relativistic, the symmetry of the radiation pattern will be slightly distorted by relativistic aberration (see below). This results in radiation at higher harmonics of the relativistic gyrofrequency $\nu_r = eB/(2\pi m_e \gamma)$ (see Longair[171], section 8.2, for details):

$$\nu_l = \frac{l}{1 - \beta_{\|} \cos\theta} \nu_r, \qquad (2.41)$$

where $l = 1, 2, 3, ...$ and $\beta_{\|} = v_{\|}/c$ is the component of the electron velocity in the observer's line of sight. The denominator $(1 - \beta_{\|} \cos\theta)$ is therefore the Doppler shift produced by the electron's velocity parallel to the field lines, $v\cos\theta$, and—since the pitch angle θ differs for different electrons—has the effect of broadening the observed lines. As the electron's velocity increases, this broadening effect becomes more significant, washing out the line structure and producing a continuous spectrum.

Cyclotron lines have been observed in some astrophysical systems, particularly X-ray pulsars and AM Herculis type binary systems; an example of the latter is shown in figure 2.35.

Synchrotron radiation

As the electron's velocity increases and it becomes more relativistic, the polar diagram of the cyclotron emission becomes more and more distorted by aberration. Defining ϕ as the angle to the electron velocity vector, the transformation between ϕ' in the instantaneous rest frame of the electron and ϕ in the lab frame is

$$\cos\phi = \frac{\cos\phi' + \beta}{1 + \beta\cos\phi'}. \qquad (2.42)$$

For $\cos\phi' = 0$, this gives $\sin\phi = \sqrt{1-\beta^2} = 1/\gamma$, i.e. the leading lobe of the emission becomes concentrated into a narrow beam of half-angle $1/\gamma$. This effect, which is known as *beaming*, is shown in figure 2.36.

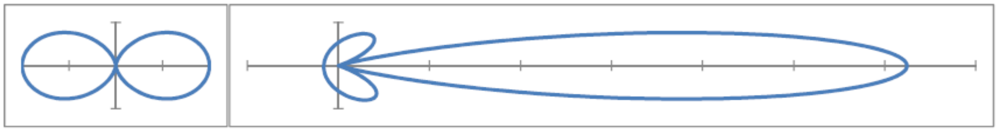


Figure 2.36: Relativistic aberration. The left panel shows the polar diagram for synchrotron emission in the instantaneous rest frame of an electron travelling left to right; the acceleration vector is directed vertically. Right, the same diagram in the lab frame, for the case $\beta = 0.95$.

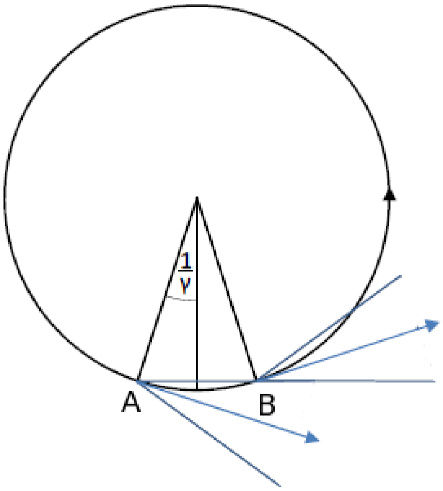


Figure 2.37: Beaming in synchrotron radiation. Because of the relativistic beaming effect, radiation from the electron is only visible to an observer at right of figure over an angle $2/\gamma$, i.e. a fraction $1/\pi\gamma$ of the electron's gyratory period.

As a consequence of beaming, the radiation from a relativistic electron is only visible for a fraction $1/\pi\gamma$ of its gyratory period. Furthermore, the *observed* duration of the pulse is much less than this, because the electron itself is moving at $v \simeq c$ and almost catches up with its own radiation.

If the distance from A to B in figure 2.37 is Δx, then the light from B is emitted $\Delta x/v$ later than the light from A, but has a distance Δx less far to travel to the observer. Therefore, the duration of the observed pulse is

$$\Delta t = \frac{\Delta x}{v} - \frac{\Delta x}{c} = \frac{\Delta x}{v}(1-\beta)$$

Now for relativistic electrons, $1-\beta^2 = (1+\beta)(1-\beta) \simeq 2(1-\beta)$, and therefore $1-\beta \simeq 1/(2\gamma^2)$. Also, $\Delta x/v = \Delta\theta/\omega_r = 2/(\gamma\omega_r)$, so the observed pulse duration is

$$\Delta t = \frac{2}{\gamma\omega_r}\frac{1}{2\gamma^2} = \frac{1}{\gamma^3\omega_r} = \frac{1}{\gamma^2\omega_g} \tag{2.43}$$

where $\omega_g = \gamma\omega_r$ is the angular cyclotron frequency. Allowing for the pitch angle of the electron with respect to the magnetic field, this becomes

$$\Delta t = \frac{1}{\gamma^2\omega_g\sin\theta} \tag{2.44}$$

Therefore, synchrotron radiation from a single relativistic electron is emitted as a series of short pulses. To obtain its spectrum, we would need to take the Fourier transform of this pulse train. This is tedious—the gory details are given in Longair[171] section 8.4—but we would expect the dominant Fourier mode to be given by

$$\nu_s \simeq \frac{1}{\Delta t} = \gamma^2\omega_g\sin\theta. \tag{2.45}$$

This is not far off: the full analysis gives

$$j(\omega) = \frac{\sqrt{3}e^3 B\sin\theta}{8\pi^2\epsilon_0 c m_e}F(x), \tag{2.46}$$

where

$$F(x) = x \int_x^\infty K_{5/3}(x)\,\mathrm{d}x, \tag{2.47}$$

$x = \nu/\nu_c$ where $\nu_c = \frac{3}{2}\gamma^2\nu_g \sin\theta$, and $K_{5/3}(x)$ is a modified Bessel function[178] of order 5/3. Note that the characteristic frequency ν_c is very similar to our estimate ν_s. This spectrum is quite sharply peaked close to ν_c (or ν_s), as shown in figure 2.38, so it is often adequate to assume that all the radiation is emitted at ν_c.

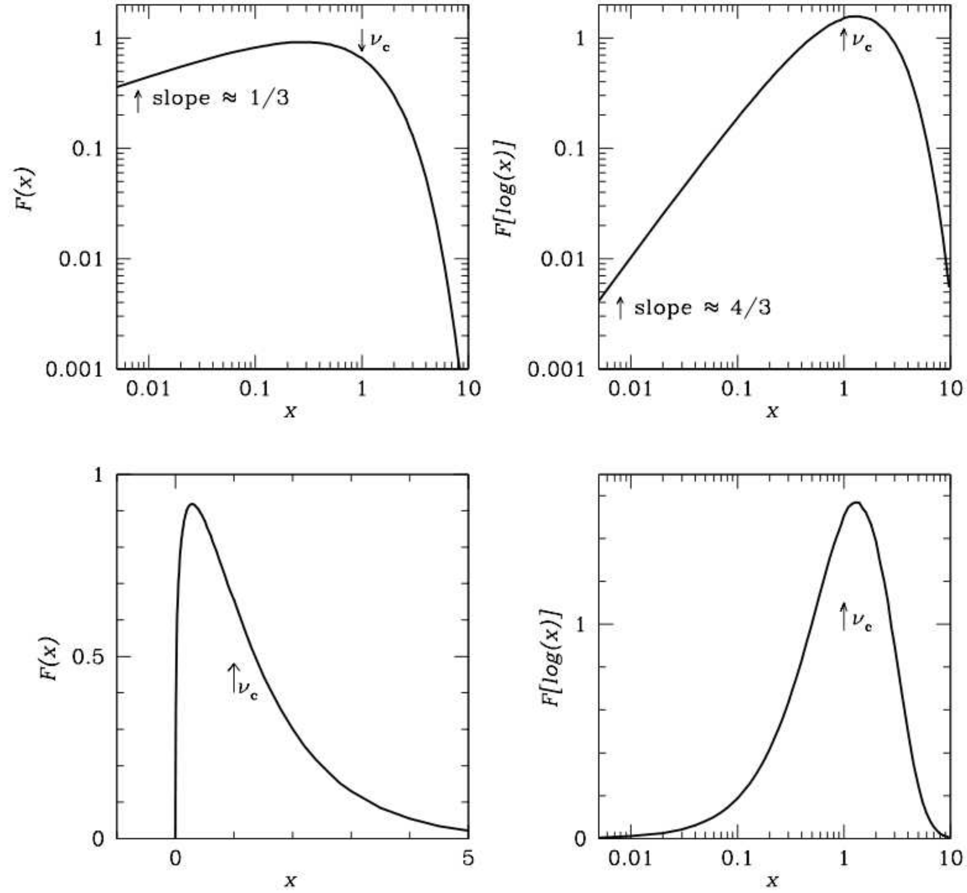


Figure 2.38: Spectrum of synchrotron radiation, in terms of $x = \nu/\nu_c$. This is shown in several different forms to bring out different features of the spectrum, but note that is is sharply peaked close to ν_c. Figure from Condon and Random[179].

As shown in figure 2.34, cosmic-ray electrons have a power-law spectrum, $N(E) \propto E^{-\delta}$ where $\delta \simeq 3$. If we assume that essentially all the synchrotron radiation from an electron of energy E is emitted at frequencies $\sim \gamma^2\nu_g$, then the spectral emissivity for synchrotron radiation is given by

$$j_\nu\,\mathrm{d}\nu = -\frac{\mathrm{d}E}{\mathrm{d}t} N(E)\,\mathrm{d}E$$

where $-\mathrm{d}E/\mathrm{d}t$ is given by equation (2.39), $E = m_e c^2\sqrt{\nu/\nu_g}$ and $\nu_g = eB/(2\pi m_e)$. Changing variables from E to ν, noting that $U_{\text{mag}} = B^2/2\mu_0$ and focusing only on the ν and B dependence of j_ν, we find

$$j_\nu \propto B^2\frac{\nu}{B}\left(\frac{\nu}{B}\right)^{-\delta/2}\nu^{-1/2}B^{-1/2},$$

since $\mathrm{d}E/\mathrm{d}\nu = m_e c^2/\sqrt{\nu\nu_g} \propto \nu^{-1/2}B^{-1/2}$. Collecting terms, we have

$$j_\nu \propto B^{(\delta+1)/2}\nu^{-(\delta-1)/2}, \tag{2.48}$$

i.e. the synchrotron radiation spectrum from cosmic-ray electrons should be a power law, $j_\nu \propto \nu^{-\alpha}$ with spectral index[3] $\alpha = \frac{1}{2}(\delta-1) \simeq 1$. The observed slope of the Galactic synchrotron radiation spectrum above a few MHz is about 0.7[180], which is in reasonable agreement with expectation.

Because of the beaming effect, the polarisation of synchrotron radiation is more complicated than than of cyclotron radiation. As can be seen in figure 2.39, the direction of polarisation rotates as the electron's velocity precesses about the direction of the magnetic field. If the velocity vector of the electron points directly along the line of sight, we should therefore see linearly polarised radiation. However, in fact we see radiation from electrons whose velocity vectors are within an angle $1/\gamma$ of the line of sight (see figure 2.37), and though this angle is small it is not zero. Velocity vectors which do not point directly towards us produce elliptically polarised light, with a small contribution parallel to the magnetic field direction as projected on the plane of the sky. (The polarisation is elliptical because the two components have a different time dependence, see Longair[171].) Considering *all* velocity vectors within $1/\gamma$ of the line of sight, we find that the components parallel to $\mathbf{B}$ cancel, having opposite senses on opposite sides of the velocity cone shown in figure 2.39, and therefore we actually observe linear polarisation, but not complete linear polarisation as expected from electrons pointing directly towards us.

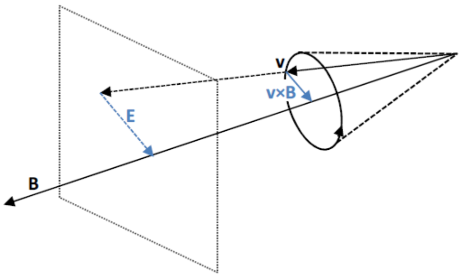


Figure 2.39: Polarisation in synchrotron radiation. As the electron velocity precesses around the magnetic field line, the direction of $\mathbf{v}\times\mathbf{B}$ changes, and therefore so does the polarisation of the radiation.

The result of the full calculation of synchrotron radiation (see Longair[171], section 8.4) is

$$j_\perp(\omega) = \frac{\sqrt{3}e^3 B\sin\theta}{16\pi^2\epsilon_0 c m_e}[F(x)+G(x)]; \tag{2.49}$$

$$j_\parallel(\omega) = \frac{\sqrt{3}e^3 B\sin\theta}{16\pi^2\epsilon_0 c m_e}[F(x)-G(x)], \tag{2.50}$$

where $j_\perp$ and $j_\parallel$ are the emissivities for polarisation perpendicular to and parallel to the magnetic field, respectively, $x = \nu/\nu_c$ as above, $F(x)$ is defined in equation (2.47) and $G(x) = xK_{2/3}(x)$. The overall ratio between perpendicular and parallel polarisation for a single electron is therefore

$$\frac{j_\perp}{j_\parallel} = \frac{\int_0^\infty [F(x)+G(x)]\,\mathrm{d}x}{\int_0^\infty [F(x)-G(x)]\,\mathrm{d}x} = \frac{\Gamma(\frac{7}{3})\Gamma(\frac{2}{3})+\Gamma(\frac{4}{3})\Gamma(\frac{2}{3})}{\Gamma(\frac{7}{3})\Gamma(\frac{2}{3})-\Gamma(\frac{4}{3})\Gamma(\frac{2}{3})},$$

[3] Warning: in defining spectral indices, some people explicitly insert the minus sign into the power law, so that a spectral index of 1 indicates $f(\nu) \propto \nu^{-1}$, while others include it in the definition of the index, so that $f(\nu) \propto \nu^{-1}$ would have spectral index -1. There is unfortunately no consensus in the literature. You just have to read the definitions in the paper or book you are reading.

where the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$ is a generalisation of the factorial $z!$ to real and complex numbers[181]; for integers, $\Gamma(n) = (n-1)!$. As one would expect from this, $\Gamma(n+1) = n\Gamma(n)$, so the above reduces to

$$\frac{j_\perp}{j_\parallel} = \frac{\frac{4}{3}+1}{\frac{4}{3}-1} = 7. \tag{2.51}$$

The fractional polarisation at angular frequency ω for a single electron is

$$\Pi(\omega) = \frac{j_\perp(\omega) - j_\parallel(\omega)}{j_\perp(\omega) + j_\parallel(\omega)} = \frac{G(x)}{F(x)}. \tag{2.52}$$

In order to calculate the polarisation expected from a given population of electrons, we need to integrate this over x, weighted appropriately by the electron spectrum. For a power law spectrum $N(E) \propto E^{-\delta}$, the result is[171]

$$\Pi = \frac{\delta+1}{\delta+\frac{7}{3}}, \tag{2.53}$$

which comes out to ~75% for $\delta \sim 3$. We therefore expect observed synchrotron radiation to be quite strongly polarised.

If synchrotron radiation is the main mechanism by which the electrons in the source lose energy, an electron of energy E will have a lifetime $\tau_s \sim E/(-\mathrm{d}E/\mathrm{d}t)$. This means that a typical synchrotron source will have a high-energy cutoff corresponding to the lifetime of its higher-energy electrons. The form of the cutoff depends on the nature of the source: if new electrons are continuously injected for some time t_0, the electron spectrum will be $\propto E^{-\delta}$ for energies such that $\tau_s > t_0$. For electrons with with energies such that $\tau_s < t_0$, the effective injection time is only τ_s—electrons injected earlier have already radiated away their energy. Since $\mathrm{d}E/\mathrm{d}t \propto E^2$, this means that the number of electrons is decreased by a factor $1/E$, and hence the power-law spectrum steepens to $\propto E^{-(\delta+1)}$. If, on the other hand, the source had an initial injection of electrons from some transient event but no subsequent supply, there will simply be a cutoff at $\tau_s = t - t_{\mathrm{inj}}$, where t is the current time and t_{inj} is the time at which the electrons were injected.

2.3.6 Self absorption

From the study of spectral lines, we are familiar with the idea that anything that produces an emission line at some wavelength can also produce the corresponding absorption line. This principle also holds for continuous distributions: both bremsstrahlung and synchrotron radiation can be *self-absorbed* if the emitting source is sufficiently dense. The key physics is the **principle of detailed balance**: in thermal equilibrium, the rates of forward and reverse reactions are equal, so emission of radiation by some physical process is balanced by absorption of radiation by the same physical process. The *absorption coefficient* χ_ν is the fractional decrease in intensity when traversing unit distance of the medium, so the loss in intensity I_ν is given by

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}z} = -\chi_\nu I_\nu,$$

where z is distance (avoiding x because of its use for ν/ν_c above!). In thermal equilibrium we must therefore have

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}z} = 0 = -\chi_\nu I_\nu + \frac{j_\nu}{4\pi} \tag{2.54}$$

where j_ν is the emissivity, and the factor of 4π comes from the fact that I_ν is defined per steradian and j_ν isn't, and in thermal equilibrium I_ν is given by the blackbody distribution

$$I_\nu = \frac{2h\nu^3}{c^2}\left(\exp\left(\frac{h\nu}{kT}\right) - 1\right)^{-1}.$$

Thermal bremsstrahlung absorption

For thermal bremsstrahlung, we substitute in j_ν from equation (2.31) and obtain

$$\chi_\nu \propto \frac{n_i n_e T^{-1/2}}{\nu^3} g(\nu, T)\left(1 - \exp\left(-\frac{h\nu}{kT}\right)\right). \tag{2.55}$$

At low frequencies, $h\nu \ll kT$, which are appropriate for radio emission and are also those where self-absorption is most likely to be important, we can expand the exponential to get

$$\chi_\nu \propto \frac{n_i n_e T^{-3/2}}{\nu^2} g(\nu, T), \tag{2.56}$$

where the Gaunt factor $g(\nu, T)$ varies only slowly and is of order unity.

The optical depth of the emitting medium is therefore

$$\tau = \int \chi_\nu \, \mathrm{d}z \propto \int n_e n_i T^{-3/2} n^{-2} \mathrm{d}z, \tag{2.57}$$

and from equation (2.54) we have, separating variables,

$$\int_0^{I_\nu} \frac{\mathrm{d}I_\nu}{j_\nu/(4\pi) - \chi_n u I_\nu} = \int_0^z \mathrm{d}z,$$

assuming $I_\nu = 0$ at $z = 0$. This gives

$$I_\nu = \frac{j_\nu}{4\pi\chi_\nu}\left(1 - e^{-\chi_\nu z}\right).$$

In a medium which is optically thick ($\chi_n u z \gg 1$) at low frequencies ($h\nu \ll kT$), this gives

$$I_\nu = \frac{j_\nu}{4\pi\chi_\nu} = \frac{2kT}{c^2}\nu^2, \tag{2.58}$$

i.e. for optically thick media at low frequencies the flat bremsstrahlung spectrum is replaced by the Rayleigh-Jeans tail of a blackbody spectrum (as one would expect for thermal radiation from an optically thick source).

Synchrotron self-absorption

Since synchrotron radiation is not thermal, its effective temperature varies with frequency. If a source has the same physical size at all frequencies, its *brightness temperature* is defined by

$$T_b = \frac{\lambda^2}{2k}\frac{S_\nu}{\Omega}, \tag{2.59}$$

where S_ν is the flux from the source at frequency $\nu = c/\lambda$ and Ω is the solid angle it subtends at the observer. This expression is obtained by equating S_ν to the blackbody flux (in the Rayleigh-Jeans limit, since brightness temperature is a concept most widely used in radio astronomy).

The effective temperature of an electron of energy E is given by

$$\gamma m_e c^2 = 3kT_e,$$

(this result is derived from the thermodynamics of a relativistic gas). Assuming that this electron radiates at frequencies $\sim \gamma^2 \nu_g$, so that $\gamma \simeq \sqrt{\nu/\nu_g}$, we have

$$T_e \simeq \frac{m_e c^2}{3k} \frac{\nu^{1/2}}{\nu_g^{1/2}}.$$

Equating this to T_b gives

$$S_\nu = \frac{2kT_e \nu^2}{c^2}\Omega = \frac{2m_e}{3\nu_g^{1/2}} \nu^{5/2}\Omega. \tag{2.60}$$

Therefore, we expect that a synchrotron-emitting source where the synchrotron radiation is generated by a population of electrons with a power-law spectrum, $N(E) \propto E^{-\delta}$, will have $S_\nu \propto \nu^{5/2}$ up to the frequency at which S_ν drops below the blackbody flux for the same effective temperature, and thereafter will have $S_\nu \propto \nu^{-(\delta-1)/2}$ as discussed earlier.

2.3.7 Summary

To summarise, the main radio emission mechanisms relevant to particle astrophysics are bremsstrahlung and, more importantly, synchrotron radiation.

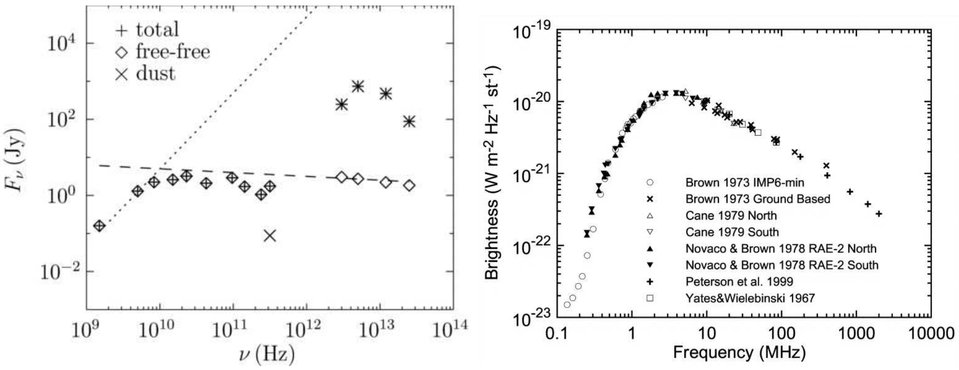


Figure 2.40: Examples of bremsstrahlung and synchrotron radiation. Left panel, radio/IR emission from a compact HII region[182]. The dotted line shows the +2 slope expected for optically thick thermal bremsstrahlung; the dashed line is the expectation for optically thin bremsstrahlung, which is roughly flat (the gradient is actually −0.1). The bump in the infrared region is thermal emission from dust. Right panel, synchrotron emission from the Milky Way[180], showing the $+\frac{5}{2}$ slope for synchrotron self-absorption up to a few MHz and a power-law slope with spectral index ∼0.7 above 10 MHz.

The flux S_ν from a thermal bremsstrahlung radio source is expected to be $\propto \nu^2$ for frequencies for which the source is optically thick, then approximately flat up to $\nu \sim kT/h$, after which it should cut off exponentially. Bremsstrahlung radiation is expected to be unpolarised.

For a source powered by synchrotron radiation, the flux should be $\propto \nu^{5/2}$ for frequencies where the source is self-absorbed, and then follow a power law $\propto \nu^{-\alpha}$ for higher frequencies, where the spectral index $\alpha \sim 1$ is determined by the spectral index of the parent population of electrons, $\alpha = \frac{1}{2}(\delta - 1)$. Finally, there will be a high-energy cutoff related to the lifetime of electrons in the source. Synchrotron radiation should be linearly polarised.

Examples of radio/IR emission displaying these spectral features are shown in figure 2.40.

2.4 High energy photons

2.4.1 High energy photons and particle astrophysics

High energy photons, defined here as photons with energies about $\sim$100 eV (i.e. X-rays and γ-rays) relate to particle astrophysics in a number of ways. First, the highest energy (TeV) γ-ray are regarded as "astroparticles" in their own right, in a similar way to high-energy neutrinos. Secondly, high-energy photons are often produced by mechanisms which require the existence of relativistic particles, usually but not always electrons; in this respect, they have more in common with radio emission than with the intermediate near IR/visual/UV wavelengths which are dominated by blackbody emission. Thirdly, although X-rays can be detected using focusing optics, albeit with a somewhat unconventional geometry (see below), γ-rays cannot: γ-ray telescopes involve the use of particle physics hardware, and thus fall under the general heading of experimental particle astrophysics.

In this section we shall separate high-energy photons into X-rays, low to intermediate energy γ-rays, and high-energy γ-rays. For astrophysical purposes, this division is motivated primarily by detection techniques: focusing optics for X-rays, space-based direct detection for low to intermediate energy γ-rays, and ground-based air shower detectors for high-energy γ-rays.

2.4.2 Mechanisms of high-energy photon emission

Both bremsstrahlung (see section 2.3.4) and synchrotron radiation (see section 2.3.5) produce high-energy photons as well as radio emission. Thermal bremsstrahlung or free-free emission is particularly important in the case of X-rays: many X-ray sources are essentially thermal emission from extremely hot gas. This is not, strictly speaking, the domain of particle astrophysics, although the X-ray emission from clusters of galaxies has provided important indirect evidence for dark matter, as discussed in more detail below.

Additional non-thermal mechanisms involved in the emission of high-energy photons are *inverse Compton scattering* and *π^0 decay*. Both of these are capable of producing very high-energy γ-rays such as are observed by imaging air Cherenkov telescopes. Inverse Compton scattering implies the presence of relativistic electrons in the source, but does not require the presence of fast hadrons, whereas in contrast pion decay is evidence for the presence of high-energy hadrons, since pions are produced when fast protons interact with ambient material. Pion decay products are therefore important signatures of potential sources of cosmic rays.

Inverse Compton scattering

Inverse Compton scattering is the scattering of a low-energy photon off a high-energy electron (as opposed to ordinary Compton scattering, where a high-energy photon—usually an X-ray—scatters off a low-energy electron). Assuming that the energy $\hbar\omega$ of the photon is such that $\hbar\omega' \ll m_e c^2$, where the primed frame is the centre of momentum frame, this can be treated as Thomson scattering in the rest frame of the electron (which, in this limit, is effectively identical to the centre of momentum frame).

Thomson scattering, the scattering of a beam of radiation by a stationary electron, is an application of Larmor's formula, equation (2.15), or more precisely the version before integrating over angle, equation (2.14). Following

Longair[171] section 9.2, we consider a beam of photons travelling in the z direction, and scattered through an angle α in the xz plane. If the incoming radiation is unpolarised, it will generate oscillating electric fields $E_{x,y} = E_{(x,y)0}e^{i\omega t}$ in the $x\ y$ directions, causing acceleration $\ddot{r}_{x,y} = eE_{x,y}/m_e$. Averaging over time gives $\langle E_{x,y}\rangle = \frac{1}{2}E_{(x,y)0}$, where for unpolarised radiation $E_{x0} = E_{y0} = E_0$. The corresponding power per unit area is given by the Poynting vector as defined in equation (2.13); in this case $S_x = S_y = \frac{1}{4}c\epsilon_0 E_0^2$.

In the x-direction, the angle θ between the acceleration vector and the radiation is given by $\theta = \frac{\pi}{2} - \alpha$; in the y-direction, the radiation is emitted perpendicular to the acceleration vector, so $\sin\theta = 1$. Adding the two components and substituting into equation (2.14) gives

$$P(\theta)\mathrm{d}\Omega = \frac{e^2}{16\pi^2\epsilon_0 c^3}\left(1+\cos^2\alpha\right)\frac{e^2E_0^2}{4m_e^2}\mathrm{d}\Omega = \frac{e^4}{16\pi^2\epsilon_0^2 m_e^2 c^4}\left(1+\cos^2\alpha\right)\frac{S}{2}\mathrm{d}\Omega, \tag{2.61}$$

where $S = S_x + S_y$.

This can be expressed in terms of a differential cross-section $\mathrm{d}\sigma_\mathrm{T}/\mathrm{d}\Omega$ defined such that

$$\frac{\mathrm{d}\sigma_\mathrm{T}}{\mathrm{d}\Omega} = \frac{\text{power radiated per unit solid angle}}{\text{incident power per unit area}}.$$

The incident power per unit area is S, so we have

$$\mathrm{d}\sigma_\mathrm{T} = \frac{e^4}{16\pi^2\epsilon_0^2 m_e^2 c^4}\frac{1}{2}\left(1+\cos^2\alpha\right)\mathrm{d}\Omega.$$

Integrating over solid angle gives

$$\sigma_\mathrm{T} = \frac{e^4}{16\pi^2\epsilon_0^2 m_e^2 c^4}\,\pi\int_{-1}^{+1}\left(1+\cos^2\alpha\right)\mathrm{d}(\cos\alpha) = \frac{e^4}{6\pi\epsilon_0^2 m_e^2 c^4},$$

which is the expression for the Thomson cross-section that we introduced in equation (2.38).

The total power in scattered radiation is therefore given by

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right) = c\sigma_\mathrm{T}U_\mathrm{rad}, \tag{2.62}$$

where the energy density in radiation, U_rad, is equal to S/c. In terms of photons, the energy density contributed by photons of frequency ν is simply $n_\nu h\nu$, where n_ν is the number density of such photons.

To apply this to inverse Compton scattering, we start be assuming that the Thomson scattering takes place in the rest frame of the high-energy electron, which we shall define as the primed frame (the unprimed frame is the lab frame). The energies of the photon in the primed and unprimed frames are related by

$$\hbar\omega' = \gamma\hbar\omega(1+\beta\cos\theta), \tag{2.63}$$

where $\beta = v/c$ is the velocity of the electron in units of c and θ is the angle between the velocity vector of the electron and the incoming photon, measured in the lab frame. The angle of incidence in the primed frame is given by

$$\cos\theta' = \frac{\cos\theta+\beta}{1+\beta\cos\theta},$$

as in equation (2.42).

We saw earlier that $\mathrm{d}E/\mathrm{d}t$ is a relativistic invariant. Therefore, if we can calculate the electromagnetic energy density in the primed frame, U'_{rad}, and apply equation (2.62), we can determine the energy transferred from the electron to the photons.

Given $U_{\mathrm{rad}} = n_\nu h\nu$, we need to transform both the frequency of the photons, as already described in equation (2.63), and their rate of incidence on the electron. By considering their coordinates in the primed and unprimed frames, it is easy to show that the time difference Δt between the arrival of successive photons transforms as

$$\Delta t = \gamma \Delta t'(1 + \beta \cos\theta). \tag{2.64}$$

This is exactly what we would expect from equation (2.63), given that frequency is inversely proportional to time.

The net result of this is that the number density of photons increases by a factor of $\gamma(1 + \beta\cos\theta)$ in the primed frame, and so does each photon's energy. Therefore

$$U'_{\mathrm{rad}} = U_{\mathrm{rad}}\gamma^2(1 + \beta\cos\theta)^2. \tag{2.65}$$

Assuming that the ambient radiation field is isotropic, we should average this over solid angle. The element of solid angle corresponding to incident angle θ is $2\pi\mathrm{d}(\cos\theta)$, so, normalising to 4π total, we have a probability $p(\theta)\mathrm{d}(\cos\theta) = \frac{1}{2}\mathrm{d}(\cos\theta)$ and hence

$$\begin{aligned} U'_{\mathrm{rad}} &= U_{\mathrm{rad}}\gamma^2 \int_{-1}^{+1} \tfrac{1}{2}(1 + \beta\cos\theta)^2 \mathrm{d}(\cos\theta) \\ &= U_{\mathrm{rad}}\gamma^2\left(1 + \tfrac{1}{3}\beta^2\right) = \tfrac{4}{3}U_{\mathrm{rad}}\left(\gamma^2 - \tfrac{1}{4}\right). \end{aligned} \tag{2.66}$$

The result of the scattering process is therefore that the ambient radiation field loses the original energy of the low-energy photons that interact, $c\sigma_{\mathrm{T}}U_{\mathrm{rad}}$, and gains the energy of the scattered photons, $\frac{4}{3}c\sigma_{\mathrm{T}}U_{\mathrm{rad}}\left(\gamma^2 - \frac{1}{4}\right)$, for a net change of

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \tfrac{4}{3}c\sigma_{\mathrm{T}}U_{\mathrm{rad}}(\gamma^2 - 1) = \tfrac{4}{3}c\sigma_{\mathrm{T}}U_{\mathrm{rad}}\beta^2\gamma^2, \tag{2.67}$$

using the identity $\gamma^2 - 1 = \beta^2\gamma^2$. Note that this is identical in form to equation (2.40), the only difference being that U_{mag} is replaced by U_{rad}. This reflects the fact that the underlying physics—a relativistic electron interacting with an ambient electromagnetic field—is the same in both cases.

The maximum energy gain for a photon of initial energy $\hbar\omega_0$ corresponds to a head-on collision with the electron. Applying equation (2.63) twice (once to transform the incoming photon to the primed frame, and once to transform the outgoing photon, which is travelling in the opposite direction, back to the lab frame) we get

$$(\hbar\omega)_{\max} = \gamma^2(1 + \beta^2)^2\hbar\omega_0 \simeq 4\gamma^2\hbar\omega_0. \tag{2.68}$$

Also, since the number of photons scattered per unit time is given by $c\sigma_{\mathrm{T}}U_{\mathrm{rad}}/\hbar\omega_0$, comparing this with equation (2.67) shows that the *average* energy of the scattered photons is

$$\langle\hbar\omega\rangle = \tfrac{4}{3}\beta^2\gamma^2\hbar\omega_0 \simeq \tfrac{4}{3}\gamma^2\hbar\omega_0. \tag{2.69}$$

This confirms that the spectrum of inverse-Compton upscattered photons is sharply peaked, as shown in figure 2.41. As electrons with γ factors of > 1000

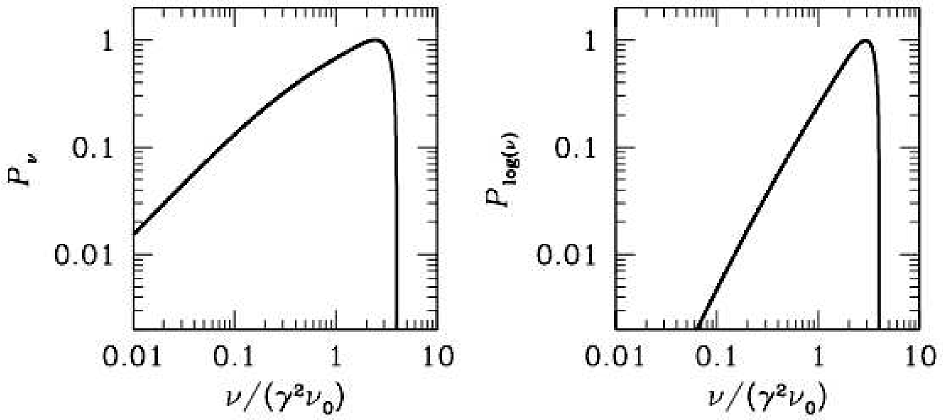


Figure 2.41: Spectrum of photons from inverse Compton scattering of an electron of energy $\gamma m_e c^2$, expressed in terms of ν/ν_0, per unit frequency (left) and log frequency (right). Note the sharp peak below the cutoff at $4\gamma^2$. Figure from Condon and Random[179], section 5E.

are known to exist in our Galaxy (see figure 2.34), inverse Compton scattering is an efficient way to create extremely high-energy photons.

This is the inverse Compton spectrum from a single electron. For a power law distribution of electrons, $N(E) \propto E^{-\delta}$, as with synchrotron radiation (see equation (2.48), the resulting photon spectrum is a power law, $j_\nu \propto \nu^{-\alpha}$ where $\alpha = \frac{1}{2}(\delta - 1)$. The source of photons can be the optical output of the source, the cosmic microwave background or the synchrotron emission from the same population of relativistic electrons (synchrotron-self-Compton or *synchro-Compton radiation*). As the CMB is always present, no population of relativistic electrons can avoid energy losses by inverse Compton scattering. This puts an upper limit on the lifetime of relativistic electrons of

$$\tau_{\rm IC} = \frac{E}{{\rm d}E/{\rm d}t} = \frac{E}{\frac{4}{3}\sigma_{\rm T} c U_{\rm CMB}}, \tag{2.70}$$

where $U_{\rm CMB} = 4\sigma T^4/c = 2.6 \times 10^5$ eV m^{-3}. Putting in all the numbers, we find that

$$\tau_{\rm IC} = \frac{2.3 \times 10^{12}}{\gamma} \text{ years.}$$

This implies that the 100 GeV electrons seen in figure 2.34 have lifetimes of at most 10^7 years.

Pion decay

Neutral pions produced when high-energy protons collide with ambient material will decay into two photons, each with energy $\frac{1}{2}m_{\pi^0}c^2$ in the pion rest frame. In the lab frame, the energies will transform according to

$$E = \gamma E'(1 - \beta\cos\theta'),$$

where θ' is the angle between the outgoing photon and the pion velocity vector, as measured in the pion rest frame, and β is the pion velocity in units of c. As the π^0 has zero spin, its decay is isotropic in its rest frame, so the energy distribution of the produced photons is flat between

$$E_{\gamma,\min} = \tfrac{1}{2}m_{\pi^0}c^2\gamma(1-\beta) \text{ and } E_{\gamma,\max} = \tfrac{1}{2}m_{\pi^0}c^2\gamma(1+\beta); \tag{2.71}$$

an ultrarelativistic π^0 ($\beta \simeq 1$) will produce photons with all energies up to E_{π^0}.

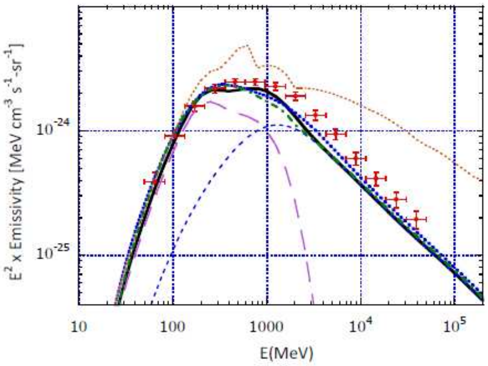


Figure 2.42: Calculation of $p + p \rightarrow \pi^0 \rightarrow \gamma\gamma$, compared with data from *Fermi*–LAT. The solid line, blue dotted line and green dot-dashed line are model calculations; the red dotted line is an isobaric model prediction, which fails to describe the data, and the purple long-dashed line and light-blue short-dashed line are components of the model described by the heavy solid line. Figure from Dermer et al.[183], where a more detailed description of the models can be found. *Fermi*–LAT data from [184].

Assuming that the π^0 is produced by a cosmic-ray proton of energy E_p hitting a stationary hydrogen nucleus, $p + p \rightarrow p + p + \pi^0$, the centre-of-mass energy $\sqrt{s}$ of the collision is given by (in particle physics units, in which $c = 1$)

$$s = E_{\rm tot}^2 - p_{\rm tot}^2 = 2m_p^2(\gamma_p + 1), \quad (2.72)$$

where γ_p is the Lorentz factor of the cosmic-ray proton; note that its energy $E_p = \gamma_p m_p$ and its momentum $p_p = m_p\sqrt{\gamma_p^2 - 1}$. The minimum centre-of-mass energy to produce a π^0 is $\sqrt{s} = 2m_p + m_{\pi^0}$, corresponding to a collision in which all three final-state particles are stationary in the c.o.m. frame. Substituting this into equation (2.72), we find

$$4m_p^2\left(1 + \frac{m_{\pi^0}}{m_p} + \frac{m_{\pi^0}^2}{4m_p^2}\right) = 2m_p^2(\gamma_p + 1),$$

which gives

$$E_{\rm thr} = m_p + 2m_{\pi^0} + \frac{m_{\pi^0}^2}{2m_p}; \quad (2.73)$$

the minimum kinetic energy ($= E_{\rm thr} - m_p$) required for π^0 production is 280 MeV, a little over twice the mass of the π^0 itself (135 MeV/c^2).

The cross-section for inclusive π^0 production, shown in figure 2.43[185], is complicated by a number of factors: at proton energies up to ~10 GeV, the reaction may go via resonances such as the $\Delta(1232)$ ($p + p \rightarrow p + \Delta^+ \rightarrow p + p + \pi^0$); the π^0 may not be produced directly but by way of heavier mesons or baryons; at higher energies, more than one π^0 may be produced in a single pp collision. These issues are discussed in detail by Dermer et al.[183], who provide the empirical fit

$$\sigma_{\pi^0 X}(\text{mb}) = 32\ln p_p + \frac{48.5}{\sqrt{p_p}} - 59.5 \quad (2.74)$$

for incoming proton momenta in the range $8 < p_p < 1000$ GeV/c.

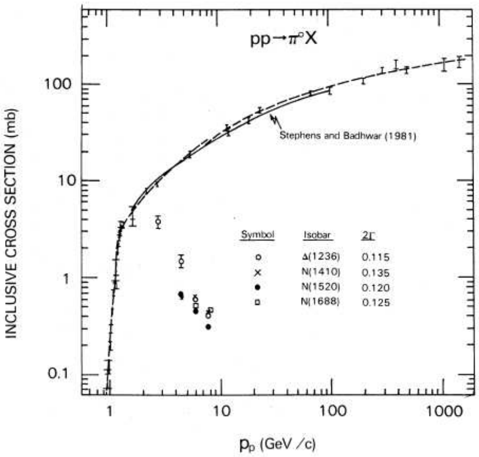


Figure 2.43: Cross-section for π^0 production in pp collisions, $p + p \rightarrow \pi^0 + X$, as a function of proton momentum p_p. From Dermer (1986)[185].

The diffuse γ-ray emission from the Galaxy above 1 GeV is believed to be produced almost entirely by π^0 decays, wih unresolved point sources contributing only 5–10%[183]. The model proton spectrum derived from the observed γ-ray spectrum is consistent at energies above ∼10 GeV with direct measurements of the cosmic-ray proton spectrum from PAMELA and AMS[183]; at lower energies, the observed proton spectrum is affected by solar modulation as discussed earlier.

The characteristic spectrum of π^0 decay photons in astrophysical sources is strong evidence for the presence in such sources of relativistic hadrons, since pions are far less likely to be produced in leptonic interactions. We should note that any environment that produces π^0 decay photons should also produce high-energy neutrinos, since the threshold for the corresponding reaction for charged pion production, $p + p \to p + n + \pi^+$, is only marginally higher, at a proton kinetic energy of 292 MeV (the difference arises because the neutron mass, 939.6 MeV/c^2, is slightly higher than the proton mass, and the $\pi^\pm$ mass, 139.6 MeV/c^2, is slightly higher than the π^0 mass). The π^+ decays almost exclusively into $\mu^+ + \nu_\mu$, but neutrino oscillations will result in an equal mix of neutrino flavours over astronomical distances. However, the weak interaction cross-sections characteristic of neutrino interactions are so much smaller than photon interaction cross-sections that we should expect to observe neutrinos only from the most intense γ-ray sources.

Summary

High-energy photons are secondary products of high-energy charged particles. They may be produced by thermal bremsstrahlung (free-free emission) in a hot plasma, or by interactions of non-thermal populations of relativistic electrons or hadrons. Multiwavelength observations soanning the range from X-rays to hard γ-rays are essential to define the spectrum of the radiation sufficiently to distinguish inverse Compton scattering (implying relativistic electrons) from π^0 decay (implying relativistic hadrons). Observations of high-energy neutrinos from a γ-ray source would also imply the presence of relativistic hadrons, but the expected rate of detection is extremely low because neutrino interactions are weak at almost all energies.

As a result, it is very important that observations of high-energy photons span a wide energy range. This requires multiple different technologies, some of them resembling conventional astronomical observations while others are more akin to particle physics experiments. This will be covered in the remainder of this section.

2.4.3 X-rays

The X-ray region of the electromagnetic spectrum runs from about 100 eV to 100 keV in energy (0.01 to 10 nm in wavelength), though different authors quote slightly different boundaries. As can be seen in figure 2.28, high-energy photons (UV and above) are completely absorbed by the atmosphere. X-rays and soft and intermediate-energy γ-rays are therefore observed using space-based platforms.

After some exploratory rocket and balloon missions in the 1960s, the first satellite specifically designed for X-ray astronomy, *Uhuru*[66], was launched in 1970. Many more missions followed[186]; significant highlights include *Einstein* (1978–81), the first fully imaging instrument, ROSAT (1990–99), the first imaging all-sky survey, and ASCA (1993–2000), the first imaging spectrometer. The principal current missions are *Chandra*[187], XMM–Newton[188] and Suzaku[189]; in addition, the γ-ray observatories INTEGRAL[190] and *Swift*[191] have X-ray telescopes on board to complement the γ-ray instruments.

If we define "conventional observational astronomy" as the detection of elec-

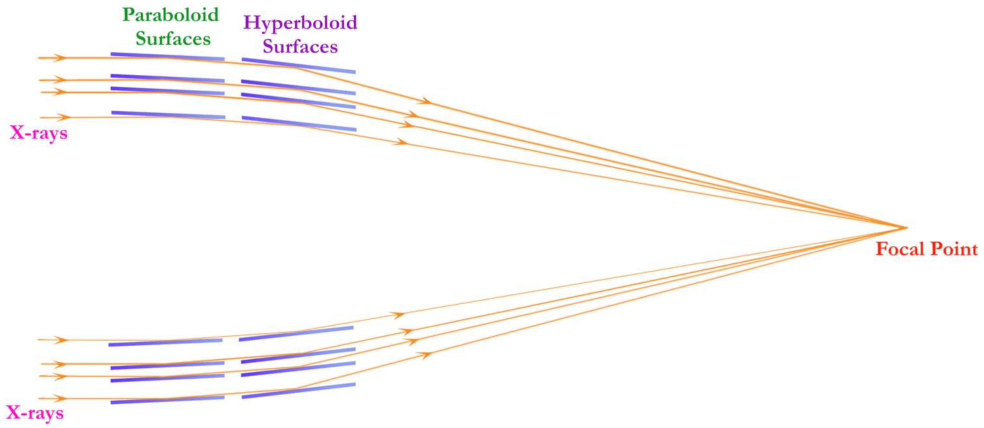


Figure 2.44: Light path for a typical X-ray telescope: nested paraboloid and hyperboloid mirrors focus X-rays on to a focal plane well behind the mirrors. This image is from *Chandra*[187], but XMM–Newton and Suzaku use essentially identical systems.

tromagnetic radiation from the source by means of focusing optics, then X-ray astronomy qualifies (whereas γ-ray astronomy does not). However, the details of the optics are unusual, because X-rays reflect only at *grazing incidence*. The net result of this is that focusing X-ray telescopes consist of a series of nested barrels which focus the incoming X-rays at a point far behind the mirrors, in contrast to optical or radio telescopes which consist of the central "cup" of the paraboloid and reflect back to a prime focus above the primary mirror. Figure 2.44 shows a schematic of the focusing system of NASA's *Chandra* X-ray telescope[187], and figure 2.45 is a schematic of ESA's XMM–Newton spacecraft[188]. The instruments lie several metres behind the primary mirror system (7.5 m for XMM–Newton, 10 m for Chandra).

The detectors in modern X-ray telescopes are usually silicon-based: both of XMM–Newton's principal instruments use CCDs, as do *Chandra*'s ACIS and Suzaku's XIS[189]. *Chandra*'s High Resolution Camera uses *microchannel plates*[192], which are similar in operation to miniature photomultiplier tubes: the incoming particle (which may be a charged particle or a high-energy photon) liberates an electron from the channel wall, and this secondary electron is then amplified into an avalanche by means of a carefully designed voltage gradient. *Chandra*'s instrumentation covers the energy range 0.07–10 keV and has a spatial resolution of order 1″, depending on the instrument used[193]; XMM–Newton has a slightly higher maximum energy (15 keV) and a somewhat worse angular resolution ($\sim 5''$)[194].

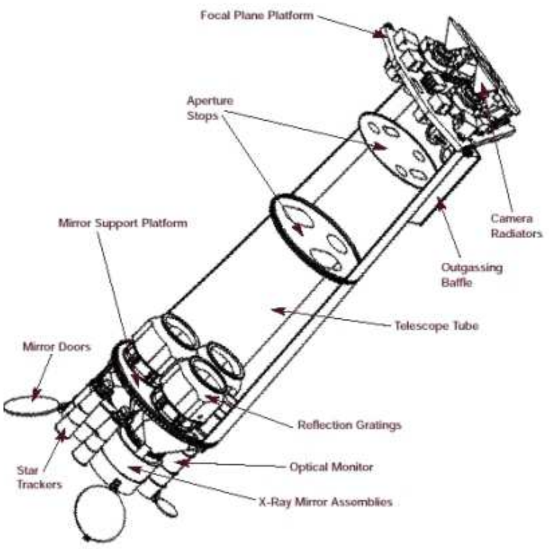


Figure 2.45: Schematic of the XMM–Newton spacecraft[188], which contains three separate mirror assemblies.

Suzaku's X-ray Imaging Spectrometer has a similar energy range (0.2–10 keV), but is supplemented by a Hard X-Ray Detector (HXD), which covers the hard X-ray–soft γ-ray energy range 10–700 keV. Grazing-incidence reflection does not work at these energies, so the HXD[195] is not an imaging device. Instead, it has *collimators*: the active detectors are located at the bottom of a well, so that the field of view is restricted to the direction in which the instrument is pointed: above 100 keV, the HXD field of view is $34' \times 34'$[189]. The HXD reads out hard X-rays (<60 keV) with silicon PIN diodes, and soft γ-rays (>30 keV) with a "phoswich" detector[196]. Phoswich detectors (a contraction of "*phos*phor sand*wich*") are combinations of different scintillators read out to a common photodetector: HXD uses gadolinium silicate and bismuth germanate. The different pulse shapes of the two scintillators allow their signals to be distinguished in the readout, giving greater dynamic range and/or background rejection than a single scintillator.

Astrophysical X-ray sources are numerous and varied[198]. Much X-ray emission is thermal bremsstrahlung or free-free emission from extremely hot plasma, either in an accretion disc or in the *intracluster medium* of rich clusters of galaxies. The latter is usually accompanied by spectral lines coming from K-shell transitions of heavy elements ("metals", according to the rather undiscriminating nomenclature of astrophysics), especially iron (which really is a metal)[197]; these lines from very highly ionised species confirm that the X-ray emission is thermal, i.e. that the temperature of the plasma is appropriate for X-ray emission.

Thermal bremsstrahlung does not in itself fall within the remit of particle astrophysics. However, X-ray free-free emission from rich clusters of galaxies has been a useful tool in determining the mass profile of the cluster, and hence providing evidence for dark matter. If we assume that the cluster is spherical and in hydrostatic equilibrium, we have

$$\frac{\mathrm{d}\Phi}{\mathrm{d}r} = \frac{GM_r}{r^2} = -\frac{1}{\rho_g}\frac{\mathrm{d}P_g}{\mathrm{d}r}, \tag{2.75}$$

where Φ is the gravitational potential of the cluster ($= -GM/r$ for a spherical cluster), M_r is the total mass contained within radius r, ρ_g is the gas density, and P_g is the gas pressure. Using the gas laws, $P_g = n_g kT_g$ where n_g is the number density of the gas, T_g is its temperature, and k is Boltzmann's constant, we have

$$M_r = \frac{r^2}{G}\frac{k}{\mu m_p n_g}\left(T_g\frac{\mathrm{d}n_g}{\mathrm{d}r} + n_g\frac{\mathrm{d}T_g}{\mathrm{d}r}\right),$$

which is usually expressed as

$$M_r = \frac{kT_g(r)r}{G\mu m_p}\left(\frac{\mathrm{d}\ln n_g}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln T_g}{\mathrm{d}\ln r}\right), \tag{2.76}$$

where μ is the mean particle mass in units of the proton mass m_p (for fully ionised gas $\mu = 0.61$), and by the chain rule $\mathrm{d}\ln y/\mathrm{d}\ln x = (x/y)\mathrm{d}y/\mathrm{d}x$.

The hot gas in clusters of galaxies is optically thin, so the number density can be determined from the X-ray luminosity using equation (2.31). Temperature and n_g can be disentangled by fitting the bremsstrahlung spectrum; it is generally found that clusters are fairly isothermal ($\mathrm{d}\ln T_g/\mathrm{d}\ln r \sim 0$ to -0.8, whereas $\mathrm{d}\ln n_g/\mathrm{d}\ln r \sim -2.0$ to -2.4)[199]. There are various methods of extracting the temperature and density profiles, each with its own set of advantages and disadvantages: working backwards from the data requires numerical calculations of derivatives, which can be subject to large errors (you are subtracting nearly equal numbers, which is a recipe for inflating errors); fitting a model generates smooth profiles, but introduces model dependence. Evidence from simulations suggests[199] that X-ray mass determinations tend to underestimate the true mass, sometimes substantially.

Mass determinations using this method were the first to demonstrate that (1) the mass in the hot gas exceeds that in the visible galaxies by around an order of magnitude, but (2) the total mass exceeds the gas mass by about a factor of 5. Thus, the hot intracluster medium is one of the main hiding places for "dark baryons" (those that make up the difference between the density parameter for stars, Ω_*, and the baryon density, Ω_b).

In more recent years, gravitational lensing has provided an alternative method of measuring cluster masses. Comparisons between lensing and X-ray mass distributions in galaxy cluster collisions (most famously the Bullet Cluster[200]) have provided evidence for the collisionless nature of dark matter, in that the intracluster medium is shocked and displaced by the collision whereas the massive cluster halos are not.

Non-thermal X-ray emission can arise from relativistic bremsstrahlung, synchrotron radiation, or inverse Compton scattering. Most of these are closely related to radio emission mechanisms, and in consequence there is a close relationship between X-ray and radio luminosity, as shown for X-ray binaries in figure 2.46. In this plot, the "inefficient branch" corresponds to an *Advection Dominated Accretion Flow* or ADAF[203], where the accretion rate is low and most of the accreted material falls into the black hole without being heated to the point where it emits X-rays, whereas the "efficient branch" corresponds to a *Luminous Hot Accretion Flow* or LHAF[204], where the accretion rate is higher and—as the name indicates—the accreted material is much hotter and hence has a much higher X-ray luminosity.

Accretion on to black holes is the fundamental power source for many high-energy astrophysical phenomena, particularly active galactic nuclei (AGN). Analogies between different classes of AGN and different classes of black-hole binaries are discussed by Feng and Narayan[204]. We shall return to this question when we discuss astrophysical sources below.

2.4.4 Soft γ-rays

Soft γ-rays, from 100 keV up to 10 MeV or so, present an observational challenge. They are too energetic to undergo even grazing-incidence reflection, so focusing optics will not work, but their energies are too low for e^+e^- pair production, so the particle physics tracking calorimeters used for harder γ-rays (see next section) will not work either. Collimators, as used in the Suzaku HXD, are a possibility, but provide poor angular resolution and necessarily limit field of view, making them an unsatisfactory solution for survey instruments. Another option, used by the COMPTEL instrument[32] on the Compton Gamma Ray Observatory (CGRO), is to make use of the kinematics of Compton scattering. COMPTEL contains two detector layers separated by 1.5 m: the incoming γ-ray scatters in the upper detector and is absorbed in the lower. The energy lost in the upper detector, combined with the measurement of the energy of the scattered photon in the lower detector, defines the scattering angle; this locates the direction of the incoming γ somewhere on a cone whose axis is defined by the direction of the scattered photon. Many γ-rays coming from a single point source can be identified by the fact that the "event circles" on the

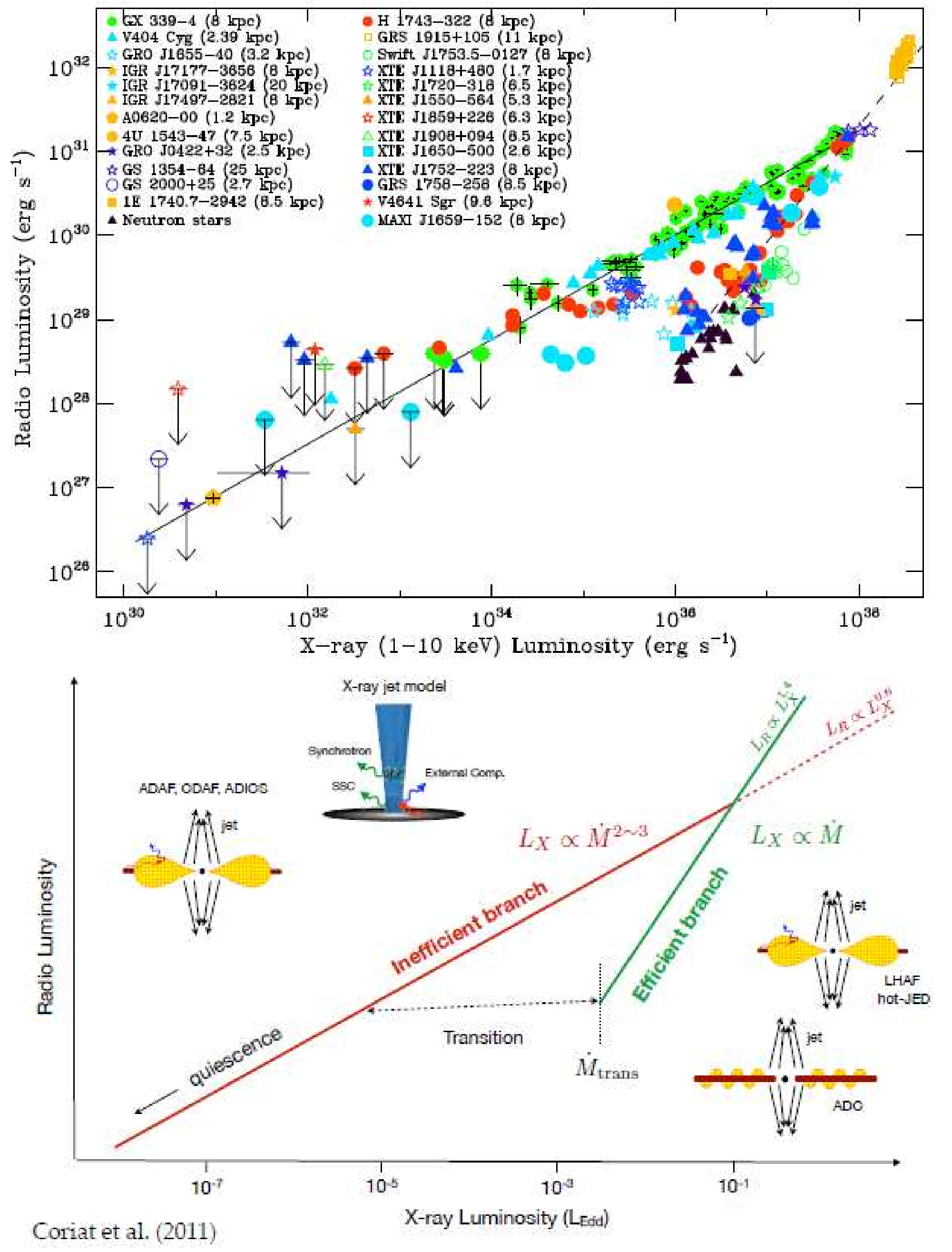


Figure 2.46: Correlation between X-ray and radio emission in X-ray binaries. Top, compilation of data from Corbel et al.[201]; bottom, schematic from Panessa[202].

sky corresponding to these cones will all intersect at a particular point, which is the location of the source. (This is, of course, an idealised description: in practice, finite detector resolution, multiple scatterings, leakage of energy out of the back of one or both detectors, etc., will make the reconstruction much more difficult.)

However, the technique employed by most currently operational soft γ-ray detectors, including the Burst Alert Telescope on *Swift*[191] and the IBIS imager and SPI spectrometer on INTEGRAL[190] is the coded mask aperture[205]. A coded mask is an array of opaque and transparent pixels which casts a shadow on the detector. The shape of the shadow depends on the direction of the incident radiation, as shown in figure 2.47. Therefore, in principle, the resulting image contains information about the directions of the incident photons. This information would be straightforward to decode if there were only one source in the field of view, but coded mask telescopes are designed to be wide-field instruments, so this is not normally the case.

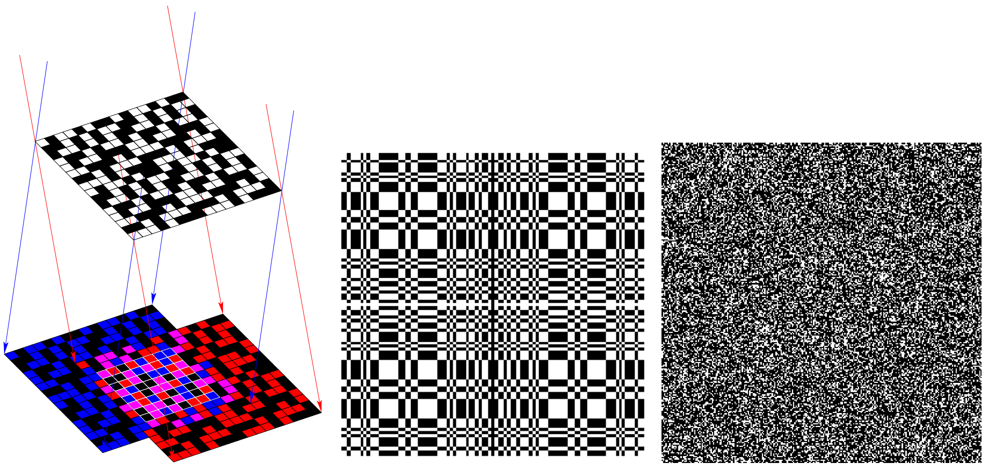


Figure 2.47: Coded mask imaging. The schematic on the left[206] shows the principle: the two stars cast different shadows, so the resulting image can be deconvolved to extract the original incident directions. Mask designs can be structured, as in the IBIS instrument[207] (centre) or pseudorandom, as in the Wide Field Camera (WFC) on Beppo-SAX[208] (right).

The image on the detector array can be described by

$$\mathbf{D} = \mathbf{O} * \mathbf{M} + \mathbf{B},$$

where $*$ is the convolution operator, $\mathbf{D}$ is an array representing the detector plane, $\mathbf{O}$ represents the object of the imaging, i.e. the sky in the field of view, $\mathbf{M}$ represents the mask and $\mathbf{B}$ represents background noise. The aim of the reconstruction is to recover $\mathbf{O}$ knowing $\mathbf{D}$ and $\mathbf{M}$.

There are various approaches to this[205]. There is an explicit deconvolution operation, but this is essentially inverting a matrix, which is always potentially hazardous when dealing with experimental data: in particular, the deconvolution is necessarily applied to $\mathbf{B}$ as well as $\mathbf{O} * \mathbf{M}$, which means that if some of the elements of the deconvolution matrix are large, the reconstructed image may be dominated by noise. A widely used alternative approach[209] is to define a *post-processing array* $\mathbf{G}$ such that convolving $\mathbf{D}$ with $\mathbf{G}$ recovers an approximation to $\mathbf{O}$:

$$\hat{\mathbf{O}} = \mathbf{D} * \mathbf{G} = (\mathbf{O} * \mathbf{M}) * \mathbf{G} + \mathbf{B} * \mathbf{G}.$$

The aim is to choose $\mathbf{G}$ such that $\mathbf{M} * \mathbf{G}$ is effectively the identity matrix and $\mathbf{B} * \mathbf{G}$ is approximately zero. In this case, the array $\mathbf{G}$ is not a formal inverse, but is empirically chosen so as to meet these conditions as closely as possible.

In addition to these analytical or semi-analytical methods, a range of maximisation techniques are also used, including maximum entropy[210] and maximum likelihood[211]. More complicated multivariate analysis tools such as neural networks have also been suggested[212].

All of these reconstruction techniques are fairly time-consuming and therefore are generally applied offline, after the data have been transmitted to Earth. In principle, multiple different reconstruction methods can therefore be applied to the same dataset, although in practice the team responsible for the instrument in question will supply a standard toolkit.

Coded mask telescopes do not provide angular resolution competitive with focusing optics, but they are significantly better than collimators and have the advantage of a wide field of view. The IBIS imager on board INTEGRAL[213], which is sensitive to hard X- and soft γ-rays between 15 keV and 10 MeV, has a field of view of $8.3^\circ \times 8.0^\circ$ fully coded, with an accuracy for point source detection that varies from $30''$ for a bright source at 100 keV to 5–$10'$ for a just-detectable source at 1 MeV. *Swift*–BAT[214], the Burst Alert Telescope on the *Swift* satellite, has a more restricted energy range of 15–150 keV, a field of view of $100^\circ \times 60^\circ$ half-coded, and an angular resolution of $17'$: it is designed for rapid identification of gamma-ray bursts (GRBs), with the secondary purpose of providing an all-sky survey in the hard X-ray waveband.

The "killer app" for soft γ-ray detection is gamma-ray bursts (GRBs). These exceptionally luminous transients[215], which compress into a few seconds more energy than the luminosity of the Sun over its entire lifetime, have a spectral energy distribution which peaks in the MeV region, at least in the initial burst; higher-energy, GeV–TeV, γ-rays are observed, but generally occur some seconds after the prompt burst and last for a much longer time, as shown in *Fermi*–LAT observations of GRB 090926A[216] (see figure 2.48).

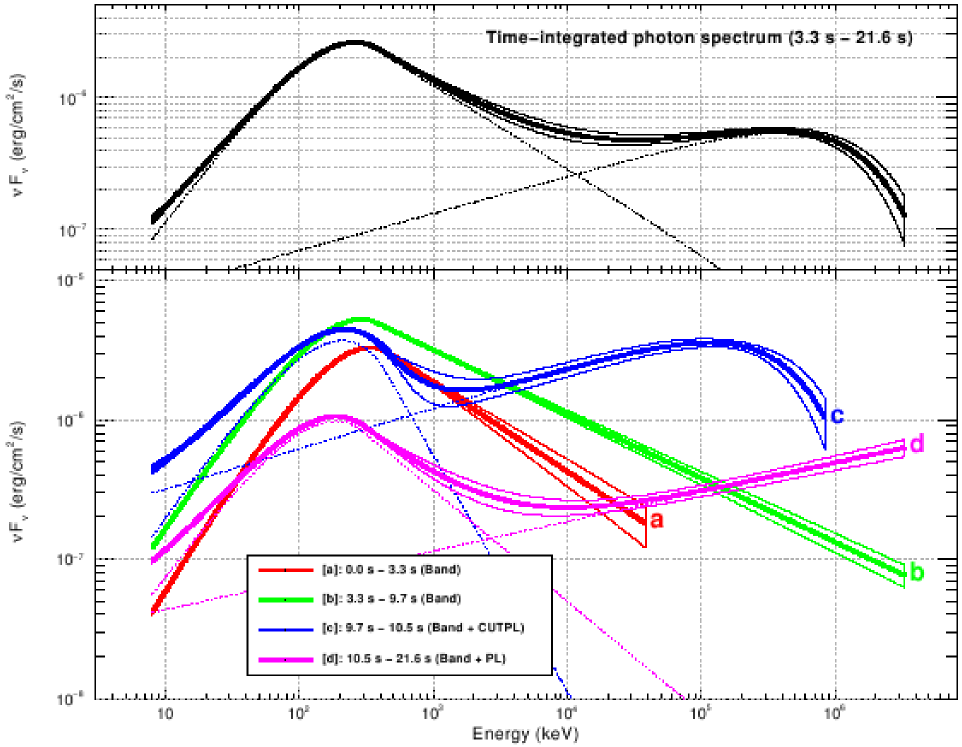


Figure 2.48: Energy spectrum of the bright GRB 090926A. Top, time-integrated spectrum showing two components, one soft and one hard; bottom, time evolution of the spectrum showing that the hard component is delayed relative to the soft component. Figure from [216].

GRB spectra are usually fitted using the **Band function**[217],

$$N_E(E) = \begin{cases} AE^\alpha \exp\left(-E/E_0\right) & E \leq (\alpha-\beta)E_0 \\ A\left[(\alpha-\beta)E_0\right]^{\alpha-\beta} E^\beta \exp(\beta-\alpha) & E \geq (\alpha-\beta)E_0, \end{cases} \tag{2.77}$$

where A, α, β and E_0 are fitted parameters; typically $\alpha \sim -1$, $\beta \sim -2$ and the spectral break E_0 is between 0.1 and 1 MeV. This is a purely empirical formula with no theoretical input or implications: in particular, Band et al.[217] stress that although E_0 is playing a role analogous to that of temperature in some thermal distributions, it is not to be interpreted as a physical temperature.

GRBs were first discovered accidentally, by military γ-ray satellites designed to search for evidence of clandestine nuclear tests[218, 219], but have since become the object of intense study. As the γ-ray signal itself does not provide any information about redshift, and the angular resolution of early instruments was not good enough to allow effective searches for optical counterparts, the nature of GRBs was at first extremely mysterious. Later instruments, particularly Beppo-SAX[208], were able to associate the γ-ray signal with a longer-lasting X-ray afterglow, which in turn, because of its greater positional accuracy, allowed the identification of optical afterglows and host galaxies. The *Swift* satellite[191], which is explicitly designed for GRB studies, is equipped with a γ-ray telescope (BAT), an X-ray telescope (XRT) and a UV/optical telescope (UVOT), all co-aligned so that the two focusing instruments can slew rapidly to point at the position of a burst as defined (to arcminute precision) by BAT. *Fermi* also has a gamma-ray burst detector, the GBM, and can follow up GBM detections with the higher-energy Large Area Telescope (LAT) discussed below. Multiwavelength studies of GRBs, covering all wavelengths from radio to TeV γ-rays, are now quite common, and neutrino telescopes have conducted (so far unsuccessful) searches for coincident neutrino bursts[220, 221], which are expected in some but not all GRB models.

The first extensive studies of GRB properties were carried out using the BATSE detector on CGRO[222]. Apart from demonstrating that the distribution of bursts on the sky was isotropic (and thus that they were not associated with the Galactic disc), BATSE's most important discovery was the distinction between "long" and "short" GRBs (see figure 2.49). The formal boundary between long and short is usually taken to be $T_{90} = 2$ s, where T_{90} is the burst duration excluding the first and last 5% of the total fluence, although there is evidence that the boundary depends on the energy band in which the GRB is studied[223].

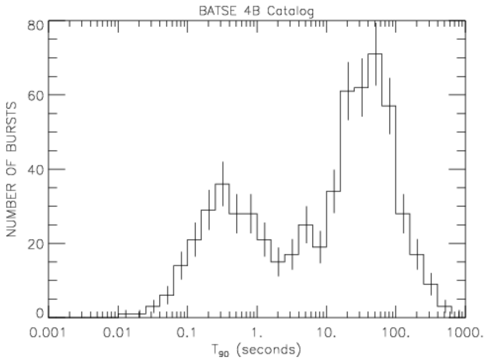


Figure 2.49: Durations of GRBs from the BATSE 4B GRB catalogue. The variable plotted is T_{90}, the time between 5% and 95% of the total fluence. Two populations corresponding to "long" and "short" bursts are clearly seen.

It turns out that the long and short bursts differ in more than just duration. Long bursts generally have a softer energy spectrum, are brighter, are found in galaxies with high star formation rate and in a few cases are unambiguously associated with ultraluminous Type Ib/c core-collapse supernovae[224]. Short bursts have a harder energy spectrum, are typically fainter, sometimes occur in elliptical galaxies and are suspected on theoretical grounds of being caused by mergers of compact objects (two neutron stars or a neutron star and a black hole)[223]. The T_{90} value is not a perfect discriminator: a few nearby "long" GRBs, such as GRB 060614[218] have no associated supernova and may be more closely related to "short" GRBs in terms of physics; some "short" GRBs have a long tail of "extended emission" which may imply that they differ physically from more typical short GRBs. Various alternative classification schemes have been proposed, but not generally accepted so far.

The observed properties of GRBs make it clear that the observed radiation must be beamed[219]. The detection of optical afterglows has allowed the host galaxies of many GRBs to be identified, permitting a determination of their redshifts and hence their distances. The isotropic luminosity inferred from this is of order 10^{52} erg s^{-1}, or 10^{45} W, for a typical burst duration of 10 s. If we assume that the burst emanates from a compact object with a mass of a few solar masses, the radius of the emitting region is likely to be around 100 km (a few Schwarzschild radii, where the Schwarzschild radius of an object of mass M solar masses is $3M$ km). The number density of photons at radius r_0 is roughly

$$n_\gamma = \frac{L_\gamma}{4\pi r_0^2 c\overline{E}_\gamma},$$

where L_γ is the luminosity in γ-rays and $\overline{E}_\gamma$ is the average γ-ray energy, of order 1 MeV. Then the *compactness parameter* ℓ', which is essentially the optical depth for photons with $E_\gamma \geq m_e c^2$ against two-photon pair production, $\gamma\gamma \to e^+e^-$, is given by[219]

$$\ell' \sim \tau_{\gamma\gamma} \sim n_\gamma \sigma_\mathrm{T} r_0 \sim \frac{f\sigma_\mathrm{T} L_\gamma}{4\pi r_0 c\overline{E}_\gamma} \sim 10^{15}, \tag{2.78}$$

where f is the fraction of L_γ coming from photons above $m_e c^2$. This is obviously an enormous optical depth, and implies that any photons with energy $> m_e c^2$ should be efficiently removed from the observed spectrum. This contradicts observations, since *Fermi*–LAT regularly observes photons of GeV energies emitted by GRBs. The escape is that the kinematics of pair production imply that the energy threshold is dependent on the angle between the directions of the photons in question: in fact pair production requires

$$2(m_e c^2)^2 \leq E_{\gamma 1} E_{\gamma 2}(1 - \cos\theta) \simeq \tfrac{1}{2} E_{\gamma 1} E_{\gamma 2} \theta^2$$

if θ is small. We saw in section 2.3.5 that photons emitted by relativistic particles are confined to a cone of half-angle $1/\gamma$. Therefore, requiring that γ-ray photons are *not* depleted by pair production implies a limit for the bulk Lorentz factor of the presumed jet (conventionally denoted Γ rather than γ) of

$$\Gamma \geq \frac{1}{2}\sqrt{\frac{E_{\gamma 1}}{m_e c^2}\,\frac{E_{\gamma 2}}{m_e c^2}}. \tag{2.79}$$

If a GRB is seen to emit photons of 30 GeV and the average photon energy is $\sim$1 MeV, this implies that $\Gamma \geq 100$, i.e. a highly relativistic jet. More precise calculations, considering the photon spectrum in detail, show that this is a lower limit, with $\Gamma \sim 250$ required even if the photon energy spectrum only extends to 100 MeV[219].

The emission mechanism for the soft γ-rays is generally assumed to be synchrotron radiation associated with internal and external shocks in the relativistically expanding fireball[219]. The internal shock is believed to be responsible for the prompt emission, primarily soft γ-rays, and the external shocks for the afterglow. The high-energy photons observed by *Fermi*–LAT are assumed to be due to synchrotron-self-Compton emission. Although this picture has attractive features, it is not without problems: the radiative efficiency of the internal shocks may not be high enough to account for the observed luminosity, and some bursts have high-energy Band spectral indices $\beta > -2/3$, which is not compatible with simple synchrotron radiation. There are a number of proposed explanations for these issues[219], but there is no single generally accepted model that fits all the data with no issues.

We will discuss GRBs in more detail in the chapter on "Astrophysical Sources".

2.4.5 Intermediate energy γ-rays

For the purposes of this course, intermediate-energy γ-rays are those between $\sim$30 MeV and $\sim$300 GeV—energies which are high enough to allow detection by e^+e^- pair production, but not so high that the rates are too small for a space-based detector (with the size limitations imposed by available launch vehicles). The main instruments covering this energy range are ESA's pioneering COS-B[225] (1975–1982) EGRET on CGRO[33] (1991–2000), and the presently operational instruments: the Italian satellite AGILE[226] and the Large Area Telescope on *Fermi*[227].

All of these instruments are remarkably similar in general concept, with only details of the technology differing. The detection principle is *pair conversion*: the incoming γ-ray converts into an e^+e^- pair in the electric field of an atom (this is two-photon pair production, $\gamma\gamma \to e^+e^-$, with the second, very low-energy, photon coming from the electromagnetic field); the outgoing electron and positron are tracked in order to reconstruct the direction of the incoming photon, and finally the photon energy is measured by absorbing the energy of the e^+e^- pair in a calorimeter. The whole detector is covered in an anticoincidence counter[4] to reject incoming charged particles. Figure 2.50 shows a schematic of the EGRET detector; compare this with figure 2.51 to see the same geometry in the other experiments.

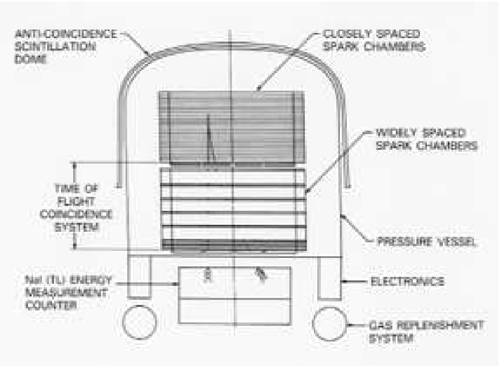

ANTI-COINCIDENCE SCINTILLATION DOME
CLOSELY SPACED SPARK CHAMBERS
WIDELY SPACED SPARK CHAMBERS
TIME OF FLIGHT COINCIDENCE SYSTEM
PRESSURE VESSEL
NaI (TL) ENERGY MEASUREMENT COUNTER
ELECTRONICS
GAS REPLENISHMENT SYSTEM

Figure 2.50: Schematic of the EGRET pair-conversion telescope on board CGRO[33].

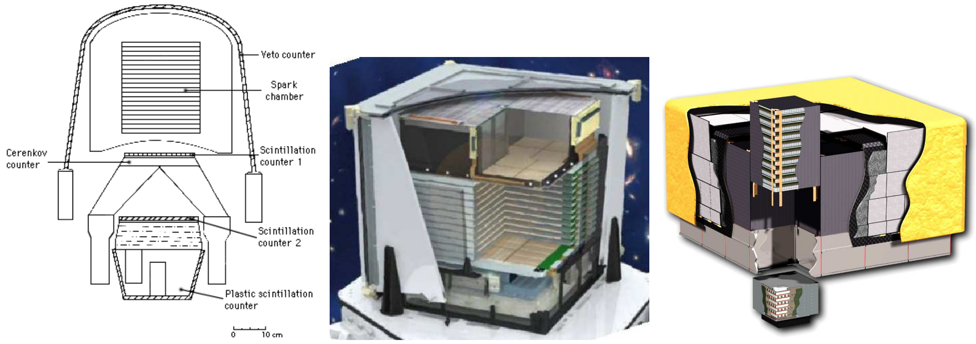


Figure 2.51: Other pair-conversion γ-ray telescopes: left, a schematic of COS-B[225]; centre, a cutaway image of AGILE[226]; right, a diagram of *Fermi*–LAT showing one tracker module and one calorimeter module[227] (the full experiment is a 4×4 array of such modules). The same basic structure is visible in all cases.

[4] "Anticoincidence" means that this detector is required *not* to detect anything when a candidate photon is registered in the main detector.

Both COS-B and EGRET used spark chambers for the converter/tracker section of the telescope. A spark chamber[228] consists of an array of parallel plates or wire grids installed in a gas-tight container filled with a noble gas such as helium or argon. When the detector is triggered, normally by signals from scintillation counters, a very high voltage is applied to the plates. If a charged particle has passed through the chamber and caused ionisation, the resulting free electrons will initiate a spark across the gap between two adjacent plates. Spark chambers were used for particle physics experiments in the 1950s and 1960s, but were already obsolete as particle detectors when COS-B was designed in the early 1970s, let alone EGRET: I assume the choice of technology was motivated by the fact that the charge signal from a spark is very large, and therefore can be read out using very simple electronics, whereas more modern gaseous ionisation detectors such as proportional counters and drift chambers typically require pre-amplifiers because of the small signals. The disadvantage of spark chambers is that the operation of the chamber degrades the gas, which must be periodically flushed and refilled: this limits the operational lifetime of the instrument. Despite this, both COS-B and EGRET managed to exceed their design lifetimes through clever management of resources. To encourage the incoming photons to convert, the active layers are interleaved with "converter" plates of high-density metal: EGRET used tantalum. For optimum resolution, the incoming γ-ray should convert as early as possible: COS-B achieved this by varying the thickness of the converter layers, with thicker plates at the front end of the chamber; EGRET, as shown in figure 2.50, divided its tracker into two sections, one with closely-spaced layers encouraging conversion, and one with wider gaps intended mainly for tracking.

AGILE and LAT, launched in 2007 and 2008 respectively, replaced spark chambers with silicon detectors. These extremely similar designs both used tungsten sheets as converters, with two orthogonal layers of silicon strip detectors below each converter, giving 2D readout. AGILE has 12 layers, the bottom two without tungsten; LAT has 18. Silicon-strip detectors are very widely used in particle physics, e.g. the ATLAS central tracker[229]. They have many advantages over spark chambers, notably that they require much lower operating voltages and are less susceptible to aging.

All four experiments used scintillating crystals for calorimetry: COS-B, AGILE and LAT all use CsI, while EGRET chose NaI. The advantage of crystal calorimeters is that the whole volume is active—unlike sampling calorimeters which consist of alternating layers of absorber and detector—and they can be relatively compact. COS-B's calorimeter was a single large crystal, but the other three are all segmented to provide some positional information. In the two older experiments, COS-B and EGRET, the scintillation light was detected using photomultiplier tubes, while the second-generation AGILE and LAT use photodiodes. The anticoincidence shields are also scintillator-based, but this time plastic scintillator, which is less expensive and lends itself to being formed into relatively thin sheets. COS-B and EGRET had one-piece anticoincidence domes, but the anticoincidence detectors of AGILE and LAT are segmented. In the case of AGILE, the segmentation was designed to facilitate effective triggering over a wide field of view and to contribute to direction reconstruction; in the LAT, the main purpose of segmenting the anticoincidence detector was to reduce the rate of false vetoes caused by backward-going particles from the electromagnetic shower induced by the e^+e^- pair in the calorimeter, which can enter the anticoincidence shield from the inside and be misinterpreted as evidence that the incoming particle was charged. This effect reduced the efficiency of EGRET at high energies; since the LAT was designed to reach higher maximum energies than EGRET, it was essential to reduce the false veto rate[227]. Segmentation means that only those segments of the anticoincidence shield located close to the reconstructed direction of the incoming particle can veto the event, while signals from elsewhere are interpreted as "backsplash" from the calorimeter.

The performance of pair-conversion spectrometers is quite strongly dependent on energy. At low energies, the e^+e^- pair are soft and may scatter in the converter-tracker, reducing the angular resolution; at high energies, there may be leakage out of the back of the calorimeter, degrading energy resolution. There will also be a dependence on angle of incidence: a normally incident photon will traverse the whole of the detector, but one coming in at an angle may be lost out of the side before passing through the entire system. The *Fermi*–LAT website at Stanford University[230] provides plots of the LAT performance as a function of energy and angle of incidence for a number of variables, including effective area, angular resolution and energy resolution. For normally incident photons, 68% of the shower is contained with 1° for all energies over 1 GeV, improving to 0.1° at 100 GeV, and the energy resolution varies between 7% and 15%, with the best resolution at energies of a few GeV. EGRET had an energy resolution of 15% (FWHM) with 67% of the gamma rays from a point source contained inside an angle $\theta = 5.85^\circ E^{-0.534}$ with E in MeV[231]: this corresponds to 0.15° at 1 GeV. Note that this is not the same as the angular resolution for locating a point source, because an identified point source will necessarily have produced multiple photons: EGRET could locate a point source to within 5–30′, *Fermi*–LAT to within 30″. This improvement in point source location is critical in identifying γ-ray sources with optical counterparts: 63% of the EGRET γ-ray point sources were unidentified, in part because the large error boxes let in too many potential candidates. The corresponding fraction of unidentified sources in the much larger (1873 objects, as opposed to EGRET's 271) *Fermi*–LAT two-year catalogue[232] is "only" 31%: still very large, but a factor of two better.

Gamma rays of these energies are most likely to be produced by inverse Compton scattering or π^0 decays; bremsstrahlung and synchrotron radiation are more significant at lower energies. Inverse-Compton sources are often radio sources as well, since provided the source has a magnetic field the relativistic electrons required for inverse Compton scattering should also produce synchrotron radiation. This is not necessarily the case for sources powered by π^0 decay, since here the parent population is relativistic hadrons, but it is by no means unlikely that relativistic hadrons would be accompanied by relativistic electrons, so the presence of synchrotron radiation is not proof of an inverse-Compton origin.

Figure 2.52 shows the point sources in the *Fermi*–LAT 2-year catalogue, with source associations. The vast majority of identified sources (57% of all sources; 83% of sources with identifications) are *blazars*[232]—a class of active galactic nucleus typified by a high degree of variability, and believed to represent a situation where the relativistic jet emerging from the AGN is directed very close to our line of sight. Most of the other identified sources are pulsars or supernova remnants in our Galaxy. In addition, there is diffuse emission from the Galactic plane, dominated by π^0 decay[231] and presumably caused by cosmic rays interacting with the interstellar medium. A striking feature discovered by *Fermi*–LAT is the presence of two "bubbles" of γ-ray emission approximately perpendicular to the Galactic plane[233], which appear to be

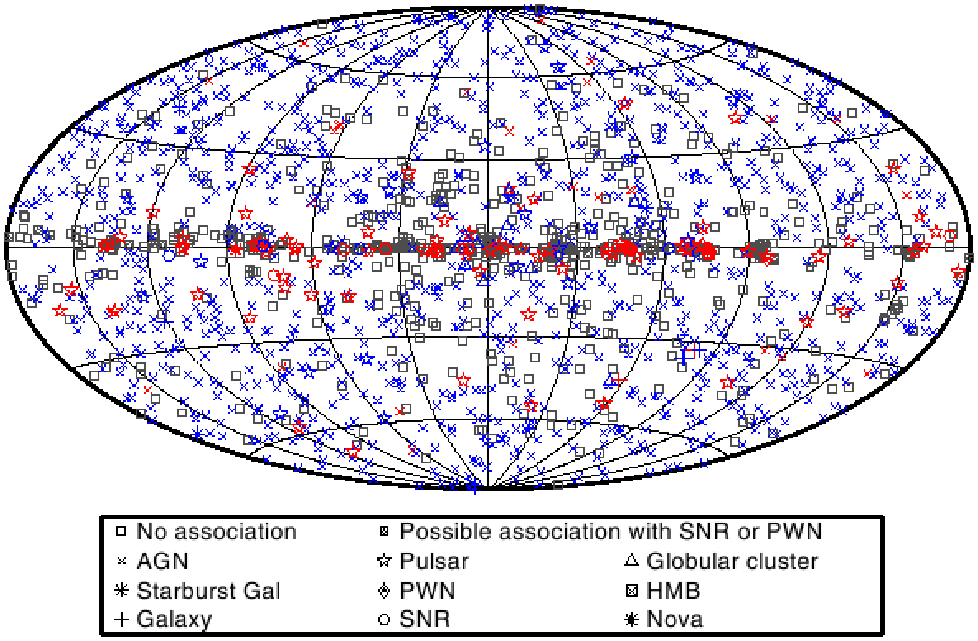


Figure 2.52: The *Fermi*–LAT 2-year catalogue of point sources, with source identifications where they exist. Blue points are "associated" sources, identified on the basis of position; red points are positively identified, either by correlated variability (for example, variation with the same time period as a pulsar within the error ellipse, or irregular outbursts at the same times as a variable AGN in the error ellipse) or by similarity in size and shape in the case of extended sources such as nearby supernova remnants. Gamma ray bursts, being transient, are not included in this catalogue. Figure from Nolan et al.[232].

correlated with features seen in soft X-rays by ROSAT and with a feature in the WMAP microwave observations. It seems likely that these structures were created by a recent burst of activity in the Galactic centre, either an accretion event on to the Sgr A* black hole or a burst of star formation[233].

2.4.6 High-energy γ-rays

Fermi–LAT detects γ-rays up to energies of $\sim$300 GeV. Higher energies are problematic for two reasons: first, the energy resolution will become increasingly degraded as the calorimeter is too small to contain the shower, and secondly, the flux of such high-energy γ-rays is so low that the statistics collected by the LAT, with its effective area of $\sim$0.8 m^2[230], would be too low for useful physics. Both of these problems are intractable for a space-based platform, because they both require a much larger and heavier instrument, whose launch costs would be unrealistic. Therefore it is very difficult to extend the energy range of space-based instrumentation beyond *Fermi*.

Given that γ-rays do not penetrate the Earth's atmosphere, this would appear to make observations of TeV-energy γ-rays impossible. Fortunately, this is not the case: as with hadronic cosmic rays (see section 2.2.2), the air shower produced by the primary particle's interaction with the atmosphere can be detected from the ground and used to infer the properties of the particle.

As shown in figure 2.53, air showers initiated by high-energy photons differ from hadron-induced showers in a number of ways:

- they require an initial pair-production, and therefore tend to start deeper in the atmosphere than hadron-induced showers;

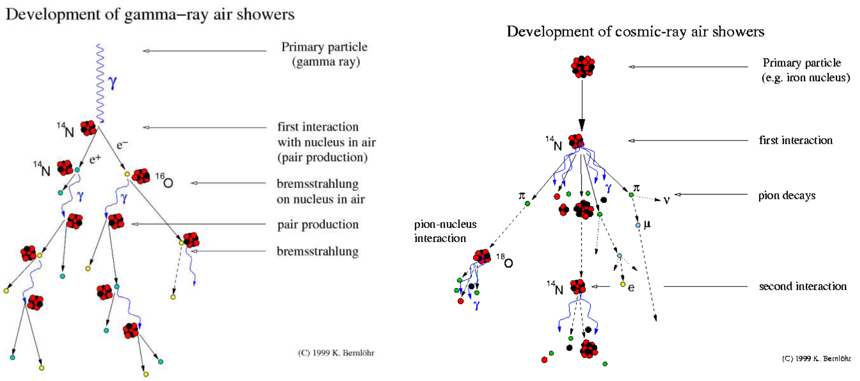


Figure 2.53: Schematic diagrams of air showers initiated by a γ-ray (left) and a hadron (right)[234].

- they contain only e^-, e^+ and γs, unlike hadron-induced showers, which will also contain $\mu^\pm$ and ν from $\pi^\pm$ decay, and probably nuclear fragments (nucleons and small nuclei) as well;
- they are narrower than hadron-induced showers, and more regular in shape (nuclear fragments from hadron-induced showers can initiate independent sub-showers;
- because of the lack of penentrating particles, little if any of the shower reaches ground level.

Electron-induced showers are very similar to photon-induced, but start earlier because the initial pair-production step is not needed.

As we saw earlier (page 50), the usual techniques for detecting hadron-induced showers are nitrogen fluorescence (detecting the shower in the atmosphere) and ground arrays (detecting shower particles that reach ground level). The fairly poor angular resolution provided by these reconstruction methods is not really an issue with charged primaries, since—as discussed above—we do not expect charged primaries to point back to their sources in any case.

TeV γ-rays, in contrast, *should* point back to their source, since they are unaffected by magnetic fields. To improve angular resolution, the preferred technique is therefore the detection of Cherenkov radiation produced by shower particles travelling faster than c/n, where n is the refractive index of air. As noted on page 49, this corresponds to an energy threshold of order 25 MeV for electrons. The Cherenkov angle $\cos^{-1}(c/n)$ corresponds to a half-angle for the Cherenkov cone of about 1.3° (depending on temperature and density of the air, and hence on the height at which the shower is initiated). This produces a pool of Cherenkov light about 250 m in diameter (depending on the height of the shower and the incidence angle of the primary photon); any telescope located within the pool will see a streak of light pointing back towards the shower direction[235]. The precision with which the primary direction is reconstructed is greatly improved if multiple telescopes are used to provide "stereoscopic" views of the shower: both H.E.S.S.[235] and VERITAS[236] have an array of four telescopes, supplemented in the former case by a single larger telescope which was added later.

Theory of Cherenkov emission

Cherenkov radiation occurs when a charged particle travels through a medium with refractive index n at a speed $v > c/n$, i.e. the speed of the particle is greater than the speed of light *in the medium*. There are a number of different approaches to deriving the spectrum of Cherenkov radiation; here we summarise Longair[171] section 9.7.

The moving electron corresponds to a current density $\mathbf{J}$. If we define coordinates such that the electron is moving along the positive x-axis, then at time t

$$\mathbf{J}(\mathbf{r},t) = e\mathbf{v}\delta(x-vt)\delta(y)\delta(z), \tag{2.80}$$

where $\delta(s)$ represents the **Dirac delta function**. The Dirac delta function[237] $\delta(s)$ is zero for all $s \neq 0$ and has an integral $\int_{-\infty}^{+\infty} \delta(s)\,\mathrm{d}s = 1$; it can be thought of as a Gaussian of zero width. In the above equation, the delta functions make the value of $\mathbf{J}$ zero except at the location of the electron.

The Fourier transform of $\mathbf{J}(\mathbf{r},t)$ is

$$\begin{aligned}\mathbf{J}_\omega(\mathbf{r}) &= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} \mathbf{J}(\mathbf{r},t)\, e^{i\omega t}\mathrm{d}t \\ &= \frac{e}{\sqrt{2\pi}}\delta(y)\delta(z)\, e^{i\omega x/v}\hat{\mathbf{x}},\end{aligned} \tag{2.81}$$

where $\hat{\mathbf{x}}$ is the unit vector in the x direction. In doing the integral we have used two properties of the Dirac delta function:

- $\int_{-\infty}^{+\infty} f(x)\delta(x-a) = f(a)$, i.e. the integral of a function times a delta function just picks out the only value of that function for which the delta function is not zero;

- $\delta\alpha x = \delta x/|\alpha|$, where α is a constant—therefore $\delta(x-vt) = \delta(\frac{x}{v}-t)/v$, which cancels out the v in the integrand.

If we express Maxwell's equations[238] in terms of the vector potential $\mathbf{A}(\mathbf{r},t)$ and the scalar potential $\phi(\mathbf{r},t)$, we find[238]

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu_0\mathbf{J} \tag{2.82}$$

in a non-magnetic medium in which the relative permeability $\mu = 1$. This form of the equation for $\mathbf{A}$ is valid in the so-called *Lorentz gauge* defined by

$$\nabla\cdot\mathbf{A} + \frac{1}{c^2}\frac{\partial\phi}{\partial t} = 0, \tag{2.83}$$

but as Maxwell's equations are *gauge invariant* we are allowed to assume this—gauge invariance implies that our results will not depend on choice of gauge (but the complexity of the working often *does* depend on choice of gauge—Lorentz gauge is usually the most convenient for working with time-dependent fields).

This differential equation has a standard general solution

$$\mathbf{A}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}',t')}{R}\,\mathrm{d}^3\mathbf{r}', \tag{2.84}$$

where $t' = t - R/c$ and $\mathbf{R} = \mathbf{r} - \mathbf{r}'$. The reason that we evaluate $\mathbf{J}$ at t' instead of t is that it takes the field a time R/c to propagate from $\mathbf{r}'$ to $\mathbf{r}$. Currents and potentials evaluated at t' are known as *retarded* currents and potentials.

The electric field corresponding to this vector potential is given by

$$\mathbf{E}(\mathbf{r}) = -\frac{\partial \mathbf{A}}{\partial t} = \frac{\mu_0}{4\pi}\int \frac{\dot{\mathbf{J}}(\mathbf{r}', t')}{R} \mathrm{d}^3\mathbf{r}, \tag{2.85}$$

where $\dot{\mathbf{J}} = \mathrm{d}\mathbf{J}/\mathrm{d}t$.

As in the case of radiation from an accelerated charge (see section 2.3.3), the electric field associated with the electromagnetic radiation is perpendicular to $\mathbf{r}$, $E_r = |\mathbf{E}(\mathbf{r}) \times \hat{\mathbf{r}}| = E\sin\theta$, where θ is the angle between $\mathbf{r}$ and $\mathbf{J}$, i.e. between $\mathbf{r}$ and the electron's velocity vector.

In a non-magnetic medium, the refractive index n is just given by the relative permeability ϵ: $n = sqrt\epsilon$. This modifies the Poynting vector flux from $c\epsilon_0 E_r^2$ as in section 2.3.3 to $nc\epsilon_0 E_r^2$ (equivalent to $c_n\epsilon\epsilon_0 E_r^2$, where $c_n = c/n = c/\sqrt{\epsilon}$ is the speed of light in the medium). Therefore, the total power radiated in Cherenkov radiation is given by integrating $nc\epsilon_0 E_r^2$ over the surface of a sphere of radius r:

$$\begin{aligned}\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right)_{\mathrm{rad}} &= nc\epsilon_0 \int E_r^2 r^2 \mathrm{d}\Omega \\ &= \frac{nc\epsilon_0\mu_0^2}{16\pi^2}\int \sin^2\theta \left|\int \frac{\dot{\mathbf{J}}(\mathbf{r}', t')}{R}\mathrm{d}^3\mathbf{r}'\right|^2 r^2\mathrm{d}\Omega \\ &= \frac{n}{16\pi^2\epsilon_0 c^3}\int \sin^2\theta \left|\int \dot{\mathbf{J}}(\mathbf{r}', t')\mathrm{d}^3\mathbf{r}'\right|^2 \mathrm{d}\Omega,\end{aligned} \tag{2.86}$$

assuming $r \gg r'$ so that $R \simeq r$, and using $\mu_0\epsilon_0 = 1/c^2$.

To obtain the spectrum, we take the same approach as we did in section 2.3.4: integrate over time and Fourier transform to convert into an integral over ω, using Parseval's theorem: this gives

$$E_{\mathrm{rad}} = \frac{n}{8\pi^2\epsilon_0 c^3}\int_0^\infty \int_\Omega \left|\int \dot{\mathbf{J}}_{\omega,\mathrm{r}}(\mathbf{r}')\mathrm{d}^3\mathbf{r}'\right|^2 \mathrm{d}\Omega\,\mathrm{d}\omega, \tag{2.87}$$

where $\mathbf{J}_{\omega,\mathrm{r}}$ is the retarded current as expressed in the frequency domain.

To evaluate the volume integral, we first evaluate the phase factor caused by the propagation of the electromagnetic waves from $\mathbf{r}'$ to $\mathbf{r}$. Using the usual expression for a plane wave, we have

$$\exp\left((i(\omega t) - \mathbf{k}\cdot\mathbf{r}')\right).$$

This is the only time dependence in $\mathbf{J}_{\omega,\mathrm{r}}$, so the volume integral becomes

$$\left|\int \dot{\mathbf{J}}_{\omega,\mathrm{r}}(\mathbf{r}')\mathrm{d}^3\mathbf{r}'\right| = \left|i\omega e^{i\omega t}\int e^{-i\mathbf{k}\cdot\mathbf{r}'}\mathbf{J}_\omega(\mathbf{r}')\mathrm{d}^3\mathbf{r}'\right|. \tag{2.88}$$

Evaluating $\mathbf{J}_\omega(\mathbf{r}')$ using equation (2.81) gives

$$\left|\int \dot{\mathbf{J}}_{\omega,\mathrm{r}}(\mathbf{r}')\mathrm{d}^3\mathbf{r}'\right| = \left|\frac{\omega e}{\sqrt{2\pi}}\int \exp\left[ikx\left((\cos\theta + \frac{\omega}{kv}\right)\right]\mathrm{d}x\right|, \tag{2.89}$$

where we have used the $\delta(y)\delta(z)$ in $\mathbf{J}_\omega$ to convert $\mathrm{d}^3\mathbf{r}'$ into $\mathrm{d}x$. The factors i and $e^{i\omega t}$ go away in taking the magnitude.

Since

$$\int_{-\infty}^{+\infty} e^{ik(a-b)}\mathrm{d}k = \frac{1}{2\pi}\delta(a-b),$$

the integral is only non-zero if

$$\cos\theta = -\frac{\omega}{kv}$$

for some value of θ. Since $\omega/k = c_n$, where $c_n = c/n$ is the speed of light in the medium, this is possible only if $v > c_n$, i.e. the particle is travelling faster than the speed of light in the medium, as stated earlier.

Substituting equation (2.89) into equation (2.87) and writing $\sin^2\theta = 1 - \cos^2\theta = 1 - 1/(n\beta)^2$ from the argument above, we have

$$\frac{\mathrm{d}E_{\mathrm{rad}}}{\mathrm{d}\omega} = \frac{n\omega^2 e^2}{16\pi^3\epsilon_0 c^3}\left(1 - \frac{1}{n^2\beta^2}\right)\int_\Omega \left|\int \exp\left[ikx\left(\cos\theta + \frac{\omega}{kv}\right)\right]\mathrm{d}x\right|^2 \mathrm{d}\Omega. \quad (2.90)$$

To do the integral over x, Longair[171] suggests integrating over the finite distance $-L < x < +L$ (it can also be done using complex residues). Writing $\alpha = k[\cos\theta + 1/(n\beta)]$, we have

$$\int_{-L}^{+L} e^{i\alpha x}\mathrm{d}x = \frac{1}{i\alpha}\left(e^{i\alpha L} - e^{-i\alpha L}\right) = \frac{2\sin\alpha L}{\alpha}.$$

The element of solid angle $\mathrm{d}\Omega = \mathrm{d}\phi\mathrm{d}(\cos\theta) = \mathrm{d}\phi\mathrm{d}\alpha/k$, so equation (2.90) can be written

$$\begin{aligned}\frac{\mathrm{d}E_{\mathrm{rad}}}{\mathrm{d}\omega} &= \frac{n\omega^2 e^2}{4\pi^3\epsilon_0 c^3}\left(1 - \frac{1}{n^2\beta^2}\right)\int \frac{\sin^2\alpha L}{\alpha^2} 2\pi\frac{\mathrm{d}\alpha}{k} \\ &= \frac{\omega e^2}{2\pi^2\epsilon_0 c^2} L \int \frac{\sin^2(\alpha L)}{(\alpha L)^2}\mathrm{d}(\alpha L),\end{aligned} \quad (2.91)$$

where we have substituted $k = n\omega/c$ in the second line, and the ϕ integral just introduces a factor of 2π as there is no ϕ dependence in the function.

The integral $\int_{-\infty}^{+\infty}(\sin^2\theta/\theta^2)\mathrm{d}\theta$ is a standard integral[239] and has the value π. Therefore, the energy radiated in Cherenkov light per unit path length is

$$\frac{\mathrm{d}E_{\mathrm{rad}}}{\mathrm{d}\omega\mathrm{d}x} = \frac{\omega e^2}{4\pi\epsilon_0 c^2}\left(1 - \frac{1}{n^2\beta^2}\right). \quad (2.92)$$

Changing variables from x to t, the Cherenkov light emitted per unit time is

$$I(\omega) = \frac{\mathrm{d}E_{\mathrm{rad}}}{\mathrm{d}\omega\mathrm{d}t} = \frac{\omega e^2 v}{4\pi\epsilon_0 c^2}\left(1 - \frac{1}{n^2\beta^2}\right). \quad (2.93)$$

The dominant functional dependence in equation (2.93) is $I(\omega) \propto \omega$, which explains why Cherenkov radiation appears blue. However, note that the refractive index n is also a function of ω, so the Cherenkov spectrum is not simply linear with ω (and does not blow up at high frequencies). Note that neither the intensity nor the spectrum depends on the mass of the particle, except in so far as more massive particles will require higher energies in order to pass the Cherenkov threshold. For ultra-relativistic particles, where $v \simeq c$, the intensity and spectrum do not depend on the energy of the particle either: for electrons in air showers, the relative intensity increases rapidly once the threshold is passed, so that above ~100 MeV there is little dependence on particle energy.

Cherenkov radiation from air showers

In the context of high-energy γ-ray detection, we are interested in Cherenkov radiation from photon-induced electromagnetic showers, which develop by a combination of pair production and bremsstrahlung as shown in the left panel of figure 2.53. The shower terminates when the mean energy of the particles decreases to the point at which pair production and bremsstrahlung are no longer the dominant energy loss mechanisms—typically a few MeV. (Of course, from the point of view of an air Cherenkov telescope the shower terminates when the electrons drop below the Cherenkov threshold in air, which is about 25 MeV.) The number of Cherenkov-radiating particles produced in a photon-induced shower is of order $E_\gamma/E_{\rm thr}$, where $E_{\rm thr} \simeq 25$ MeV; since we saw above that the intensity of the Cherenkov radiation is not strongly dependent on the energy of the radiating particle except very close to threshold, this means that the observed intensity of Cherenkov radiation from a photon-induced air shower is proportional to the number of particles in the shower, which in turn is proportional to the energy of the parent photon. Therefore, the energy of a very high-energy γ-ray can be deduced from the intensity of the Cherenkov radiation.

The lack of dependence of intensity on particle energy also implies that most of the Cherenkov radiation will come from the tail end of the shower, when the $e^\pm$ are a factor of a few above Cherenkov threshold. The higher the energy of the incoming photon, the greater the depth of shower maximum (since each generation of shower particles has about half the mean energy of the preceding one, the higher the initial energy, the more shower generations are possible before Cherenkov threshold is reached).

The threshold energy, cone angle and intensity of Cherenkov radiation all depend on the refractive index of the medium. The refractive index of air depends on its pressure and temperature, both of which vary with altitude. Figure 2.55 shows the variation of these properties with altitude, neglecting humidity (which also affects n) and assuming the atmospheric temperature and pressure profile provided by Sable Systems International[242].

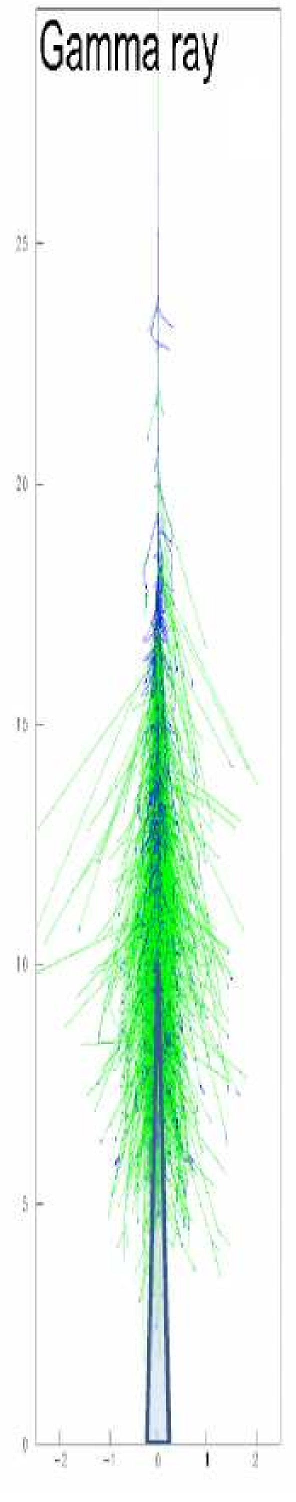


Figure 2.54: Photon-induced air shower with schematic Cherenkov light cone overlaid. Photon shower from STACEE web page [240].

As with cosmic ray air showers, it is common to express the location of and depth within the shower in terms of air mass rather than height above sea level. The density of air as a function of height, $\rho(h)$ is given approximately by

$$\rho(h) \simeq \frac{\mu p_0}{kT} \exp\left(-\frac{\mu g h}{kT}\right), \qquad (2.94)$$

where μ is the average molecular mass, p_0 is the sea level pressure, T is the temperature and k is Boltzmann's constant. Putting in numbers gives

$$\rho(h) \simeq 1.225 \exp\left(h/8400\right),$$

where h is measured in metres and ρ in kg m^{-3}. A more precise estimate can be obtained by applying the gas laws directly to a pressure and temperature profile such as [242], but this approximation is good to within 10% up to heights of 15 km or so. The column density or air mass corresponding to height h is then given by

$$X = \int_h^\infty \rho x \, \mathrm{d}x = 10300 e^{-h/8400}, \tag{2.95}$$

where X is measured in kg m^{-2}. The quantity

$$h_0 = \frac{kT}{\mu g} = 8400 \text{ m}$$

is called the *scale height* of the atmosphere.

Electromagnetic showers are characterised by the *radiation length* X_0, which is the column density through which an electron will lose $1/e$ of its energy to bremsstrahlung, or $\frac{7}{9}$ of the mean free path before pair production of a high-energy photon[243]. The radiation length of air is 371.5 kg m^{-2}[244], so we would expect photon showers to start at a height of about 25 km. It is usually assumed, as an order of magnitude estimate, that the bulk of the Cherenkov light comes from a height of about 10 km.

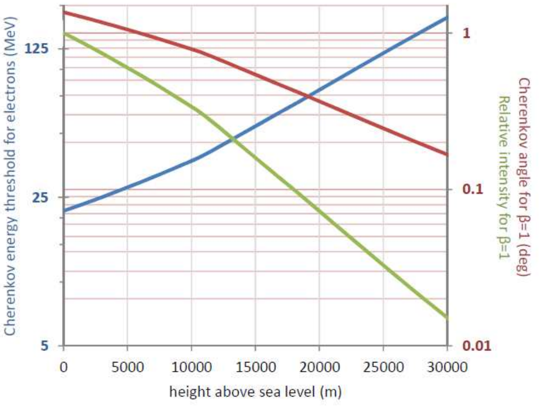


Figure 2.55: Variation of Cherenkov light properties with altitude in the atmosphere: threshold energy for $e^{\pm}$ in MeV (blue, left scale); Cherenkov angle for $\beta \simeq 1$ in degrees (red, right scale); relative intensity compared to sea level (green, right scale). The variation of the refractive index of air with temperature and pressure is taken from Kaye and Laby[241]; the variation of temperature and pressure with height are taken from the table at [242]. A wavelength of 450 nm is assumed, and the relative intensity and Cherenkov angle assume $1 - \beta \ll n - 1$.

Cherenkov radiation from air showers is extremely faint—a TeV-energy primary photon produces only about 100 Cherenkov photons per square metre in the ground-level Cherenkov light pool[235]. This gives air Cherenkov telescopes a poor duty cycle: they can operate only on dark, clear nights, in contrast to space-based platforms such as *Fermi* which are "on" all the time. On the other hand, the light pool produced by a high-energy γ-ray is about 140 m in radius (assuming a Cherenkov angle of 0.8° at a height of ∼10 km), which gives an effective area of ∼60000 square metres, since a Cherenkov telescope sited anywhere in the light pool will detect the shower. This is to be compared with of order 1 m^2 for space-based telescopes. We conclude that for high-energy (and correspondingly low flux) γ-rays, the air shower technique is much more effective than space-based observation.

Imaging air Cherenkov telescopes

The aim of an IACT is to detect photon-induced air showers, distinguish them from hadron- and (if possible) electron-induced showers[5], and measure the energy and direction of the incoming γ-ray. Although fluorescence detectors and non-imaging Cherenkov arrays like Tibet–ASγ[137] are perfectly capable of detecting photon-induced showers, IACTs are preferred in this branch of experimental particle astrophysics because of their superior angular resolution: unlike charged cosmic rays, γ-rays are not deflected by magnetic fields, so if the incoming direction is accurately reconstructed the astrophysical source can be identified.

The basic design of an IACT is similar to that of a large optical telescope. The principal differences come from the fact that air showers are extended objects (so arcsecond angular resolution is unachievable in principle, and hence need not be attempted in practice) and are *extremely* faint (so the detection system must be designed to maximise sensitivity). As a consequence, IACTS have very large, segmented primary mirrors—the H.E.S.S.–I and VERITAS telescopes, which are quite small by IACT standards, have 12-metre mirrors—and focal-plane instrumentation consisting of arrays of photomultiplier tubes, which are highly efficient detectors of faint blue light. The large primary mirrors are not as accurately figured as an optical telescope, and PMT arrays produce very large pixels compared to CCD cameras, but these are not serious drawbacks because arcsecond angular resolution is not required. The size of the primary mirror does not determine the angular resolution (mainly set by the physical size of the air shower) or the effective area of the telescope (set by the size of the Cherenkov light pool), but rather the energy threshold: larger telescopes will collect more photons, and thus will be able to detect fainter showers induced by lower-energy primaries.

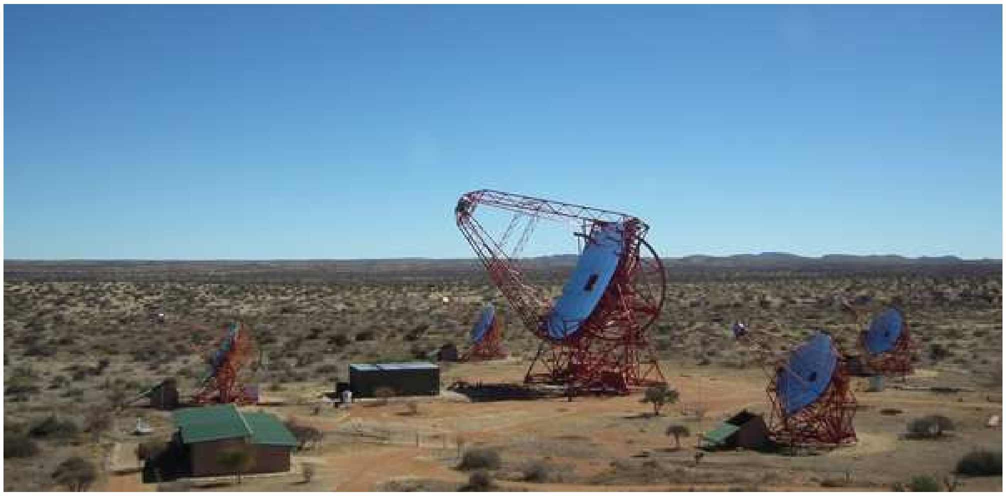


Figure 2.56: The H.E.S.S. IACT array in Namibia, showing the four 12-metre telescopes of HESS-I and the new 28-metre HESS-II. Photograph from [245]

The IACT technique was pioneered by the Whipple telescope in Arizona[246], which began operations in 1968 and has only recently been decommissioned. The Whipple telescope had a 10-metre primary mirror with an effective col-

[5] The latter is much more difficult, because an $e^{\pm}$ will initiate an electromagnetic shower just like a photon-induced shower, but some separation can be achieved because electron-induced showers do not require the initial pair-production step and therefore start slightly earlier than photon-induced showers.

lecting area of 75 m^2, and a focal plane array of 379 PMTs; it had a field of view of 2.6° and an angular resolution of $7'$. More recent telescopes have a very similar basic design, improved by larger primary mirrors and more finely segmented focal-plane instrumentation. The principal IACTS currently operating are the H.E.S.S. array in Namibia[235] (see figure 2.56), the VERITAS array in Arizona[236] and MAGIC[247] in the Canary Islands. All have more than one telescope, so that showers can be reconstructed more precisely by using stereoscopic imaging by more than one instrument, although MAGIC was a single telescope until quite recently (2010) and the MAGIC telescopes can still operate independently. In the future, these instruments will be superseded by the ambitious Cherenkov Telescope Array (CTA)[248] project, which aims for an order of magnitude improvement in sensitivity over the existing arrays.

The most advanced of the existing facilities is probably the H.E.S.S. array, which combines the single largest telescope (HESS-II, with a collecting area of 614 m^2[245]) with a square array of four smaller telescopes (108 m^2 each) to provide both a low energy threshold (30 GeV for HESS-II) and a good angular resolution ($\sim 5'$ at 20° zenith angle from stereo combination of the four telescopes of HESS-I). The focal-plane instrumentation of the H.E.S.S. telescopes consists of $1\frac{1}{4}''$ photomultiplier tubes: 960 for each of the HESS-I telescopes, giving a pixel size of $6'$ and a 5° field of view, and 2048 for HESS-II ($4'$ pixel size, 3.2° field of view). As with space-based γ-ray detection, the angular resolution for source location is much better than that for the reconstruction of individual γ-rays: H.E.S.S. can locate sources to within a few arcseconds, limited by the pointing precision of the telescopes themselves[249].

The VERITAS array is very similar to HESS-I, with four telescopes each with a collecting area of 110 m^2. The focal-plane instrumentation is slightly less sophisticated, comprising 499 PMTs with a pixel size of $9'$ and a field of view of 3.5°. Its source location accuracy is $50''$. VERITAS and H.E.S.S. are complementary, owing to their locations in the northern and southern hemisphere respectively: together, they cover the whole sky, with a substantial overlap in the equatorial region for cross-comparison.

The MAGIC telescopes[247] are midway between HESS-I and HESS-II, with an effective collecting area of 236 m^2 each and focal-plane instrumentation comprising 1039 $1''$ PMTs (pixel size $6'$, field of view 3.5°). As expected for larger telescopes, the energy threshold of MAGIC is lower than for HESS-I or VERITAS, about 50 GeV compared to ∼100 GeV for the latter.

CTA is intended to be a composite array, with a few large (24 m, slightly smaller than HESS-II) telescopes for low-energy γ-rays (tens of GeV), rather more medium-sized (10–12 m, similar to HESS-I or VERITAS) telescopes optimised for the energy range 100 GeV–1 TeV, and a large number of small telescopes (4–6 m diameter) to cover the TeV energy range where it is essential to cover a large ground area to compensate for the low flux. The exact numbers of telescopes of each class are not yet fixed, and will depend on available funding and on the selected telescope designs: in particular, for the high-energy, small-telescope array, there is a performance trade-off between the number of telescopes and their size (more small telescopes would give a greater effective area at the expense of a higher threshold energy compared to fewer, larger instruments). CTA is currently in the design stage, with proposed designs for the different telescopes being evaluated[250].

In order to analyse the high-energy γ-ray flux, it is necessary to distinguish photon-induced showers from hadron-induced cosmic-ray events. As can be seen from figure 2.53, there are differences in the shower shape: photon-induced showers have a smoother and more regular development, whereas nuclear fragments from hadron-induced showers may initiate sub-showers at some lateral distance from the primary shower. In addition, hadronic showers will include some $\mu^{\pm}$ produced from $\pi^{\pm}$ decay; owing to their long lifetime, muons are much more likely to reach the ground that the $e^{\pm}$ and γs of an electromagnetic shower, and can therefore be detected by ground arrays. IACTs are not, in general, associated with ground arrays, but muons may nevertheless be detected if they happen to hit the telescope (see figure 2.57).

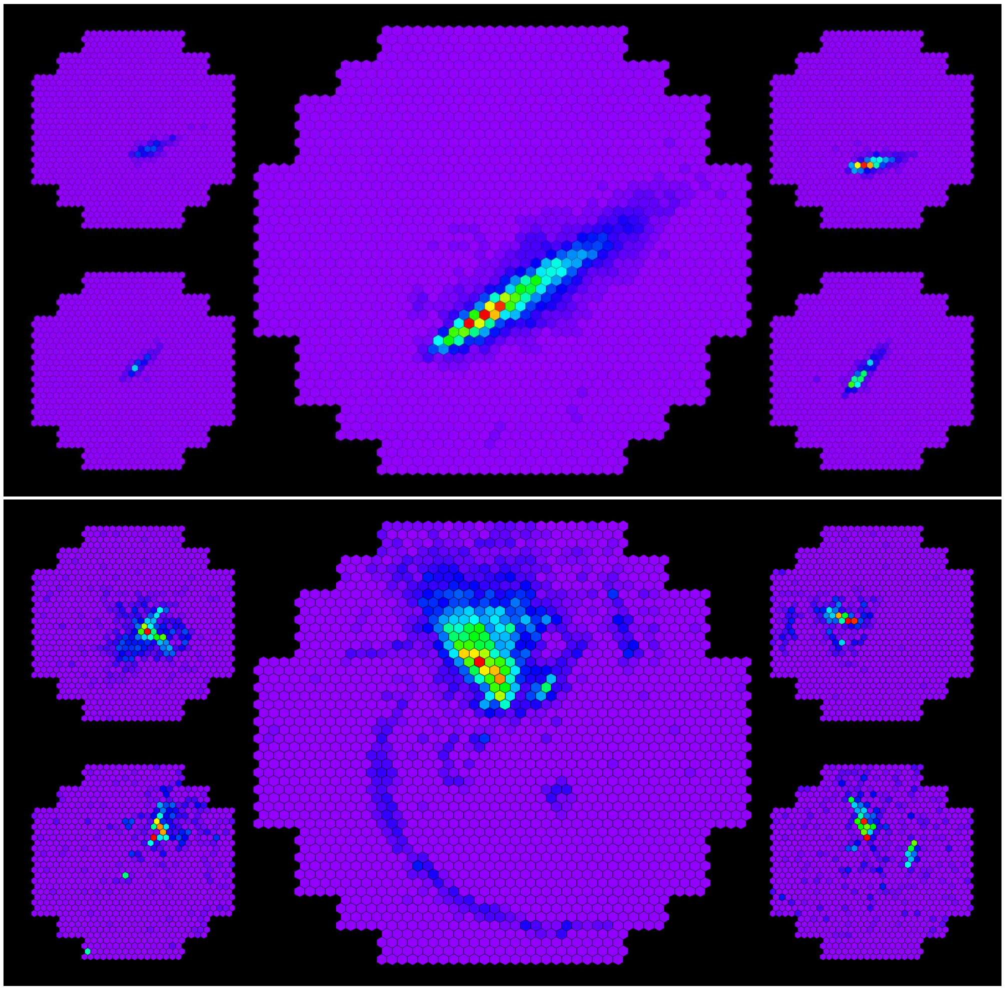


Figure 2.57: Event displays from the H.E.S.S. array[245]. Although individual events cannot be positively identified as photon- or hadron-induced without numerical analysis, the upper display corresponds with what one might expect from a photon. The lower plot, with a much less regular shower profile, is more consistent with a hadron-induced shower. Furthermore, the circular arc visible in the HESS-II display (centre) is consistent with a Cherenkov ring generated by a muon hitting the telescope, which would again imply a hadron-induced shower.

Although some events, such as that displayed in the lower panel of figure 2.57, are immediately identifiable as probably due to charged cosmic rays, in general identification requires quantitative analysis. The experiments use multivariate analyses based on the width and length of the elliptical images of the shower[251]; these can be cuts-based as in [251] or use more sophisticated techniques such as boosted decision trees[252].

As discussed above, the total Cherenkov intensity is proportional to the energy of the incoming photon, via the number of radiating particles in the shower. As this is basically a count of shower particles, we would expect the energy resolution to be approximately $\propto \sqrt{E_\gamma}$ by Poisson statistics; however, there are also uncertainties arising from instrumental and reconstruction sources which may not have this dependence. H.E.S.S. quotes a constant fractional uncertainty, $\Delta E/E \simeq 15\%$, for energy reconstruction above a threshold that depends strongly on zenith angle[251] (see figure 2.58).

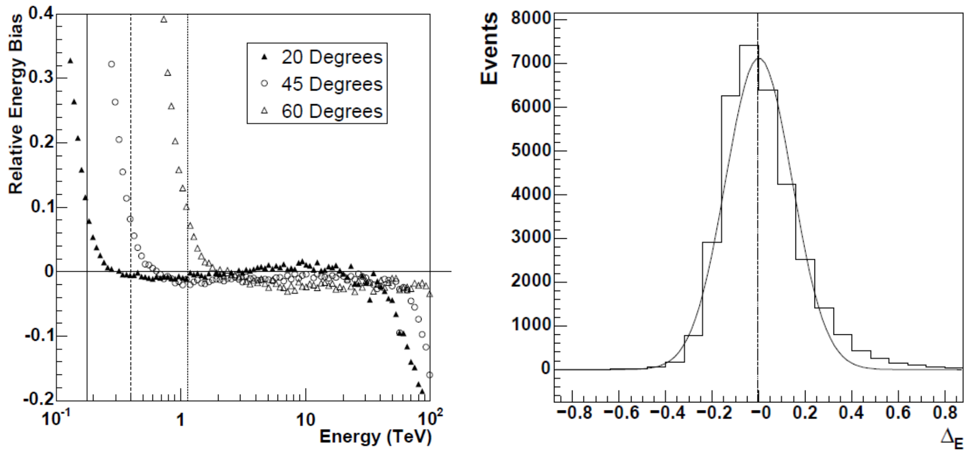


Figure 2.58: Energy reconstruction in H.E.S.S.[251]. The left panel shows the bias in energy reconstruction for four representative zenith angles; "safe" energy reconstruction is only possible for the energy range where this is small. The right panel shows the resolution for γ-rays sampled from a power law distribution $N(E) \propto E^{-2.6}$ at 50° zenith distance, above the appropriate "safe" energy of 440 GeV.

The direction of the incoming photon is reconstructed from the long axis of the shower image. In the case of stereoscopic reconstruction by two or more telescopes, the reconstructed shower direction is the intersection of the projected long axes. For a single telescope, the shower image must be compared with models to estimate the position of the shower: this is considerably less accurate, which is why stereo systems are generally preferred.

Although IACTs are primarily used for TeV γ-ray imaging, they detect and can analyse showers with other progenitors. Electron showers may be statistically distinguished from γ showers by the fact that electrons shower earlier (owing to the lack of an initial pair production). Because photons are not deflected by magnetic fields, whereas electrons are, electromagnetic showers detected when the telescopes are *not* pointing at a γ-ray source should be dominated by $e^\pm$. H.E.S.S.[253] confirmed this in a measurement of the cosmic-ray electron spectrum in 2008: the X_{max} distribution of the off-source data is clearly more consistent with electron-like showers than with expectations from γ-rays, as shown in figure 2.59.

In analyses by cosmic-ray air shower detectors such as Auger and the Telescope Array, a statistical measure of primary composition is obtained from the mean and standard deviation of the X_{max} distribution. IACTs have an additional, unique, method of identifying heavy-ion primaries, which has also been exploited by H.E.S.S.[254].

As can be seen from equation (2.92), the amount of Cherenkov light radiated depends on the square of the charge. Therefore, a heavy ion such as iron ($Z = 26$) generates hundreds of times more Cherenkov light than an electron or proton. As a consequence, IACTs can sometimes detect the direct Cherenkov

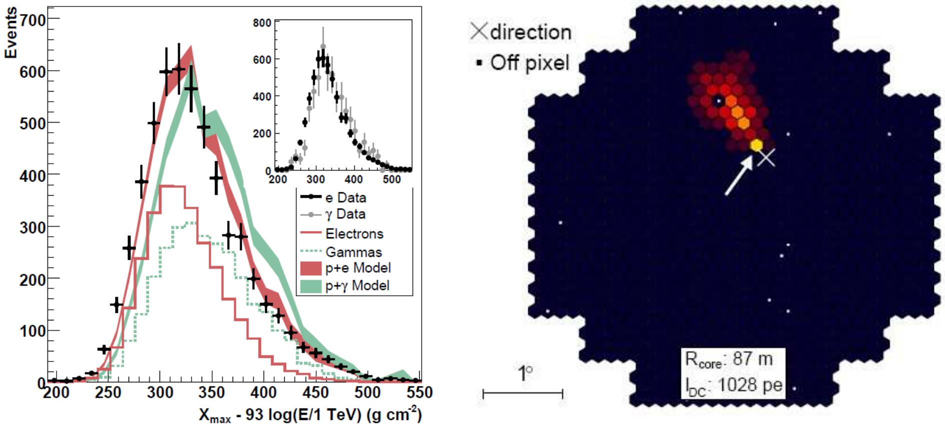


Figure 2.59: Cosmic-ray showers in H.E.S.S. The left panel[253] shows the $X_{\rm max}$ distribution for off-source showers, along with expectations from $e^{\pm}$ (red) and γ (green) showers combined with the hadronic background. The data clearly agree better with the "p + e" hypothesis; note the lower $X_{\rm max}$ value, implying an earlier shower. The right panel shows an event display for a shower initiated by a heavy ion[254]. The "direct" Cherenkov radiation, originating high in the atmosphere where the Cherenkov angle is very small, is seen as the anomalously bright pixel (arrowed) at one end of the shower. The X marks the reconstructed shower direction, which is very close to the bright pixel, as expected.

light from a heavy-ion primary as well as the light from the subsequent hadronic shower. Because the refractive index is smaller at high altitudes, this light is concentrated in a very narrow cone (see figure 2.55), and is seen in a single pixel of a telescope such as the HESS-I or VERITAS 10-metre instruments. An image of such a heavy-ion-induced shower is shown in the right panel of figure 2.59.

Analyses similar to these are also being undertaken by VERITAS[255], but have not yet been published; MAGIC has presented an electron energy spectrum at conferences[256]. Both topics are part of the science case for the Cherenkov Telescope Array[248].

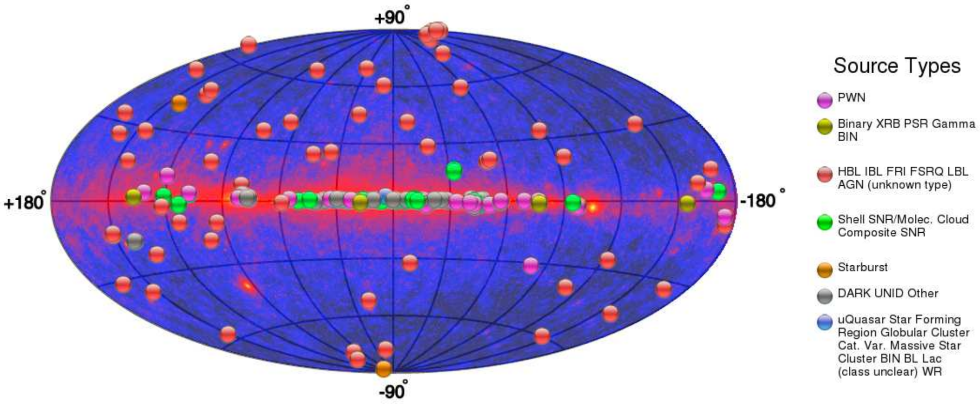

+90°
-90°
+180°
-180°
Source Types
PWN
Binary XRB PSR Gamma BIN
HBL IBL FRI FSRQ LBL AGN (unknown type)
Shell SNR/Molec. Cloud Composite SNR
Starburst
DARK UNID Other
uQuasar Star Forming Region Globular Cluster Cat. Var. Massive Star Cluster BIN BL Lac (class unclear) WR

Figure 2.60: TeV γ-ray sources superimposed on the *Fermi* γ-ray sky map, from the TeVCat web page[257].

Very high energy γ-rays are generally produced by the same processes as lower-energy γ-rays: inverse Compton scattering, bremsstrahlung, and π^0 decay (synchrotron radiation is not efficient at producing such high-energy photons directly, though synchrotron photons serve as a "seed" population for inverse

Compton scattering). Therefore, it is not surprising that TeV γ-ray sources are generally also sources of lower-energy γ-rays and X rays, and not infrequently radio. The converse is not necessarily true: not all sources of low and intermediate-energy γ-rays are observed as TeV sources.

Figure 2.60 shows the TeVCat[257] catalogue of TeV γ-ray sources, overlaid on the *Fermi* γ-ray sky map in Galactic coordinates. The distribution on the sky immediately implies that both Galactic and extragalactic sources are present: there is a clear concentration along the Galactic plane, but also a substantial number of high-latitude sources. As noted in the legend to the plot, most of the high-latitude sources are active galactic nuclei (HBL, IBL and LBL stand for "high", "intermediate" and "low" energy BL Lac object[258], where the adjective refers to the frequency at which the synchrotron emission peaks; FRI is Fanaroff-Riley class I (low-luminosity) radio galaxy, and FSRQ is "flat spectrum radio-loud quasar"), and most of the Galactic sources are supernova remnants: a "pulsar wind nebula", PWN, is a supernova remnant powered by a young pulsar, like the Crab. The Crab Nebula itself is an extremely intense source of TeV *gamma*-rays, to the extent that it is often used as a unit for γ-ray flux ("a sensitivity of 0.01 Crab").

Some AGN observed at intermediate energies by *Fermi* are not seen at TeV energies[259]. This may be partly because e^+e^- pair production off extragalactic background light attenuates high-energy γ-rays in a manner analogous to the GZK limit (see page 57) on high-energy protons: a high-energy photon and a background photon can convert into an e^+e^- pair when

$$E_\gamma \varepsilon (1 - \cos\theta) = 2(m_e c^2)^2 \simeq 0.52 \text{ MeV}^2, \tag{2.96}$$

where E_γ is the energy of the high-energy γ-ray, ε is the energy of the background photon, and θ is the opening angle between their trajectories. This restricts the effective range of photons with energies over 1 TeV to the fairly local universe ($z < 0.1$ or so, $d < 500 Mpc$, for an optical depth of 1). The effect can be observed, for sources with redshift $z \sim 0.1$, as a steepening of the observed photon spectrum compared to expectations, as higher-energy photons are more severely affected. Unfortunately, the spectral indices of blazars span a fairly wide range, so a precise measurement (which would require prior knowledge of the unattenuated spectrum) is not possible, but estimates can be made based on plausible models[260].

A consequence of this effect is that it is not reasonable to expect observable TeV emission from candidate sources at high redshift, such as AGN at $z \sim 1{-}2$ or many gamma-ray bursts. However, this is not a complete explanation of the γ-ray sources that are not observed at TeV energies: the H.E.S.S. report on non-observation of 47 AGN seen by *Fermi*[259] notes that two of these sources still fall significantly below the extrapolated *Fermi* spectrum even after attenuation corrections. Such findings, if corroborated by more observations, help to constrain the population of fast particles in the sources.

2.5 High-energy neutrinos

Astrophysical neutrinos are expected to span a huge energy range, from the 1.9 K Cosmic Neutrino Background produced shortly after the Big Bang to PeV neutrinos associated with ultra-high-energy cosmic rays. The best-known and most well-studied astrophysical neutrinos are those produced in solar fusion reactions, which have energies from 0.4 to ~10 MeV and have been studied using a variety of detection techniques from radiochemical to water Cherenkov[261].

The neutrino flux from Supernova 1987A in the Large Magellanic Cloud, observed by the Kamiokande-II, IMB and Baksan detectors, was also the subject of intense study[262]. However, from the point of view of high-energy particle astrophysics, the most significant neutrinos are those produced by $\pi^\pm$ decay following the interaction of a high-energy cosmic-ray proton with ambient material or radiation. These high-energy neutrinos are intimately associated with the high-energy cosmic ray flux, since the population of fast protons that we observe as cosmic rays is presumably the same population that generates the neutrinos.

2.5.1 Neutrino production and the Waxman–Bahcall bound

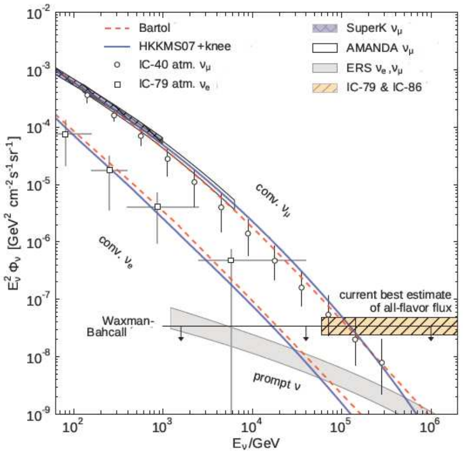


Figure 2.61: Atmospheric neutrino spectrum as estimated by models and as measured by IceCube. Figure from [264]. Note that the flux is weighted by E_ν^2.

Astrophysical neutrinos—like neutrino beams from terrestrial particle accelerators—are produced by charged pion decay. It is usually assumed that the pions are generated by *photoproduction*:

$$p + \gamma \to \Delta^+ \to \begin{cases} n + \pi^+ & (33\%) \\ p + \pi^0 & (67\%) \end{cases} \qquad (2.97)$$

where the 1:2 ratio of decay modes is a consequence of isospin[263]. This reaction can also go non-resonantly, or through resonances higher than the $\Delta^+(1232)$, but the Δ^+ dominates the cross-section. For ultrarelativistic protons interacting with relatively low-energy photons, it was shown on page 57 that the threshold for Δ^+ production is approximately

$$E_p = \frac{M^2 - m^2}{4E_\gamma}$$

where M is the mass of the Δ (1232 MeV/c^2) and m is the mass of the proton (938 MeV/c^2). For CMB photons, this implies a minimum proton energy of order 10^{20} eV, and even for optical photons with energies of around 2 eV, proton energies over 10^{17} eV are required. Although production of pions by Galactic cosmic rays is attested by the γ-ray spectrum in the Galactic plane, as shown in figure 2.42, it is likely that the bulk of detectable high-energy neutrinos will come from high-energy protons of extragalactic origin, interacting with the ambient radiation and matter in their astrophysical sources. The reason for this is that the background from atmospheric neutrinos, produced when cosmic-ray protons interact in the Earth's atmosphere (see section 1.5.4), dominates the expected astrophysical signal below energies of 10^{14} eV, as shown in figure 2.61, whereas the γ-rays from Galactic cosmic-ray-induced pion production peak around 10^9 eV.

As pion production by $p\gamma$ interactions has been measured in the laboratory (see figure 2.62), and the cosmic-ray flux at high energies has also been measured (see figures 2.1, 2.4 and 2.13), it is possible to deduce an expectation of the high-energy neutrino flux. This was first explicitly calculated by Eli Waxman and John Bahcall[266] and is known as the **Waxman-Bahcall bound**.

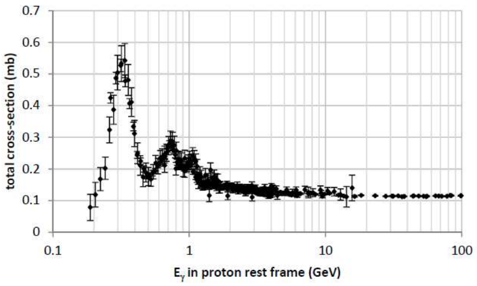


Figure 2.62: Total cross-section for $\gamma p \rightarrow X$, from the compilation by the Particle Data Group[265]. As this is based on laboratory measurements, it is plotted in terms of the photon energy in the proton rest frame. Note the large peak near threshold: this is the Δ^+ (1232) resonance. The approximately constant cross-section above 5 GeV is dominated by multipion production.

The principles of the Waxman-Bahcall bound are fairly straightforward. Following [266], we assume that the injection spectrum of high-energy cosmic-ray protons is given by

$$\frac{\mathrm{d}\dot{N}_p}{\mathrm{d}E_p} = \dot{N}_0 E_p^{-2},$$

where $\dot{N}_p$ is the number of protons per unit time and $\dot{N}_0$ is a proportionality constant. The spectral index of 2 is what we expect from diffusive shock acceleration (see next chapter); the *observed* spectrum is steeper than this, but this is a reasonable estimate for the *injected* spectrum.

The energy production rate in cosmic-ray protons of energy between E_p and $E_p + \mathrm{d}E_p$ is then

$$\dot{\mathcal{E}}(E_p)\mathrm{d}E_p = \dot{N}_p(E_p) \times E_p \mathrm{d}E_p = \frac{\dot{N}_0}{E_p}\mathrm{d}E_p.$$

If we integrate this between 10^{19} and 10^{21} eV (corresponding to those cosmic rays capable of creating pions by photoproduction off the CMB), we get

$$\dot{\mathcal{E}} = \dot{N}_0 \ln\left(10^{21}/10^{19}\right).$$

This quantity is measured, and is about 5×10^{37} J Mpc^{-3} yr^{-1}[266]. Therefore

$$\dot{N}_0 \simeq 5 \times 10^{37}/\ln(100) \simeq 10^{37} \text{ J Mpc}^{-3} \text{ yr}^{-1}.$$

Now suppose that each proton loses some fraction $\eta < 1$ of its energy in pion photoproduction before escaping from the source, and that about 25% of this lost energy goes into neutrinos (estimating that of order half the pions produced are charged, and of order half of the energy of a charged pion is carried off, when the pion decays, by neutrinos; the former is a bit of an overestimate, since π^0s are favoured if the photoproduction goes through the Δ^+, but the latter is an underestimate, since the subsequent decay of the muon produces more neutrinos). We obtain a present-day energy density of neutrinos of

$$E_\nu^2 \frac{\mathrm{d}N_\nu}{\mathrm{d}E_\nu} \simeq \frac{1}{4}\xi_z \eta t_H E_p^2 \frac{\mathrm{d}\dot{N}_p}{\mathrm{d}E_p}, \tag{2.98}$$

where t_H is the Hubble time ($\sim 10^{10}$ years) and ξ_z is an evolution factor introduced to account for the redshift of neutrino energies from distant sources and the possible evolution of $\dot{N}_p$ over cosmic time—for example, if (as seems likely) AGN are the sources of extragalactic cosmic rays, we expect that the rate of cosmic-ray production would have been higher at early times, because the number density of AGN is much greater at $z \sim 2$ than at the present time. Waxman and Bahcall[266] attempt to estimate ξ_z, and conclude that it lies between about 0.6 (no change in $\dot{N}_p$) and 3 (evolution of $\dot{N}_p$ similar to evolution of star formation rate). The observed flux (measured per steradian) is $c/4\pi$

times the energy density (the c comes from the speed of the neutrinos and the 4π from solid angle). Putting in the numbers, and converting to units more suitable for particle spectra, we obtain

$$E_\nu^2 \Phi_{\nu_\mu} \simeq \frac{c}{4\pi} E_n u^2 \frac{\mathrm{d}N_{\nu_\mu}}{\mathrm{d}E_\nu} \simeq \xi_z \eta \times 10^{-4} \text{ GeV m}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \tag{2.99}$$

as an order of magnitude estimate. Neutrino oscillations will ensure that the fluxes of all neutrino types are more-or-less equalised as they travel over cosmic distances, so $\Phi_{\nu_\tau} \simeq \Phi_{\nu_\mu} \simeq \Phi_{\nu_e}$. For an upper bound we set $\eta = 1$ and $\xi_z = 3$, getting $E_\nu^2 \Phi_\nu \simeq 3 \times 10^{-4}$ GeV m^{-2} s^{-1} sr^{-1} as shown in figure 2.61. The currently measured value[267] is consistent with 10^{-4} GeV m^{-2} s^{-1} sr^{-1} per flavour, though the statistical errors are still very high (only 37 events observed, with an estimated background of 8.4 ± 4.2 cosmic-ray muons and $6.6^{+5.9}_{-1.6}$ atmospheric neutrinos).

2.5.2 Interaction of high-energy neutrinos with matter

Neutrinos interact only via the weak interaction (and gravity, but gravity is negligible on the scale of particle interactions). The weak interaction is so called because it *is* weak, at least at the energy scales appropriate to particle decays; this is because the weak interaction carriers, the W and Z bosons, are massive, and therefore highly virtual at such energy scales.

As the energy of the neutrino increases, the W mass becomes less of a problem, and consequently the neutrino-nucleon interaction cross-section is approximately proportional to the neutrino energy, as shown in figure 2.63. Nevertheless, the mean free path for a neutrino of energy $\sim 10^{15}$ eV in water is

$$\ell = \frac{1}{n\sigma} \simeq \frac{1.7 \times 10^{-27}}{1000 \times 10^{-37}} = 1.7 \times 10^7 \text{ m},$$

where $n = \rho/\mu$ is the number density of nucleons in water, $\rho = 1000$ kg m^{-3} is the density of water, $\mu = 1.7 \times 10^{-27}$ kg is the mass of a nucleon, $\sigma = 10^{-37}$ m^2 is the cross-section, and the numerical result is about 1.3 times the diameter of the Earth. It is therefore clear that a very large detector is required in order to acquire useful statistics: even the IceCube detector, with an instrumented volume of about 1 km^3 (therefore containing about 6×10^{38} nucleons) has an effective area of only about 60 m^2 as a telescope for neutrinos of energy around 10^{15} eV. Since the expected flux of neutrinos from equation(2.99) is only about 0.4 per square metre per year over the energy range from 10^5 to 10^7 GeV, it is not surprising that only a small number of events have been recorded to date.

Although the steady increase of νN cross-sections with energy is clearly visible, the most striking feature of figure 2.63 is the sharp peak in the $\bar{n}u_e e$ cross-section. This is the **Glashow resonance**, caused by the production of a real W^- boson in the reaction $\bar{\nu}_e + e^- \to W^- \to \bar{\nu}_X + \ell_X$, where X can be any lepton flavour (e, μ or τ). The expected energy of the Glashow resonance is easy to calculate. Assuming that the electron is initially stationary, that the mass of the neutrino is negligible, and that $m_\ell \ll E_\nu$, and taking $c = 1$ as is usual in particle physics calculations, we have, from $E^2 = p^2 + m^2$,

$$M_W^2 = (E_\nu + m_e)^2 - E_\nu^2 \simeq 2E_\nu m_e,$$

which gives

$$E_n u = \frac{80.4^2}{2 \times 0.511 \times 10^{-3}} = 6.3 \times 10^6 \text{ GeV},$$

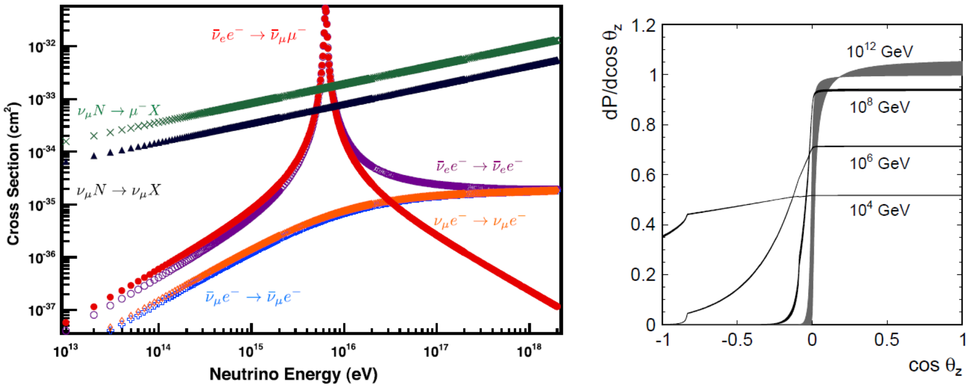


Figure 2.63: Interaction cross-section for high-energy neutrinos. Left panel, cross-sections for neutrino-nucleon (νN) and neutrino-electron (νe) interactions, separated by flavour, from [268]. Note the Glashow resonance, caused by formation of a real W boson, in $\bar{\nu}_e e$ scattering. Right panel, survival probability of neutrinos of various energies as a function of zenith angle $\cos\theta_z$, from [269]. Positive $\cos\theta_z$ corresponds to down-going neutrinos, and negative $\cos\theta_z$ to up-going neutrinos. The noticeable kink at $\cos\theta_z \simeq -0.8$ is caused by the Earth's core, which is significantly denser than the mantle.

or 6.3×10^{15} eV, as shown in the figure. At the peak of the Glashow resonance, the $\bar{\nu}_e e^-$ cross-section is about a factor of 30 higher than non-resonant cross-sections; this should produce a noticeable rise in the overall event rate if the flux of antineutrinos is comparable to the neutrino flux. However, if the pions whose decays create the neutrinos are produced primarily by $p\gamma$ photoproduction on low-energy photons, we do *not* expect this to be the case: as the proton is positively charged, π^+ (which decay to neutrinos) should be produced more often than π^- (which decay to antineutrinos). The situation is less clear if the pions are produced by interactions with other nuclei: in this case, multipion production is more common, and positive and negative pions are produced at nearly equal rates. Hence, observation (or significant non-observation) of the Glashow resonance would provide important information on the environment in which cosmic rays are accelerated[270].

Although the interaction cross-section is small enough to make detection of high-energy neutrinos a challenge, it is large enough to be significant in terms of neutrino absorption in bodies such as the Earth: a mean free path of 1.3 Earth diameters *in water* corresponds to only about half the Earth's radius in the Earth itself, given that the mean density of the Earth is about 5.5 times that of water. This is shown in the right panel of figure 2.63, which shows the survival probability for neutrinos of various energies as a function of $\cos\theta_z$, the angle away from the zenith. From this figure[269], it is clear that neutrinos of energies 10^{15} eV and above are essentially completely absorbed by the Earth, and can only be detected when they enter the detector from above—the frequently-quoted statement that it would take about a light-year of lead to stop a neutrino is true at low energies, where the interaction cross-section is only of order 10^{-44} m^2[268], but not at the energies of interest here. This is unfortunate, because it means that the astrophysical signal has to be distinguished not only from the cosmic-ray *neutrino* flux (atmospheric neutrinos, see section 1.5.4), but also from the cosmic-ray *muon* flux, which is negligible for upgoing neutrinos because the Earth acts as a very effective shield. On the other hand, the increasing cross-section opens up new avenues for the detection of ultra-high-energy neutrinos: for example, the potentially very cost-effective technique of acoustic detection[271] relies on the fact that ultra-high-energy neutrinos have a mean free path in water short enough that they induce an electroweak shower on impact with the ocean.

2.5.3 Detection of high-energy neutrinos

Because of the low predicted fluxes, detection of high-energy neutrinos requires very large detectors. Atmospheric (and accelerator-generated) neutrinos of 1 GeV to 1 TeV energy are typically detected using large water Cherenkov detectors such as Super-Kamiokande[272], but these facilities are not large enough to collect useful statistics at higher energies. The best strategy to date has been to retain the water Cherenkov technique, but to replace artificial water tanks with natural bodies of water: Lake Baikal[273], the Mediterranean Sea[274] and the Antarctic icecap[44]. These experiments are all very similar in design and physical principles: the IceCube experiment is the largest and, to date, the only one to report a signal, so we will focus on IceCube as the "type" example.

In contrast to atmospheric and solar neutrino experiments such as Super-Kamiokande and SNO[275], which have a volume of water surrounded by closely packed photomultiplier tubes to reconstruct the Cherenkov ring produced by a charged lepton, the large-volume neutrino telescopes instrument the entire volume fairly sparsely using "strings" of PMTs (encased in suitable pressure-resistant housings). The PMTs on each string are separated by $\sim$15 m; the separation between strings is 125 m for the main IceCube detector, though ANTARES[274] and the DeepCore subdetector of IceCube[276] have closer spacings of 60 and 72 m respectively, for a lower energy threshold. High-energy muons travel hundreds of metres in ice or water, radiating Cherenkov light as they do so; the arrival time of the light at different PMTs can be used to reconstruct the track direction to better than 1°, ANTARES performing slightly better than IceCube because the light scatters less in water than in ice. For interactions that do not create a muon, IceCube observes the particle shower produced at the interaction point, caused by secondaries from the struck nucleus and/or the electromagnetic shower left by the produced $e^\pm$ (in the case of ν_e or $\bar{\nu}_e$ charged-current interactions). The visible energy of such interactions can be reconstructed quite well (though in the case of neutral-current, i.e. Z exchange, interactions it will underestimate the incident neutrino energy, because the final-state neutrino is not detected), but the angular resolution is poor, typically around 15°[267].

In order to minimise the background from down-going cosmic-ray muons, neutrino telescopes are deployed deep below the surface of the ice or water: both IceCube and ANTARES are located at depths of about 1.5 to 2.5 km beneath the surface, providing 1.5 km of water shielding. The Baikal neutrino telescope is deployed at a somewhat shallower depth, simply because Lake Baikal is only 1600 m deep. IceCube additionally has the IceTop surface array[112], which can be used in anticoincidence with IceCube to study neutrinos or in coincidence to study cosmic rays; a surface array of this kind is not practical for ANTARES or Baikal because the surface is liquid water for at least part of the year (Lake Baikal does freeze over in winter). Analyses may require the event to be contained within the detector volume, with no hit PMTs in the uppermost layers of the experiment, to reject muons that are not stopped by the overburden.

Searches for astrophysical neutrinos can be done in two ways: searching for a diffuse neutrino flux, identified as an excess of high-energy events over the more steeply falling astrophysical neutrino background,[6] or searching for signals from specific point sources, identified by an excess of neutrinos from within the point spread function of the source. Both types of searches have been carried out, but so far only the diffuse flux analysis has produced a positive result, and only from IceCube (much the largest of the three operating natural-water neutrino telescopes).

IceCube selects[267] events with at least 6000 photoelectrons detected (corresponding to a threshold energy of about 30 TeV deposited in the detector), with not more than 3 of the first 250 photoelectrons detected lying on the boundary of the detector (a veto against cosmic muons coming in from outside). The same cut is applied to all boundaries, even though cosmic muons are only expected to come from above, in order to ensure that the selection is not biased in terms of the direction of the incoming neutrino. It rejects about 99.999% of high-energy cosmic muons (the efficiency is determined by defining an artificial "boundary layer" within the body of IceCube and seeing how many cosmic muons it identifies). The cosmic-ray veto also reduces the down-going atmospheric neutrino background, because the air shower that produces the neutrinos will also generate muons which may trigger the veto. (It will not suppress the up-going atmospheric neutrino background, because the muons will obviously not penetrate the Earth!)

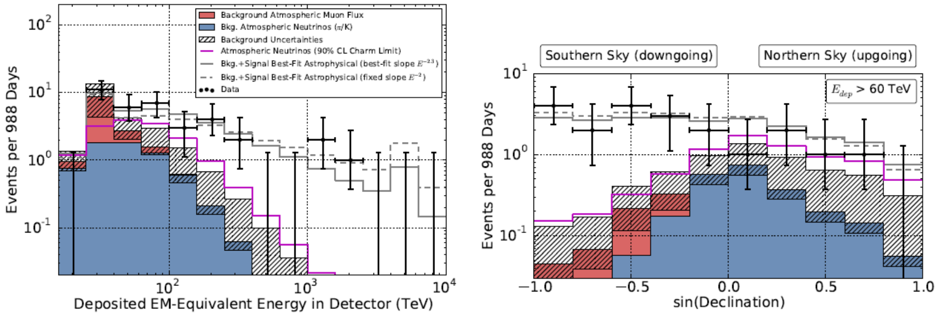


Figure 2.64: High-energy neutrinos observed by IceCube[267]. The left panel shows the energy spectrum, with a clear excess of events above background (shaded histogram) at energies higher than 100 TeV. The Glashow resonance is shown in the best-fit signal spectrum (unshaded histogram) as a high bin at the appropriate energy: there is no evidence for it in the data, but the absence is not statistically significant at present. The right panel shows the zenith angle distribution (declination $\delta = 90° - \theta_z$, so sin(declination) $= \cos\theta_z$) for events with $E > 60$ TeV: as predicted above, the signal is mostly down-going, while the up-going events are mostly atmospheric neutrino background.

The event selection identifies 37 events, one of which can only be interpreted by assuming that it records not one single particle, but two coincident muons from different directions (i.e. not part of the same air shower). This event is certainly part of the cosmic-ray muon background—there were matching hits in the IceTop array, but they did not veto the event because they were below the threshold energy for well-reconstructed data. Several other events with

[6] You may wonder why the atmospheric neutrino spectrum falls off more steeply than the astrophysical neutrino spectrum, when both sets of neutrinos are generated by cosmic ray interactions from (presumably) the same seed cosmic ray population. The reason is that as the energies of the produced pions increase, their lifetime in the lab frame is increased by time dilation. In the case of atmospheric neutrinos, it then becomes increasingly likely that the pion will interact strongly with another nucleus before it has time to decay, thereby suppressing the production of high-energy neutrinos. This would not happen in a lower-density environment.

reconstructed muon tracks are also suspected of being cosmic-ray background: the efficiency study mentioned above predicts a total background rate from this source of 8.4 ± 4.2 events, along with an atmospheric neutrino background of $6.6^{+5.9}_{-1.6}$ events.

Figure 2.64 shows the energy spectrum and zenith angle distribution of the selected events. The apparent "hole" in the energy spectrum, with three events between 1000 and 2000 TeV but none between 400 and 1000, is not statistically significant—[267] report that simulations of a sample of this size drawn from the best-fit histogram produce gaps at least this large nearly half the time (43% of all cases). There is no evidence of an enhancement at the Glashow resonance, but again this absence is not statistically significant at this time. It does indicate that the proportion of neutrinos that are $\bar{\nu}_e$ is not much *larger* than anticipated, and so may suggest that $p\gamma$ pion production dominates over pp; on the other hand, [267] contend that the observed spectrum is more consistent with pp production than with $p\gamma$, where a more peaked spectrum might be expected from the combination of a thermal photon spectrum (which is peaked) and production via the Δ^+ resonance (which prefers a specific centre-of-mass energy). The safest conclusion is probably to argue that more statistics are needed before anything useful can be said. The zenith angle distribution shows that, as expected, the signal is mostly in down-going events.

The spatial distribution of the signal events, shown in figure 2.65, is consistent with an isotropic parent population. There is an excess of events close to the Galactic centre, but it is not statistically significant[267], and no corresponding excess is seen by ANTARES[277]. The lack of any tendency to concentrate along the Galactic plane *is* significant, and suggests that at least some of the sources are extragalactic (compare figure 2.65 with figure 2.60).

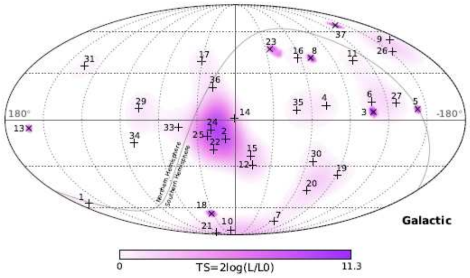


Figure 2.65: Map of the arrival directions of the IceCube high-energy neutrino events: + indicates a shower-like event (angular resolution ∼15° and × a track event (angular resolution $\leq 1°$, but a number of these may be cosmic muon background). The map is in Galactic coordinates; the celestial equator is shown as the thin line running just to the left of the most significant concentration of events.

The lack of clustering in this signal, coupled with negative results from targeted point-source searches, suggests that the neutrino flux is coming from a large number of individually faint sources, so that no single source is contributing enough neutrinos to make a significant cluster. This is perfectly possible, especially if most of the sources are extragalactic. Unlike high-energy photons and protons, neutrinos are not significantly attenuated by interactions with intergalactic matter or radiation, so the diffuse neutrino flux could be coming from very high redshifts—this is the reason for the t_H in equation (2.98)—in contrast to photons and protons which must be fairly local. Hence a very large number of very faint sources is entirely plausible, if not particularly encouraging—we would obviously rather have a small number of powerful sources that we could positively identify.

Transient sources of high-energy neutrinos might be identifiable even in the absence of a clear spatial clustering, if neutrinos were seen to arrive at appropriate times: even if the number of neutrinos from any individual source was not statistically significant, an excess of neutrinos arriving in coincidence with transient sources could be built up by multiple observations. The obvious candidate for association with transient neutrino sources is gamma-ray bursts, which are regarded as possible sources of cosmic rays and would, if this were correct, also be expected to produce neutrinos.

IceCube carried out a search for such an association based on the first two years of IceCube data[278], and found none. ANTARES also conducted an unsuccessful search[221], although their limit is less restrictive owing to their smaller volume. The IceCube limit constrains some models of GRBs as cosmic-ray sources, although the strength of the constraint is model-dependent: on the basis of the original Waxman-Bahcall analytical estimate[266], there is a conflict, but on the basis of a recent Monte Carlo calculation of the neutrino spectrum[280], there is not (see figure 2.66 from [221]). These limits, particularly that from IceCube, are close enough to the predictions to suggest that a few more years' data should resolve this issue.

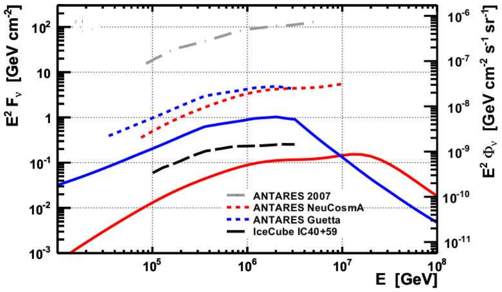


Figure 2.66: Limits on association of neutrinos with GRBs[221]. The blue and red solid lines represent the predicted neutrino signal using an analytical model[279] and the NeuCosmA Monte Carlo code[280] respectively. The blue and red dotted lines are the ANTARES upper limit, calculated using the two different model spectra. The black dashed line is the IceCube limit from [278], which rules out the blue model spectrum but not the red. The grey dash-dotted line is an earlier limit from ANTARES.

2.5.4 Future prospects

It is worth concluding this section with a brief summary of future prospects for high-energy neutrino astrophysics (see also section 1.5.5). The main problem of this field is the need for large active volumes, and associated high cost. Feasibility studies have been carried out for much larger detectors in both the Mediterranean Sea (KM3NeT[281]) and in Lake Baikal (NT-1000[273]), but it is not obvious that the funds needed to construct these can be obtained in a difficult economic climate. Other techniques such as the Askaryan effect[45] and acoustic detection[271] are potentially more cost-effective, since they can instrument larger target volumes more cheaply, but have even higher energy thresholds than the large water Cherenkovs. The medium-term future for this branch of particle astrophysics is therefore uncertain. Nevertheless, the successful observation of astrophysical neutrinos by IceCube—including one extremely high-energy event, with a visible energy of 2000 TeV, which is encouraging for alternative detection techniques—has at least demonstrated that the signal does exist, at about the predicted rate, and can be identified. IceCube would also be an invaluable detector in the event of a core-collapse supernova in our Galaxy: although it is not equipped to measure the properties of low-energy neutrinos well, it has a low enough noise level to detect the signal easily as a large increase in individual hit PMTs, and could study its time evolution over the course of the burst.

2.5.5 Summary

The detection of high-energy astrophysical neutrinos is difficult. The cross-sections are small enough to require an extremely large detector, yet large enough that the established neutrino-detection strategy of using the Earth as a shield against cosmic muons fails in the most favourable energy range. There are intractable backgrounds, particularly from atmospheric neutrinos, which are genuine neutrinos and therefore impossible to reject. Except in the case of charged-current ν_μ interactions, which produce a muon track with a well-reconstructed direction, the angular resolution of reconstructed events is very poor. Nevertheless, several large-volume water Cherenkov neutrino telescopes have been built, and in the past couple of years the largest of them, IceCube, has finally presented good evidence for high-energy astrophysical sources, albeit as a diffuse flux with no identified point sources.

Despite the difficulties, high-energy neutrino astrophysics has some unique features. Any astrophysical object identified as a point source of neutrinos is necessarily accelerating hadrons, and is therefore a source of charged cosmic rays. In principle, neutrinos have a much longer range than ultra-high-energy charged cosmic rays or TeV γ-rays, because they are not degraded by interaction with the cosmic microwave background—though in practice such distant sources are not likely to contribute more than a diffuse background to the neutrino sky. It is also possible that neutrino signals could be detected from astrophysical accelerators that are opaque to the accompanying γ-ray signal. In some respects, the opening of the neutrino window for astronomy has been disappointing—the bright, unexpected new types of source that were found on the opening of other windows, such as radio and γ-rays, have not presented themselves this time—but it may yet provide important insights.

2.6 An overview of sources

We will discuss the physics of astrophysical sources of high-energy radiation and particles in the next two chapters. However, at this point it is useful to summarise the various types of astrophysical object that have been identified in this chapter. We should recall here that, at present, the only way of identifying a source unambiguously is through its photon emission: charged cosmic rays lose directional information as a result of deflection in the Galactic magnetic field, and high-energy neutrinos have so far only been detected as a diffuse flux, with no identified point or extended sources.

From the viewpoint of this chapter, relevant astrophysical sources are those that accelerate charged particles to high energies. The mechanisms by which they do this will be explored in the next chapter. Other sources of interest to particle astrophysics are low-energy neutrino emitters (the Sun and supernovae), since neutrino production is always associated with "particle physics" processes, and possible signals from concentrations of dark matter, e.g. intermediate-energy (tens of GeV) neutrinos from $\chi\chi$ annihilation in the Sun, or γ-ray signatures from the vicinity of the Galactic centre. These are not *high energy* particle astrophysics, and will not be considered here.

The mechanisms by which fast particles are associated with photon emission are often interrelated: a population of relativistic electrons in a magnetic field will produce radio and X-ray photons by synchrotron radiation, and γ-ray photons by synchrotron-self-Compton emission. Sources of charged cosmic rays should also be sources of high-energy neutrino emission, from pion production in $p\gamma$ or pp interactions; but, unless the source is extremely optically thick, they should also be associated with γ radiation since the same interactions that produce $\pi^{\pm}$ also yield π^0. Therefore, it is not surprising that the same objects have recurred multiple times in different sections of this chapter. It is in fact very unusual for an astrophysical object to be associated with only one type of non-thermal emission[7].

Sources of non-thermal photon emission can be Galactic or extragalactic, as clearly shown in figure 2.60, and either continuous (though often variable) or transient. Many have associations with compact objects (neutron stars or black holes), but some do not (the remnants of Type Ia supernovae). Essentially all non-thermal sources have some component which can be modelled by synchrotron radiation, implying the presence of magnetic fields. This is significant: we shall see in the next chapter that magnetic fields are required by our models of particle acceleration.

Figure 2.60 provides a convenient checklist of source types. Examples of most source types seen in other wavebands emit TeV γ-rays, although not every example of a given source type will be represented (e.g. some but not all blazars are TeV sources, as are some but not all radio galaxies; most X-ray binaries are *not* TeV sources, but some are). It should be noted that transient sources are not represented on the sky map.

2.6.1 Galactic sources

Galactic sources fall into two main categories: those associated with compact objects (neutron stars and perhaps stellar-mass black holes), and gaseous supernova remnants (obviously, pulsars are also “supernova remnants” in a sense, but the term is usually used to refer to the expanding gas cloud).

Pulsars and pulsar wind nebulae

Pulsars, spinning neutron stars observed by way of a “lighthouse beam” of emission from their magnetic poles, are observed to slow down over time; i.e. they are losing their rotational kinematic energy. This is believed to occur by way of a relativistic wind composed mainly of electrons and positrons accelerated to very high energies[282]. The presence of fast electrons and a magnetic field sets up the appropriate conditions for synchrotron radiation (see section 2.3.5). If the wind is confined, for example by the more slowly expanding gas shell of a core-collapse supernova, the result is a pulsar wind nebula (PWN). Pulsar wind nebulae (PWNe) are the most common class of Galactic TeV γ-ray source; the Crab Nebula, the first-discovered TeV source and still the classic test target for Cherenkov telescopes, is a PWN. Pulsar wind nebulae are usually found in young supernova remnants—the Crab is the remnant of a supernova observed by the Chines in AD 1054—but can also be generated when pulsars with high space velocities interact with the interstellar medium[282]. Gamma-ray binaries (see below) have been described[283] as “pulsar wind nebulae in a binary environment.”

Figure 2.67 shows the spectral energy distribution (νF_ν, where F_ν is the flux at frequency ν) of the Crab Nebula, the prototype PWN. The spectrum extends from radio frequencies up to TeV γ-rays, and is well described by a model in which the photons are produced by synchrotron radiation and inverse Compton

[7] Since free-free emission is thermal in nature, it *is* possible for sources to emit only thermal X-rays, and not radio emission or γ-rays—the hot intracluster medium of rich clusters of galaxies is an example of this.

scattering. The average magnetic field required by the fit is $124 \pm 6^{+15}_{-6}$ μG (12.4 nT), where the first error is statistical and the second systematic[284].

The means by which PWNe accelerate electrons to very high energies is still not well understood. The presence of magnetic fields and shock fronts would initially suggest appropriate conditions for diffusive shock acceleration (see later), but in fact a magnetised relativistic shock is not a good site for particle acceleration by this mechanism[285], and the problem is still under study. It is also not clear whether, or to what extent, protons and positive ions are accelerated in addition to $e^{\pm}$.

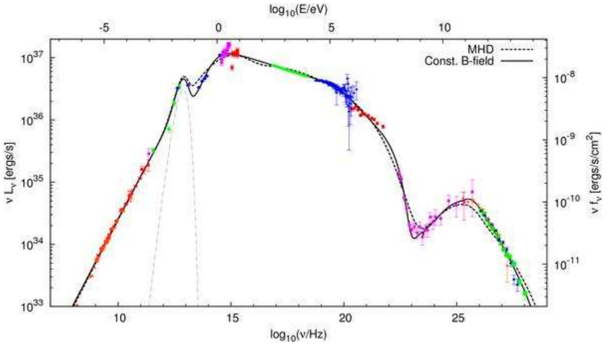


Figure 2.67: Spectral energy distribution for the Crab Nebula[284], showing emission from radio to TeV energies. See [284] and references therein for the data sources. The data are well described by a magnetohydrodynamic (MHD) model and by a simplified model assuming a constant magnetic field (see [284] for details).

Supernova remnants

High-energy γ-ray emission from supernova remnants (SNRs) is not necessarily associated with PWNe. A clear counterexample is the remnant of Tycho's supernova, securely identified as a Type Ia from spectroscopy of its "light echo"[286]: SNe Ia do not leave a compact remnant (the explosion disrupts the entire star) but the supernova remnant associated with Tycho's SN is a TeV γ-ray source.

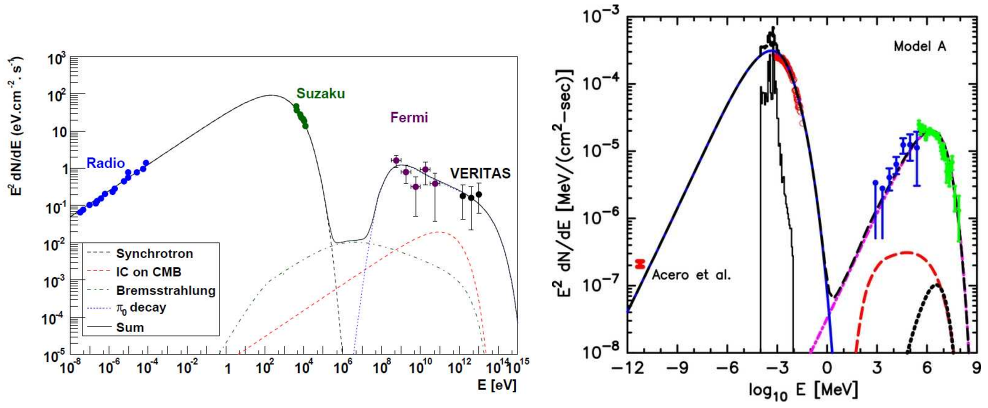


Figure 2.68: Spectral energy distributions for SNRs. Left panel, Tycho's supernova (SN 1572)[287]. Models suggest that π^0 decay dominates the high-energy γ-ray production in this SNR. Right panel, RX J1713.7–3946[288]. In this SNR, inverse Compton emission appears to dominate the high-energy γ-ray production, although the authors argue that "the efficient production of CR ions is an essential part of our leptonic model."

On the general grounds of their size and magnetic fields, SNRs have long been postulated as the main sites for Galactic cosmic-ray acceleration, at least up to energies of 10^{15} eV. In this context, it is encouraging that the spectral energy distribution of Tycho's supernova strongly suggests that the GeV and TeV γ-ray flux is coming from π^0 decay and not from inverse Compton (see figure

2.68). On the other hand, some SNRs, such as RX J1713.7–3946 (right panel of figure 2.68), have GeV–TeV photon emission which is dominated by inverse Compton, and *not* by π^0 decay. However, models of these lepton-dominated SNRs may still imply the acceleration of positive ions[288].

X-ray binaries

X-ray binaries are close binary systems in which material from a companion star is being accreted on to a compact object (neutron star or black hole). The accreted material is heated up by friction and emits thermal X-rays. Most XRBs are not really in the domain of particle astrophysics, but some are known to emit much more energetic radiation. "Microquasars" are XRBs where the compact object emits relativistic jets from its poles, analogous—hence the name—to the jets emitted by the supermassive black holes in radio galaxies.

Only a small minority of XRBs (five, to date) have been observed to emit high-energy (GeV–TeV) γ-rays. They are all "gamma-ray binaries"[283], much brighter in γ-rays than elsewhere in the electromagnetic spectrum (three of the five were discovered in γ-rays). All have massive O or B0e class main-sequence companion stars. In one case, the compact object is known to be a neutron star, because it is a radio pulsar with a period of 47.76 ms; in the other four cases, the compact object is generally assumed to be a neutron star but could, if the orbit is close to face-on to our line of sight, be a small stellar-mass black hole. None of the confirmed gamma-ray binaries is *known* to contain a black hole[8]. In all cases, the γ-ray emission is modulated over the orbital period, which confirms the identification of the source as a binary.

The gamma-ray binaries are all X-ray sources, as expected from binary systems with accretion on to a compact object. They are also all radio sources, which is *not* a normal feature of high-mass X-ray binaries, but is consistent with the presence of the non-thermal γ-ray emission; the properties of the radio emission are consistent with synchrotron radiation.

The two possible models for γ-ray emission from these systems[283] are pulsar wind nebulae (the wind from the pulsar being confined by the strong stellar wind from the O/Be companion) or accreting microquasars. The established presence of a pulsar in PSR-B1259–63 suggests the former scenario for this system, and the overall similarity of the five systems would tend to prefer the same model for all of them. The similarity of the properties of gamma-ray binaries and PWNe is further indirect evidence in support of this interpretation[283].

Diffuse Galactic emission

In addition to Galactic point sources, *Fermi*–LAT and EGRET detected a diffuse flux of intermediate-energy γ-rays from the Galactic plane. This is known to be dominated by π^0 decay (see section 2.4.2 above), and is presumably caused by Galactic cosmic rays interacting with ambient material or background light. At lower energies, bremsstrahlung and synchrotron radiation from Galactic cosmic-ray electrons also contribute.

This diffuse emission does not emanate from the *sources* of the Galactic cosmic rays, and therefore is of limited value in the context of high-energy particle astrophysics. It does imply that we should expect nearby galaxies to

[8] A TeV γ-ray flare coincident with the well-known black-hole binary Cygnus X-1 was detected by MAGIC in 2007[257], but only at 4.1σ significance. This is the only detection, so clearly Cyg X-1 is at best an episodic γ-ray source.

be weak sources of high-energy emission, even in the absence of active accretion around their central supermassive black holes.

2.6.2 Extragalactic sources

Extragalactic sources dominate the *Fermi*–LAT point source catalogue at GeV energies, and are responsible for slightly less than half of the identified TeV sources. The vast majority of continuous extragalactic γ-ray sources are blazars or closely related objects.

Blazars

Blazar is a composite word combining *BL Lac*[258] and *quasar* (with the spelling presumably influenced by "blaze"). Most quasars do not qualify as blazars: the class only includes the category usually known as *flat spectrum radio quasars* (FSRQs).

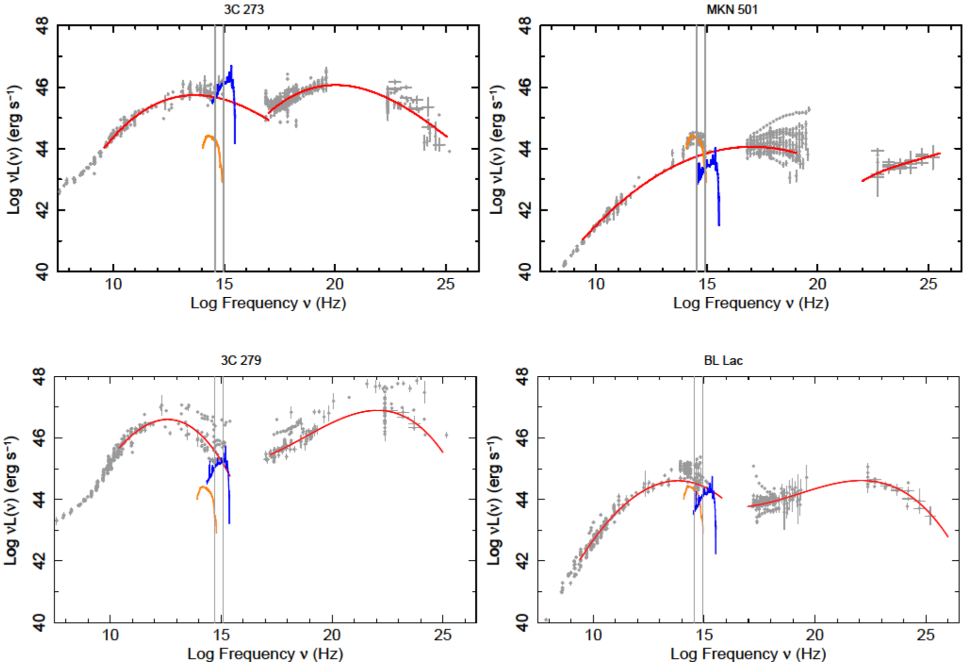


Figure 2.69: Spectral energy distributions for blazars, from Giommi et al.[289]. Left panels, two FSRQs; right panels, two BL Lacs. The emission is dominated by the jet (red), with the typical quasar emission from the accretion disc and broad line region (blue) and the elliptical host galaxy (orange) contributing comparatively little, especially in the BL Lacs. Note the different positions of the peak synchrotron frequency: Mkn 501 is an HBL, whereas BL Lac itself is borderline between an LBL and an FSRQ (emission lines can sometimes be seen in BL Lac's spectrum, when the continuum is in a low state). The vertical lines on the plots indicate the optical region (380–800 nm).

Blazars are compact, radio-loud objects with radio spectral indices ≤ 0.5 (hence "flat spectrum"). They are characterised by strong, irregular variability and highly polarised emission at radio and optical wavelengths. Their continuous spectra extend from radio wavelengths up to γ-rays, with a double-humped shape characteristic of synchrotron plus inverse Compton emission[289] (see figure 2.69. Historically, BL Lac objects are characterised by weak or absent emission lines, whereas FSRQs have the strong emission lines characteristic of quasar optical spectra[290]; the usual rule of thumb is that BL Lacs have no emission lines with rest-frame equivalent width > 5 Å (0.5 nm). BL Lacs are further subdivided according to the peak frequency of their synchrotron emission, ν_S: "high" for $\nu_S > 10^{15}$ Hz (UV or higher), "intermediate" for $10^{14} < \nu_S < 10^{15}$ Hz (optical/near IR) and "low" for $\nu_S < 10^{14}$ Hz (IR); the lowest peak frequencies are around $10^{12.5}$ Hz (100 μm), in the far IR[289]. These subdivisions are somewhat artificial, as the ν_S distribution is continuous from low to high—an earlier impression that there were two distinct categories, high and low, was created by selection effects.

In contrast, the distinction between BL Lacs and FSRQs is real, although the somewhat arbitrary equivalent-width selection criterion is probably not the best way to make it. Compared to BL Lacs, FSRQs have higher radio luminosities and different radio shapes, tend to occur at higher redshifts (compared to those BL Lacs with measured redshift; it should be noted that about half of all BL Lacs do not have measured redshifts owing to the lack of any spectral lines to measure the redshifts of), and generally have lower synchrotron peak frequencies. As a result of the last point, the relative proportions of BL Lacs and FSRQs in any given sample depend strongly on the selection criteria: X-ray selected samples favour high ν_S, and so have much higher fractions of BL Lacs than radio-selected samples.

It is generally accepted that blazars are active galactic nuclei where we are looking more or less straight down the axis of the radio jet. As a result of relativistic beaming (see section 2.3.5), this amplifies the emission from the jet itself relative to the unbeamed emission from the rest of the object, resulting in the smooth, synchrotron-dominated blazar spectrum. The "parent population"—that is, the objects we see when the jet axis is *not* aligned to our line of sight—are the classical double-lobed radio galaxies. In this picture, the BL Lacs are beamed versions of the low-luminosity Fanaroff-Riley class I (FR-I) radio galaxies[291], while the FSRQs are the beamed equivalents of the more luminous FR-II radio galaxies.

Other AGN

Since most AGN are radio-quiet, and only properly aligned radio-loud AGN will be observed as blazars, it follows that blazars must make up a fairly small minority of AGN (~5% is often quoted). However, as can be seen from figure 2.60 and the *Fermi*–LAT 2-year catalogue[232], blazars make up the vast majority of AGN detected in γ-rays. This is largely because of the beaming effect: the apparent luminosity of a beamed source is highly amplified by the beaming, and an unbeamed compact source of comparable luminosity would be optically thick to high-energy photons. However, it is noteworthy that radio-quiet equivalents of BL Lacs do not seem to exist: the existence of a relativistic jet seems to be highly correlated with radio emission. It therefore seems likely that radio-quiet AGN are not associated with acceleration of particles to high energies, or at least not on anything like the scale of radio-loud AGN.

As noted above, blazars are generally agreed to be the same objects as FR-I and FR-II double-lobed radio galaxies, but seen in a particular orientation. It follows that all FR-I and FR-II radio galaxies should be viewed as sites for particle acceleration. Steep-spectrum radio quasars (SSRQs) are seen in unified models[292] as intermediate in orientation between FSRQs and FR-II radio galaxies, and therefore also qualify.

In short, it seems fair to conclude that all radio-loud AGN are accelerating particles to high energies and are potential sources of cosmic rays (and neutrinos). Radio-quiet AGN, while equally energetic in terms of overall energy output, do not seem to host significant jet activity and are probably not particle accelerators. The only radio-quiet AGN observed at TeV energies are starburst galaxies, and the TeV emission is most likely associated with the starburst rather than the AGN engine.

Starburst galaxies

Starburst galaxies, as their name indicates, are galaxies that are forming stars at an unusually high rate. High rates of star formation imply a higher than normal proportion of short-lived massive stars, and therefore a larger than usual number of supernovae. Young supernova remnants are the dominant sources of γ-ray emission, and probably of cosmic rays, within our own Galaxy. Therefore, we might reasonably expect that starburst galaxies would be detected as γ-ray sources. This turns out to be the case: a small number of nearby starburst galaxies have been detected at GeV energies by *Fermi*–LAT[293], and two of these (M82 and NGC 253) have also been detected at TeV energies, by VERITAS and H.E.S.S. respectively. Based on scaling relations, the *Fermi*–LAT Collaboration infer[293] that between 4 and 23% of the diffuse GeV γ-ray background could be contributed by unresolved starburst galaxies. Unlike the AGN emission, the high-energy emission from starbursts is qualitatively similar to Galactic emission—just scaled up—and comes from similar sources. Starbursts are not likely to be the sources of the highest-energy cosmic rays, and are probably not good targets for point source neutrino searches.

2.6.3 Transient sources

Most sources of high-energy emission are variable, and it is possible that faint sources detectable only in very rare high-intensity flares might register as transient. A possible example of this is Cygnus X-1, which is not normally an emitter of TeV γ-rays, but which was possibly detected (at 4σ significance) once by MAGIC. However, the only incontestably transient sources of high-energy emission are the gamma-ray bursts (GRBs; see also page 100).

The defining feature of GRBs is the short initial burst of intense luminosity, dominated by low-energy γ-rays of energies around 1 MeV. They are conventionally divided into "long" and "short" bursts, according to whether this initial burst does or does not (respectively) last longer than 2 s. As with blazars, the conventional boundary is somewhat arbitrary, and may not be the best possible classifier, but there is no doubt that the two classes are real and represent different phenomena. Long bursts have softer spectra than short bursts (the two classes are often called "long-soft" and "short-hard" in recognition of this), and are exclusively located in star-forming galaxies, whereas short bursts sometimes occur in elliptical galaxies[223]; long bursts are clearly associated with overluminous Type Ibc supernovae (see below), whereas short bursts definitely lack such an association: no supernova has been found for any of the short GRBs with redshifts small enough that the SN would have been detectable if present[223]. It is worth noting that two nearby "long" GRBs, GRB 060505 and GRB 060614, also have no associated supernova: these may represent cases for which the simple 2 s selection criterion is inadequate. Short GRBs with measured redshifts are more local than long GRBs (median redshifts of *Swift* samples $\sim$0.5 and $\sim$2 respectively), though some of this is a selection effect because (a) short GRBs are also less luminous on average and (b) owing to redshift, some distant GRBs that are classed as "long" would be "short" in their own rest frames.

A minority of bursts of both types are detected at higher energies by *Fermi*–

LAT[294]. Most bursts do not produce detectable emission above ~100 MeV: the First *Fermi*–LAT GRB Catalogue[294] contains only 35 GRBs (only 2 of them short) but was based on 733 *Fermi*–GBM burst alerts (admittedly only about half of them in the LAT field of view). The bursts detected by the LAT are brighter than the average GBM burst (though there may be selection effects at play here); the onset of the high-energy emission is delayed relative to the ~1 MeV prompt burst, and it extends over a longer period. Although the statistics are small, it is noticeable that the high-energy emission accounts for a higher proportion of the energies of the short GRBs than of the long ones.

No GRBs have been securely detected at TeV energies (the highest-energy γ-ray detected by the LAT was 94 GeV), though weak signals were claimed for GRB 970417a (Milagrito, >0.1 TeV, 3σ) and GRB 971110 (GRAND, 2.7σ); both of these were ground arrays rather than IACTs. H.E.S.S. and MAGIC have conducted searches with negative results. This cannot, however, be regarded as proof that TeV emission does not occur. It should first be noted that the poor duty cycle of IACTs, caused by the need for a clear dark sky, means that many bursts cannot be followed up: the burst may be below the horizon, it may be local daytime, the Moon may be up, or the weather may be bad. Where the burst can be followed up, it may be too distant to expect TeV emission: recall that $\gamma\gamma \to e^+e^-$ interactions with the extragalactic background light limit the distance over which TeV γs can propagate through space.

As discussed above, neutrino signals from GRBs have also been sought and not found, at a level which is coming close to the point at which a signal might be expected if GRBs do in fact accelerate the highest-energy cosmic rays.

Long GRBs and SNe Ic-BL

The astrophysical counterparts of long-soft GRBs are now agreed to be a particular subset of core-collapse supernovae. The first evidence for the association of long GRBs and supernovae was the location of SN 1998bw within the positional error box of GRB 980425. This was not regarded as definitive at the time, because GRB 980425 was very subluminous for a GRB and thus not guaranteed to be typical of the class. The situation was much improved by GRB 030329, a perfectly normal GRB associated in position and time with the spectroscopically confirmed Type Ic supernova SN 2003dh. Since then, a number of spectroscopically confirmed supernovae have been observed in association with long GRBs, and the light curves of many other long GRBs, when followed to later times, display “bumps” which look like superimposed SN light curves[224]. Most recently, the extremely bright, nearby GRB 130427A ($z = 0.34$) is associated with the Type Ic supernova SN 2013cq[295]. This is by far the most energetic GRB ($E_{\gamma,\mathrm{iso}} \sim 10^{47}$ J) securely associated with a spectroscopically confirmed supernova.

The supernovae associated with long GRBs, and occasionally with XRFs[9], are not ordinary core-collapse supernovae. They are broad-lined Type Ic (no hydrogen, silicon or helium lines in the early spectrum), somewhat brighter than average SNe Ic in the optical and much brighter in the radio, and apparently located in low-metallicity host galaxies (compared to similar SNe that are not associated with GRBs)[299]. Although the peak absolute magnitudes of the supernovae show much less dispersion than the energies of their associated

[9] X-Ray Flashes (XRFs)[297] are softer versions of long GRBs, with a peak energy of a few rather than hundreds of keV and the bulk of the prompt energy in the X-ray region. It is not clear whether they are inherently softer, or just observed slightly off the axis of the beamed emission, though in the case of GRB 060218/SN 2006aj the former is favoured[298].

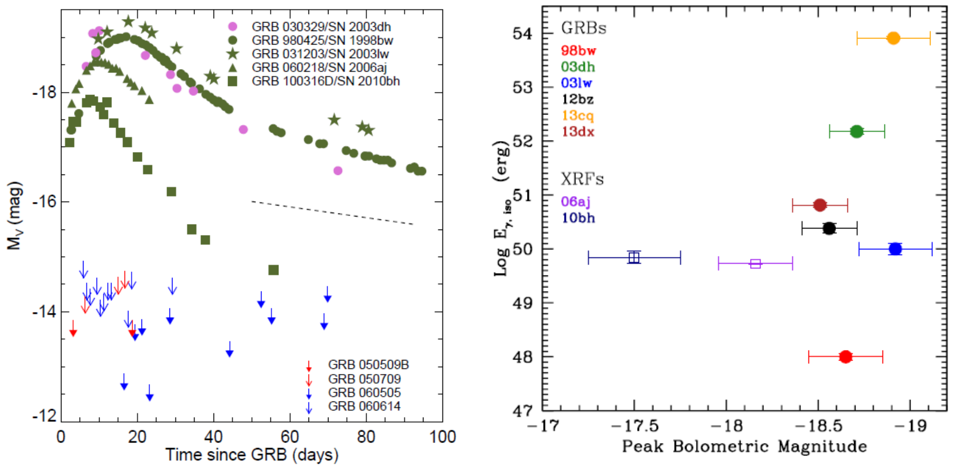


Figure 2.70: Supernovae associated with GRBs. Left panel, light curves of SNe associated with GRBs, along with upper limits for short (red) and long (blue) nearby GRBs for which no associated SN is observed[296]. Note the correlation of peak luminosity and rate of decline, similar to that observed in SNe Ia. Right panel, $E_{\gamma,\mathrm{iso}}$ for GRB/XRF against peak bolometric absolute magnitude of associated SN[295]. The dispersion in magnitudes for the SNe is *much* less than the dispersion in GRB energies, especially if the two X-ray flashes are disregarded.

GRBs (see right panel of figure 2.70), they are not “standard candles”; however, the clear correlation between peak luminosity and decline rate (left panel) suggests that they may be “standardisable candles” in the manner of Type Ia supernovae. As noted above, there are some long GRBs that are definitely not associated with supernovae (see upper limits in left panel of figure 2.70); there are also some very similar supernovae that are not associated with a GRB, particularly SN 2012ap[300].

Since GRB emission is beamed, the supernovae that produce them must somehow generate a relativistic jet (presumably two, in opposite directions, but we would only see one of them). This is not a typical feature of core-collapse supernovae. The generally agreed model[301] is a “central engine driven explosion”[300], in which the central collapsed object in the supernova powers a relativistic jet which tries to force its way through the material of the early-stage supernova. If the jet makes it to shock breakout, a GRB occurs; if it does not, a GRB does not happen, though the supernova may still have detectable mildly-relativistic ejecta, as in the case of SN 2012ap. This model explains why the associated supernovae are type Ic, i.e. stars which have lost their outer hydrogen and helium envelopes before exploding: had the envelopes still been present, the jets would have stalled before reaching the surface of the star. Long GRBs with no associated SN may occur when the supernova explosion stalls and falls back, creating a black hole without a luminous explosion.

The two preferred central engine models[301] are the *millisecond magnetar*, in which the central compact object is an extremely rapidly rotating (∼1 ms) neutron star with an exceptionally strong magnetic field ($B \sim 10^{15}$ G, 10^{11} T), and the *collapsar*, in which the central object is a promptly formed, rapidly rotating black hole. Both models can supply the energy needed to power a GRB; neither is entirely free of problems. The magnetar model may be the more testable, in the sense that magnetars (fast-spinning pulsars with very large magnetic fields) are observed in association with young supernova remnants, so that magnetar models have an observed population to test against; on the other hand, it is unlikely that every magnetar is born in association with a GRB, so the observed population and the GRB-generating population may be different.

Short GRBs and compact object mergers

Short-hard GRBs are definitely not associated with core-collapse supernovae: no supernova is observed in association with nearby short GRBs, where a supernova would be visible if one existed, and about 20% of short GRBs occur in early-type (elliptical) host galaxies, which have no ongoing star formation and therefore no core-collapse supernovae[223]. On the other hand, the fact that the fraction in early-type galaxies is as low as 20% does suggest that star formation increases the rate of short GRBs (i.e. at least some of them are associated with moderately massive stars). Unlike the hosts of long GRBs, there is no preference for low metallicity, or for high star formation rate. In addition, short GRBs tend to be located further from the centres of their host galaxies than long GRBs, and have less luminous optical and radio afterglows, indicating[223] a lower-density environment.

The preferred model for short GRB progenitors is the merger of two compact objects (NS–NS, and possibly NS–BH). The expected distribution of such systems agrees with that of short GRBs, and the energetics are appropriate. The main issue is that about 15% of "short" GRBs actually have a short initial spike (containing most of the γ-ray energy, and therefore resulting in the "short" classification) followed by a period of "extended emission" lasting tens of seconds. The extended emission is softer than either the prompt spike or long GRBs of similar duration. It is not entirely clear that this feature can be accommodated in compact object mergers[223], though there have been various attempts to do so. Possibly this subset of short GRBs has a different origin (magnetars have been suggested), though their other properties are not obviously distinct from typical short GRBs.

The association of short GRBs with NS–NS mergers is interesting for a number of reasons. First, these are also the prime candidates for detectable gravitational wave signals[302], although the opportunities for coincident detections are limited as Advanced LIGO's range for NS–NS mergers is only about 300 Mpc ($z \sim 0.07$). Second, NS–NS mergers are expected to produce a neutron-rich wind which is one of the most promising candidate sites for r-process nucleosynthesis. The radioactive decay of the super-neutron-rich isotopes produced by the r-process should power a late burst of optical/IR emission about a week after the GRB. The luminosity of this burst is of order 1000 times that of a classical nova, so it has been dubbed a "kilonova" (or sometimes "mini-SN"). The short GRB 130603b was found to have excess near-IR emission about a week after the burst, consistent with a kilonova producing of order $0.05M_\odot$ of r-process ejecta[223]. This is impressive, and if typical would imply that NS–NS mergers may be the principal source of r-process nuclides.

Short GRBs have been harder to study than long GRBs: there are fewer of them, and their afterglows are fainter. However, since the launch of the *Swift* satellite, the situation has improved markedly, and the understanding of short GRBs is rapidly catching up with their more common cousins. A coincident detection of a short GRB and a gravitational-wave signal from Advanced LIGO would be an important advance, confirming the NS–NS merger model: such coincidences are expected to occur at a rate of order 0.3 per year[223], which is small but measurable (provided some appropriate low-energy γ-ray instrument, ideally with a wide field of view, is in operation when Advanced LIGO has finished commissioning).

2.7 Summary

The evidence that certain classes of astrophysical object accelerate charged particles to extreme energies is gathered from many sources. Some of these are thoroughly embedded in mainstream observational astronomy: synchrotron emission has been observed at radio wavelengths for half a century, and X-ray astronomy is also well established. Despite its reliance on technology much more familiar to particle physicists than to astronomers, γ-ray astronomy is also becoming part of the standard astronomical toolkit. On the other hand, cosmic-ray physics, though established for a century, has never been assimilated into mainstream astronomy, though until the advent of particle accelerators it was part of the mainstream of nuclear and particle physics (the muon, the pion and strange particles were all discovered in cosmic rays), and neutrino astronomy has likewise remained a specialist preserve. Looking at the history of radio astronomy, it seems clear that new observational windows are accepted into observational astronomy at the point where their sources can be identified with known objects: radio astronomy was very much the preserve of physicists and radio engineers until the point sources began to be identified with supernova remnants and peculiar galaxies. Cosmic-ray physicists cannot do this, because of the scrambling effect of the Galaxy's magnetic field, and therefore seems doomed to remain for the most part a separate field (perhaps the very highest-energy cosmic rays, which may be traceable to their sources, might be adopted), whereas neutrino astrophysics does have the potential to identify sources and connect with the rest of astronomy, though it has not yet done so.

The signature of astrophysical particle accelerators is non-thermal radiation. As the region of the electromagnetic spectrum from the UV to the near infra-red is dominated by blackbody radiation from stars, much of our information has to come from other wavebands: radio, X-ray and γ-ray, the latter two largely inaccessible to ground-based telescopes. Cosmic rays provide information about the products of acceleration, and thus provide a testbed for accelerator models, but can give little information about their origins, and the extreme difficulty of detecting high-energy neutrinos has meant that to date they have contributed little of substance.

Nevertheless, a reasonably clear picture has built up over the last few decades. Synchrotron radiation is ubiquitous among non-thermal sources, and points to the presence of significant magnetic fields. Within our Galaxy, non-thermal emission—at all wavelengths—is typically associated with either compact objects, especially neutron stars, or supernova remnants. Outside the Galaxy, gamma-ray bursts seem to have similar associations, with long GRBs clearly linked to a specific type of SN Ic and short bursts strongly suspected to be caused by compact object mergers; here, advances in gravitational wave detection may provide confirmation in the next decade. The other class of extragalactic source, radio-loud AGN, is powered by accretion onto a supermassive black hole, and thus repeats on a much larger scale the association with compact objects. However, it should be noted that supernova remnants do not need to contain a central black hole or neutron star to act as accelerators: a counter-example is the remnant of Tycho's supernova, containing no compact object since the SN was a Type Ia, but a strong source of non-thermal emission from radio to TeV γ-rays, and almost certainly, from its γ-ray spectral energy distribution, accelerating cosmic rays. Conversely, the large class of radio-quiet AGN are powered by accretion onto a supermassive black hole, but do not seem to be strong sources of non-thermal radiation and are probably not particle accelerators. Compact objects are therefore neither a necessary nor a sufficient condition for particle acceleration.

Terrestrial particle accelerators use magnetic fields for steering and confining the accelerated particles, and electric fields, in the shape of RF cavities, to accelerate them[303]. This seems difficult to envisage for astrophysical accelerators: large-scale electric fields do not tend to occur in nature. The issue of *how* particle acceleration can be achieved in astrophysical objects is therefore not trivial, and is the subject of the next chapter.

2.8 Questions and Problems

1. Compare and contrast the fluorescence and Cherenkov techniques for detecting air showers. Why is fluorescence preferred for the detection of charged cosmic rays, while Cherenkov emission is favoured for studies of TeV γ-rays?

2. The PAMELA time-of-flight (TOF) system has a precision of 100 ps[36], and the TOF scintillators are separated by 77.3 cm. Up to what kinetic energy can PAMELA reliably separate (a) protons and deuterons; (b) oxygen-16 and oxygen-18?

3. Ionisation energy loss ($\mathrm{d}E/\mathrm{d}x$) can be used to identify particles based on their charge. Compare and contrast particle identification using TOF with identification using $\mathrm{d}E/\mathrm{d}x$.

4. Assuming a Galactic magnetic field of order 1 μG (0.1 nT), calculate the gyroradius for cosmic ray protons of energy 1 TeV, and comment on your result. It is reasonable to expect that cosmic rays with gyroradii more than about 5 kpc would not be confined in the Galaxy: to what proton energy (in eV) does this correspond? How would your answers differ if you were considering iron nuclei ($Z = 26$; $A = 56$) instead of protons?

5. Calculate the threshold proton energy for the reaction $p + \gamma \to n + \pi^+$, assuming that the photon comes from the cosmic microwave background. What is the threshold energy at which this reaction can go via the Δ^+? [The relevant masses, in MeV/c^2 are: proton 938.27; neutron 939.57; π^+ 139.57; Δ^+ 1232.]

6. Briefly explain why the particles responsible for bremsstrahlung and synchrotron radiation are assumed to be electrons, even though one might reasonably expect that the emitting region is electrically neutral overall, and thus that there must be approximately equal numbers of electrons and protons.

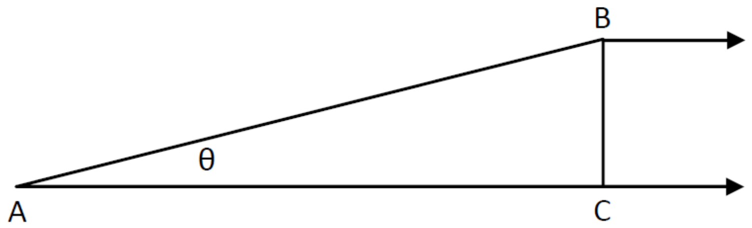
7. Explain *in words* why the spectrum of an optically thick bremsstrahlung source is proportional to ν^2, as expected from blackbody radiation, whereas the spectrum of an optically thick synchrotron source is not.

8. Show, using energy and momentum conservation, that a single high-energy photon cannot spontaneously convert into an e^+e^- pair.

9. Each of the pixels in the HESS-I telescopes subtends an angle of $6'$. If an iron nucleus interacts in the atmosphere at a height of 25 km, and its Cherenkov light is confined to a single pixel, use figure 2.55 to estimate its kinetic energy, given that the atomic mass of iron is 56 u.

10. It is often said that you would need a light-year of lead in order to stop most neutrinos. Given that lead has a density of 11.34 g cm^{-3} and an atomic mass of 207.2 u, calculate the implied neutrino interaction cross-section. By looking up appropriate references, estimate the energy of the neutrinos for which this statement is valid.

11. A characteristic of blazars is *apparent superluminal motion*: features in the blazar jet appear to move substantially faster than light. This is an optical illusion caused by the fact that the jet is oriented at a small angle to our line of sight.



The above figure shows the geometry of a superluminal jet; our line of sight is marked by the arrows. If a luminous blob moving at speed v is first observed at A, and a time Δt later has reached B, show that its *observed* speed across the sky $v_{\text{obs}} = |CB|/\Delta t_{\text{obs}}$ is given by

$$v_{\text{obs}} = \frac{\beta \sin\theta}{1 - \beta\cos\theta}\, c,$$

where $\beta = v/c$. [Hint: take into account the fact that the blob is closer to the observer at B than it is at A.]

For the case of beamed emission we can assume that $\sin\theta \simeq 1/\gamma$ (see section 2.3.5); show that in this case the apparent speed of the jet across the sky is $v_{\text{obs}} \simeq \beta\gamma$.

Chapter 3

Astrophysical Accelerators: Acceleration Mechanisms

3.1 Introduction

As discussed in the previous chapter, there is extensive evidence that some astrophysical objects are capable of accelerating both protons (and heavier nuclei) and electrons to extremely high energies. In this chapter, we will consider possible mechanisms by which this can be achieved. The key requirements, as established from observations, are:

- the proposed mechanism must generate a power-law spectrum, as the observed spectra of cosmic rays (both protons/nuclei and electrons) are approximately power laws, and the observed (non-thermal) photon spectra are consistent with a power-law distribution of the parent population;
- the spectral index of the power law should be about 2–3 (this is probably a feature of the mechanism, rather than an accident, since cosmic rays in energy ranges believed to originate from different source classes, e.g. $< 10^{15}$ and $> 10^{18}$ eV, have very similar spectral indices);
- this acceleration must be democratic, in that $e^{\pm}$, protons and heavy nuclei are all observed in cosmic rays, with abundances approximating to cosmic composition (plus the effects of spallation);
- the rate of energy gain must exceed energy losses from processes such as ionisation and bremsstrahlung (otherwise acceleration to high energies is clearly not possible);
- in at least some sites, it must be possible to achieve very high energies ($> 10^{19}$ eV for protons);
- in some cases, the timescale for acceleration must be quite short (Tycho's supernova remnant is less than 500 years old, and the Crab PWN is less than 1000 years old, but both are sources of TeV γ-ray emission).

The equation of motion of a particle of mass m and charge ze in the presence of electric and magnetic fields is

$$\frac{\mathrm{d}(\gamma m\mathbf{v})}{\mathrm{d}t} = ze(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \qquad (3.1)$$

This presents problems in accelerating particles. The term $\mathbf{v} \times \mathbf{B}$ represents a force that is always perpendicular to $\mathbf{v}$, and consequently can do no work on the particle. Therefore it does not appear that **B**-fields can accelerate particles (and, indeed, in terrestrial particle accelerators the magnets are there to steer and focus the particle beam, not to accelerate it). On the other hand, large-scale **E**-fields are very difficult to maintain in astrophysical objects, because charged plasma is highly conductive: any fixed non-zero charge will rapidly attract free ions or electrons of the opposite charge. On the face of it, this seems to rule out particle acceleration—which is unfortunate, as particles clearly *are* being accelerated!

The solution to this impasse is to consider *time-varying* electric and magnetic fields. A time-varying magnetic field will generate an electric field according to Maxwell's equations,

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{3.2}$$

and a large-scale electric field might manage to exist if it were time-varying (e.g. an intense electromagnetic wave) or transient. In this context, it is worth noting that particle acceleration in modern accelerators is achieved using electromagnetic waves in RF cavities[303], and not by DC electric fields.

3.2 The diffusion-loss equation

If cosmic rays are accelerated in astrophysical objects, then clearly

1. they must be *confined* within the source for long enough to build up their energies to the observed values (it seems most unlikely that any mechanism would be able to accelerate protons from thermal energies to 10^{19} eV in one shot), and yet

2. they must eventually *escape* from the source so that they can be observed elsewhere, specifically here in the solar system.

The observed energy spectrum, $\mathrm{d}N/\mathrm{d}E$, is a consequence of the balance between confinement and escape.

A useful tool for describing this situation is the *diffusion-loss equation* (Longair[171] section 7.5). This is a differential equation which considers both the change in energy with time (as a result of acceleration or energy losses or both) and the migration of particles into and out of the region of interest.

Following the argument of Longair section 7.5.1, we start by injecting particles into a volume $\mathrm{d}V$ at a rate $Q(E,t)\mathrm{d}V$. In this volume $\mathrm{d}V$ there are energy loss or gain processes operating, such that

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = b(E) \tag{3.3}$$

where $b(E)$ is some function of E, the form of which will depend on the processes that are operating. The signs are set up such that $b(E) > 0$ indicates that the particles lose energy.

Suppose that at time t_0 there are $N(E)\mathrm{d}E$ particles in volume $\mathrm{d}V$ which have energies between E and $E + \mathrm{d}E$. Because $\mathrm{d}E/\mathrm{d}t \neq 0$, particles migrate in and out of this energy range as their energies change. The number of particles gained or lost (depending on the sign of b) at the bottom end of the range is

$$\mathrm{d}N_1\mathrm{d}E = -N(E)\frac{\mathrm{d}E}{\mathrm{d}t}\Delta t = -b(E)N(E)\Delta t$$

and the fraction lost or gained at the top end is

$$\begin{aligned}
\mathrm{d}N_2\mathrm{d}E &= N(E+\mathrm{d}E)b(E+\mathrm{d}E)\Delta t \\
&= \left(N(E)+\frac{\mathrm{d}N}{\mathrm{d}E}\mathrm{d}E\right)\left(b(E)+\frac{\mathrm{d}b}{\mathrm{d}E}\mathrm{d}E\right)\Delta t \\
&\simeq \left(N(E)b(E)+N(E)\frac{\mathrm{d}b}{\mathrm{d}E}\mathrm{d}E+b(E)\frac{\mathrm{d}N}{\mathrm{d}E}\mathrm{d}E\right),
\end{aligned}$$

where in the last line we have neglected the term of order $(\mathrm{d}E)^2$.

The overall change in the number of particles in time Δt is therefore

$$\mathrm{d}N(E)\mathrm{d}E = \left(N(E)\frac{\mathrm{d}b}{\mathrm{d}E}+b(E)\frac{\mathrm{d}N}{\mathrm{d}E}\right)\mathrm{d}E\Delta t,$$

which gives, in dividing through by $\mathrm{d}E\Delta t$,

$$\frac{\mathrm{d}N(E)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}E}[b(E)N(E)] + Q(E,t), \tag{3.4}$$

where the second term is the rate of injection of new particles of energy E (the source term).

In addition to changes in energy, particles also diffuse in and out of the region of interest. This spatial diffusion is given by **Fick's second law of diffusion**[305]

$$\frac{\mathrm{d}N(E)}{\mathrm{d}t} = D\nabla^2 N(E),$$

where D is the diffusion coefficient. Adding this term to equation (3.4) gives the full diffusion-loss equation,

$$\frac{\mathrm{d}N(E)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}E}[b(E)N(E)] + Q(E,t) + D\nabla^2 N(E). \tag{3.5}$$

This is not quite the whole story: in addition to the source term $Q(E,t)$ representing injection of new particles, there may also be "sink" terms representing removal of particles, e.g. by radioactive decay. These are easily added to the basic equation. In the case of particle acceleration, the escape of particles from the source region might be expressed as a sink term: if, on average, particles take a time τ_{esc} to escape, then the number escaping per unit time is N/τ_{esc} and the equation becomes

$$\frac{\mathrm{d}N(E)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}E}[b(E)N(E)] + Q(E,t) + D\nabla^2 N(E) - \frac{N(E)}{\tau_{\mathrm{esc}}}. \tag{3.6}$$

3.3 Fermi second-order acceleration

An early model of particle acceleration in astrophysics was provided by Enrico Fermi in 1949[304]. Fermi postulated that protons are accelerated by being reflected off irregularities in the Galactic magnetic field, which one might nowadays associate with interstellar gas clouds, although in 1949 Fermi simply calls them "wandering magnetic fields". The key feature of this argument is the distinction between the observer's frame, in which the cloud is moving, and the centre-of-mass frame, in which the cloud is stationary (since the mass of the cloud is many, many orders of magnitude greater than that of a single particle).

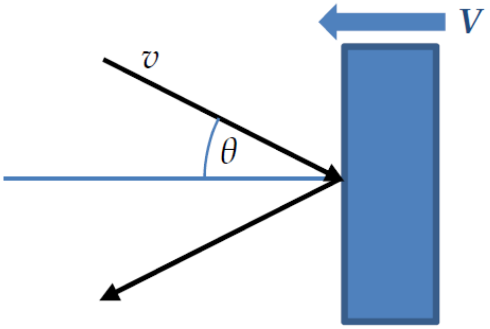


Figure 3.1: Schematic diagram of the kinematics for a particle moving at speed v reflecting from a magnetic field irregularity moving at speed V. This diagram shows a head-on collision, but an "overtaking" collision with v and V at < 90° is also possible.

Suppose the cloud moves at speed $V \ll c$, and the particle at speed v, with relative angle θ as shown in figure 3.1. In the centre-of-mass frame, the initial energy of the particle is

$$E_i' = \gamma_V(E_i + \beta_V cp\cos\theta), \qquad (3.7)$$

where γ_V is the Einstein γ factor for the cloud, $\beta_V = V/c$, and p is the momentum of the particle. The x-component of the particle momentum, $p_x = p\cos\theta$, transforms as

$$cp_x' = \gamma_V(cp_x + \beta_V E). \qquad (3.8)$$

After the collision, the particle's energy in this frame is unchanged, $E_f' = E_i'$, but the x-component of its momentum is reversed, $p_{x,f}' = -p_x'$. Therefore, transforming E_f' back into the lab frame gives

$$E_f = \gamma_V(E_f' + \beta_V cp_x'). \qquad (3.9)$$

Substituting in for E' from equation (3.7), and for cp_x' from equation (3.8), we have

$$E_f = \gamma_V^2(E_i + 2\beta_V cp\cos\theta + E_i\beta_V^2). \qquad (3.10)$$

Assuming (for simplicity) that the particle is already ultra-relativistic, $E_i \simeq cp$. Also, given that $V \ll c$, $\gamma_V^2 = 1/(1-\beta_V^2) \simeq 1 + \beta_V^2$. In this case, equation (3.10) becomes, to order β_V^2,

$$\frac{\Delta E}{E_i} = 2\beta_V\cos\theta + 2\beta_V^2, \qquad (3.11)$$

where $\Delta E = E_f - E_i$.

Assuming that the particles approach the cloud from random directions, we need to average over $\cos\theta$. One might expect that this would simply remove the first term in equation (3.11), but this is not quite the case, because of relativistic aberration. We saw on page 91 that the number density, and hence the probability of collision, of photons as seen in a frame moving at speed V is $\propto \gamma_V(1+\beta_V\cos\theta)$. Hence, the average value of $\cos\theta$ is given by

$$\langle\cos\theta\rangle = \frac{\int_{-1}^{+1}\cos\theta(1+\beta_V\cos\theta)\mathrm{d}(\cos\theta)}{\int_{-1}^{+1}(1+\beta_V\cos\theta)\mathrm{d}(\cos\theta)} = \frac{1}{3}\beta_V. \qquad (3.12)$$

Substituting this into equation (3.12) gives the final result

$$\frac{\Delta E}{E} = \frac{8}{3}\beta_V^2, \qquad (3.13)$$

the fractional increase in energy is proportional to β_V^2. This is why this acceleration mechanism is usually called *Fermi second-order acceleration*. Since it is caused by random encounters between particles and moving magnetic fields in interstellar space, it is also sometimes called *stochastic acceleration*.

Equation (3.13) tells us that the rate of change of energy

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \alpha E,$$

where $\alpha = \frac{8}{3}\beta_V^2/\tau_{\mathrm{coll}}$ is a constant and τ_{coll} is the average time between successive reflections. The rate at which particles escape from the system is $N(E)/\tau_{\mathrm{esc}}$. If we assume that there are no sources of particles of energy E, $Q(E) = 0$, and that diffusion is negligible, $D\nabla^2 E \simeq 0$, then the diffusion-loss equation (3.6) gives

$$\frac{\mathrm{d}N(E)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}E}[-\alpha E N(E)] - \frac{N(E)}{\tau_{\mathrm{esc}}}. \tag{3.14}$$

The system will eventually settle down into a steady state in which $\mathrm{d}N/\mathrm{d}t = 0$. In this case,

$$-\alpha N(E) - \alpha E \frac{\mathrm{d}N}{\mathrm{d}E} = \frac{N}{\tau_{\mathrm{esc}}}.$$

Collecting terms and dividing through by α gives

$$E\frac{\mathrm{d}N}{\mathrm{d}E} = -N\left(1 + \frac{1}{\alpha\tau_{\mathrm{esc}}}\right).$$

We can now separate the variables to get

$$\frac{\mathrm{d}N}{N} = -\left(1 + \frac{1}{\alpha\tau_{\mathrm{esc}}}\right)\frac{\mathrm{d}E}{E}.$$

Integrating and taking antilogs gives

$$N(E) \propto E^{-k}, \tag{3.15}$$

where $k = 1 + (\alpha\tau_{\mathrm{esc}})^{-1}$ is a constant. Thus Fermi second-order acceleration leads to a power law energy spectrum.

This version of Fermi second-order acceleration, which broadly follows Fermi's original treatment, is somewhat oversimplified, as discussed in Longair[171] section 17.3. The issue is that, although the *systematic increase* in energy is $\mathcal{O}(\beta_V^2)$, in *any given collision* the energy change is $\mathcal{O}(\beta_V)$ as shown in equation (3.11). To treat this properly we need to add an extra term to the diffusion-loss equation,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = D\nabla^2 N + \frac{\partial}{\partial E}[b(E)N(E)] - \frac{N}{\tau_{\mathrm{esc}}} + Q(E) + \frac{1}{2}\frac{\partial^2}{\partial E^2}[d(E)N(E)], \tag{3.16}$$

where $d(E)$ is the mean square change in energy per unit time,

$$d(E) = \frac{\mathrm{d}}{\mathrm{d}t}\left\langle(\Delta E)^2\right\rangle.$$

The value of $d(E)$ is straightforward to calculate from equation (3.11), and the result is a second-order differential equation relating N and E, whose solution is still a power law, but with a somewhat different spectral index[171],

$$k = \frac{3}{2}\sqrt{1 + \frac{16}{9\alpha\tau_{\mathrm{esc}}}} - \frac{1}{2}.$$

Thus, stochastic acceleration off magnetic field irregularities in interstellar space could in principle produce a power-law spectrum of cosmic rays. There are, however, serious problems with this model.

1. It is very slow, especially in the early stages: the peculiar velocities of interstellar gas clouds are only tens of kilometres per second, so $\beta_V \sim 10^{-4}$. Even if we were to assume that the injected particles are not participating in the rotation of the Galaxy, so that the relative velocity would be a few hundred kilometres per second, this still only implies $\beta_V \sim 10^{-3}$. In Fermi's original conception, where the acceleration takes place in interstellar space, the mean free path of cosmic rays is very long, ~0.1 pc, so we are talking about a few collisions per year, each increasing the particle's energy by at best a few parts in a million (more likely a few parts in 10^8). This does not appear promising.

2. The spectral index k depends on the speed of the magnetic "mirrors", the time (or mean free path) between collisions and the escape time. These depend on the environment, and there is no obvious reason why they would conspire to give a consistent result of 2–3 in different source types.

3. At low energies, it appears impossible for the $\mathrm{d}E/\mathrm{d}t$ generated by this process to exceed ionisation losses. This was noted by Fermi in the original paper[304], where he quoted the minimum injection energy for a proton as around 200 MeV. As ionisation losses depend on z^2, where z is the charge of the particle, this presents an even more serious problem with respect to the acceleration of heavier nuclei—recall that the composition of cosmic rays broadly reflects that of the interstellar medium.

Owing to these problems, Fermi second-order acceleration as originally envisaged by Fermi is not satisfactory as the principal acceleration mechanism for cosmic rays. It may play a role in regions where the velocities are higher, and the mean free paths shorter, than in interstellar space, e.g. in young supernova remnants, since shorter time and distance scales would allow particles to build up their energies more quickly; however, the lack of a natural route to the observed similarity of spectral indices from different source types remains a difficulty.

Clearly, acceleration would proceed much faster if the particle velocities were not isotropic with respect to the velocity of the magnetic mirrors, so that the β_V term in equation (3.11) did not get demoted to β_V^2 by averaging. Such an asymmetric situation can be realised if the acceleration takes place in the neighbourhood of a shock front. This mechanism, which is believed to be the dominant source of high-energy particles in astrophysics, is known as *Fermi first-order acceleration* or *diffusive shock acceleration.*

3.4 Astrophysical shocks

A shock, shock wave or shock front is similar to an ordinary wave, in that it propagates through a medium and has associated energy. The difference between an ordinary wave and a shock is that the properties of the medium change very abruptly at the shock front: pressure, temperature and density are very different upstream and downstream of the shock.

Shocks are common features of astrophysical objects. They are generated when a fast fluid flow encounters a solid obstacle or collides with another fluid flow, or when an initially supersonic flow decelerates to subsonic speeds. For example, in the solar system there are shocks associated with the production of solar flares, co-rotating shocks caused by fast and slow solar wind streams colliding, transient interplanetary shocks caused by the deceleration of coronal mass ejections, planetary bow shocks where the solar wind hits a planetary magnetosphere, and the solar wind termination shock where the solar wind hits the interstellar medium[306]. Most of these shocks have been directly observed by spacecraft; figure 3.2[307] shows data from Voyager 2 for the solar termination shock (crosses) and Neptune's bow shock (diamonds).

3.4.1 Shock jump conditions

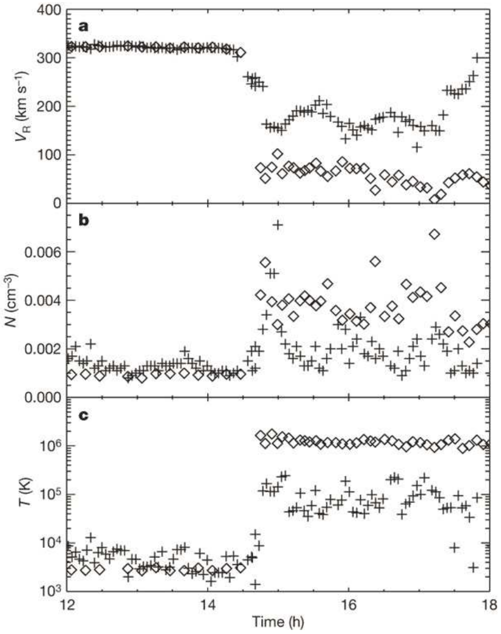


Figure 3.2: Two examples of solar system shocks. The data show the passage of the Voyager 2 spacecraft through Neptune's bow shock in August 1989 (diamonds), and subsequently, in August 2007, through the solar termination shock at 84 AU (crosses)[307]. The Neptune data have been normalised so that the upstream values match those of the solar termination shock; this required dividing the Neptune data by 1.3, 5 and 2 for panels a, b and c respectively. The solar termination shock is much weaker than the planetary bow shock; Richardson et al.[307] attribute this to transfer of energy to "pickup ions", hot protons created by ionisation in the heliosphere.

A shock produces a discontinuity in physical conditions, but clearly it cannot violate conservation laws—for non-relativistic shocks where pair production is not an issue, mass, momentum and kinetic energy must be conserved. These constraints result in relations between the properties of the gas on either side of the shock, known as *shock jump conditions.* The shock jump conditions relevant to diffusive shock acceleration are the **Rankine-Hugoniot conditions** for one-dimensional steady-state flow. ("Steady state" means that the rate of flow and the properties of the gas are not changing with time.)

To derive the Rankine-Hugionot conditions, it is simplest to work in the rest frame of the shock, as shown in figure 3.3. From the shock's perspective, gas flows in from the left, passes through the shock and exits at the right. We assume that the shock front itself is sufficiently thin that the change from pre-shock to post-shock conditions is essentially discontinuous, and that viscosity can be neglected outside the shock.

The gas mass per unit area flowing into the shock is $\rho_1 u_1$. Conservation of mass requires

$$\rho_1 u_1 = \rho_2 u_2, \tag{3.17}$$

the mass flowing out is the same as the mass flowing in.

Next, consider conservation of momentum. The mass of gas (per unit cross-sectional area) crossing the shock in time Δt is $\rho_1 u_1 \Delta t$, so its momentum is $\rho_1 u_1^2 \Delta t$. On exiting the shock it has momentum $\rho_2 u_2^2 \Delta t$, which is not the same. This change must be accounted for by a difference in gas pressure across the shock: the pressure differential $(p_2 - p_1)\Delta t$ supplies the force needed to cause the momentum change. Therefore we have

$$\rho_1 u_1^2 + p_1 = \rho_2 u_2^2 + p_2. \tag{3.18}$$

The energy of the gas has two components: its bulk kinetic energy and its internal thermal energy. The energy per unit mass is therefore

$$\mathcal{E}_{1,2} = c_V T_{1,2} + \tfrac{1}{2}u_{1,2}^2,$$

where c_V is the specific heat at constant volume. According to the first law of thermodynamics, the change in energy when the gas crosses the shock must be accounted for by the work done on the gas, $\Delta p dV$ where Δp is the difference in pressure between the two sides. The volume of gas crossing unit area of the shock front in a time Δt is $u_1 \Delta t$, so we must have

$$\rho_1 u_1 \mathcal{E}_1 + p_1 u_1 = \rho_2 u_2 \mathcal{E}_2 + p_2 u_2. \tag{3.19}$$

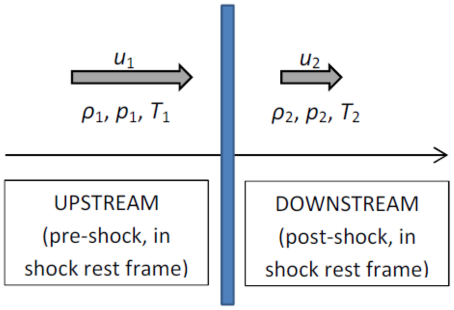

u_1
u_2
ρ_1, p_1, T_1
ρ_2, p_2, T_2
UPSTREAM (pre-shock, in shock rest frame)
DOWNSTREAM (post-shock, in shock rest frame)

Figure 3.3: Schematic diagram of shock geometry. In the rest frame of the shock, gas enters from the left with speed u_1, density ρ_1, pressure p_1 and temperature T_1, passes through the shock, and exits to the right. (In the lab frame, the shock is moving from right to left with some speed $V \ll c$.)

From thermodynamics, the ideal gas law, $pV = NRT$, can be written

$$p = \rho(\gamma_g - 1)c_V T,$$

where $\gamma_g = c_p/c_V$ is the ratio of specific heats (I have added the subscript g here to distinguish it from the Einstein γ factor). Therefore we can write $c_V T = p/[\rho(\gamma_g - 1)]$. Substituting this into $\mathcal{E}$, we can rewrite equation (3.19) as

$$p_1 u_1 \left(\frac{\gamma_g}{\gamma_g - 1}\right) + \frac{1}{2}\rho_1 u_1^3 = p_2 u_2 \left(\frac{\gamma_g}{\gamma_g - 1}\right) + \frac{1}{2}\rho_2 u_2^3. \tag{3.20}$$

If we divide the left-hand side by $\rho_1 u_1$ and the right-hand side by $\rho_2 u_2$, which are equal by equation (3.17), we get

$$\frac{p_1 \gamma_g}{\rho_1(\gamma_g - 1)} + \frac{1}{2}u_1^2 = \frac{p_2 \gamma_g}{\rho_2(\gamma_g - 1)} + \frac{1}{2}u_2^2. \tag{3.21}$$

Equations (3.17), (3.18) and (3.21) are the Rankine-Hugoniot conditions for a plane-parallel shock.

We can solve these equations to get the post-shock conditions ρ_2, p_2 and u_2 in terms of the pre-shock conditions ρ_1, p_1 and u_1 (then we can use the ideal gas law to derive T_2/T_1). If we define $q = \rho_1/\rho_2$ and $s = p_2/p_1$, then from equation (3.17) $u_2/u_1 = q$ and hence $\rho_2 u_2^2 = q\rho_1 u_1^2$. Substituting this into equation (3.18) gives

$$\rho_1 u_1^2(1 - q) = p_1(s - 1)$$

and hence

$$s = 1 + \frac{\rho_1 u_1^2}{p_1}(1 - q). \tag{3.22}$$

Equation (3.21) becomes

$$u_1^2(1 - q^2) = \frac{2p_1\gamma_g}{\rho_1(\gamma_g - 1)}(qs - 1)$$

and substituting in for s gives

$$\begin{aligned} u_1^2(1 - q)(1 + q) &= \frac{2p_1\gamma_g}{\rho_1(\gamma_g - 1)}\left[q + \frac{\rho_1 u_1^2}{p_1}q(1 - q) - 1\right] \\ &= \frac{2p_1\gamma_g}{\rho_1(\gamma_g - 1)}(1 - q)\left[\frac{\rho_1 u_1^2}{p_1}q - 1\right]. \end{aligned}$$

We now cancel the common factor of $(1-q)$ and collect all the terms in q on the left-hand side:

$$qu_1^2\left(\frac{2\gamma_g}{\gamma_g-1}-1\right)=\frac{2p_1\gamma_g}{\rho_1(\gamma_g-1)}+u_1^2$$

Multiplying through by $\rho_1(\gamma_g-1)$ gives

$$q\rho_1u_1^2(\gamma_g+1)=2\gamma_gp_1+\rho_1u_1^2(\gamma_g-1)$$

Therefore we conclude

$$\frac{\rho_2}{\rho_1}=\frac{1}{q}=\frac{\rho_1u_1^2(\gamma_g+1)}{2\gamma_gp_1+\rho_1u_1^2(\gamma_g-1)}. \tag{3.23}$$

Substituting back into equation (3.22) gives

$$\frac{p_2}{p_1}=s=\frac{1-\gamma_g}{1+\gamma_g}+\frac{2\rho_1u_1^2}{p_1(1+\gamma_g)}. \tag{3.24}$$

It is convenient to re-express these solutions in terms of the *Mach number* of the shock, which is the speed of the shock in units of the speed of sound in the unshocked gas (recall that, although we have analysed the problem in the rest frame of the shock, in the lab frame the shock front is moving!). This is given by

$$M_1=\sqrt{\frac{\rho_1u_1^2}{\gamma_gp_1}}. \tag{3.25}$$

In terms of M_1, equations (3.23) and (3.24) become

$$\frac{\rho_2}{\rho_1}=\frac{(\gamma_g+1)M_1^2}{(\gamma_g-1)M_1^2+2}; \tag{3.26}$$

$$\frac{p_2}{p_1}=\frac{2\gamma_gM_1^2-(\gamma_g-1)}{\gamma_g+1}. \tag{3.27}$$

This is useful because in the case of a *strong shock*, $M_1\gg1$, the above equations can be approximated by

$$\frac{\rho_2}{\rho_1}\simeq\frac{\gamma_g+1}{\gamma_g-1}; \tag{3.28}$$

$$p_2\simeq\frac{2}{\gamma_g+1}\rho_1u_1^2. \tag{3.29}$$

For a monatomic ideal gas, $\gamma_g=\frac{5}{3}$. Therefore, even in strong shocks, the bulk gas is compressed by only a factor of 4 in passing through the shock. In the shock rest frame, $u_2/u_1=\frac{1}{4}$; this means that in the rest frame of the unshocked gas (in which the shock is moving with velocity $-u_1$), the shocked gas is accelerated to $\frac{3}{4}$ of the speed of the shock.

3.4.2 The role of collisionless shocks

If particle-particle collisions are important in the gas dynamics of the shock, any fast particle population will tend to settle back into thermal equilibrium with the rest of the gas. Since it is clear from the results of the previous subsection that non-relativistic shocks do not accelerate the bulk gas to relativistic speeds, such a situation will not result in the acceleration of particles to cosmic-ray energies. This is not a fatal blow, however, because many astrophysical environments contain gas that is of such low density that particle-particle collisions are extremely rare[1]. Shocks that occur in such environments are called *collisionless shocks*. Although collisions between gas particles are negligible in collisionless shocks, some form of interaction must take place, because the shock front is associated with a rise in the entropy of the gas, and some form of interaction is necessary to produce this[308].

The lack of direct inter-particle collisions in collisionless shocks allows any population of fast particles to remain out of thermal equilibrium with the bulk gas. The fast particles interact with the ambient magnetic field rather than with other particles; "collisions" with magnetic fields are perfectly elastic and will not dissipate the kinetic energy of the fast particle. Therefore, any population of "superthermal" particles in a collisionless plasma can maintain its excess energy and, if conditions permit, even gain more. Collisionless shocks are therefore capable of supporting particle acceleration. An important parameter is the *criticality* of the shock[308]. In a subcritical shock, the time for which the gas flow is actually within the (extremely thin) shock front proper is sufficient to generate the necessary changes in momentum, energy and entropy to satisfy the shock jump conditions. In a supercritical shock, the gas is moving too fast relative to the shock for this to be possible, and the shock needs to dissipate energy (or generate entropy) by some other means. The natural way to do this is for the shock to reflect some of the incoming gas back on itself, similar to an impedance mismatch in AC theory. As we will see below, crossing and recrossing the shock is the crucial ingredient in particle acceleration at shock fronts, so this implies that *supercritical collisionless shocks will always accelerate particles*[309]. The boundary between subcritical and supercritical shocks is at Mach number $\mathcal{M}_c \simeq 2.76$.

If shocks are to play a role in accelerating cosmic rays, the sources of accelerated particles need to contain collisionless shocks, i.e. involve supersonic motion in very low-density gas. Fortunately for the model, such conditions are fairly easy to set up: supersonic speeds are normal even in the solar wind, and much faster supersonic outflows are seen in pulsar winds, supernova blast waves, and the radio jets of radio-loud active galactic nuclei—all locations suspected of being involved in particle acceleration. The shock must also be associated with magnetic fields: specifically, with turbulent magnetic fields that can supply energy for particle acceleration. Before continuing to look at the evidence for shocks in appropriate astrophysical sources, we should consider the effect that the presence of magnetic fields has on our derived shock jump conditions.

3.4.3 Effect of magnetic fields

In order for shocks to produce acceleration, they need to be associated with magnetic fields. This was not taken into account in our derivation of the shock jump conditions, which assumed that the only forces acting were pressure forces.

In the simple case where the magnetic field is parallel to the velocity of the shock, $q\mathbf{V} \times \mathbf{B} = 0$: the magnetic field is unaffected by the shock and does not appear in the shock jump conditions: equations (3.17), (3.18) and (3.21)

[1] This is clearly seen from the ubiquity of the 21 cm line in astrophysics: this is a highly forbidden transition which would normally de-excite collisionally rather than radiatively. It is never seen in the laboratory and is only common in astrophysics because of the prevalence of very low density atomic hydrogen in the interstellar medium.

still hold. This is not the case if the shock velocity and the magnetic field are not parallel (*oblique shock*). In this case the magnetic forces do play a role, and the Rankine-Hugoniot conditions become much more complicated, with six equations instead of three[311]:

$$\begin{aligned}
&\rho_1 u_{n1} = \rho_2 u_{n2};\\
&\rho_1 u_{n1}^2 + p_1 + \frac{B_{t1}^2}{2\mu_0} = \rho_2 u_{n2}^2 + p_2 + \frac{B_{t2}^2}{2\mu_0};\\
&\rho_1 u_{n1}\left(\frac{p_1\gamma_g}{\rho_1(\gamma_g-1)} + \frac{u_1^2}{2} + \frac{B_{t1}^2}{\rho_1\mu_0}\right) - \frac{B_{n1}\mathbf{B}_{t1}\cdot\mathbf{u}_{t1}}{\mu_0} =\\
&\qquad \rho_2 u_{n2}\left(\frac{p_2\gamma_g}{\rho_2(\gamma_g-1)} + \frac{u_2^2}{2} + \frac{B_{t2}^2}{\rho_2\mu_0}\right) - \frac{B_{n2}\mathbf{B}_{t2}\cdot\mathbf{u}_{t2}}{\mu_0};\\
&B_{n1} = B_{n2};\\
&(\mathbf{u}_1 \times \mathbf{B}_1)_t = (\mathbf{u}_2 \times \mathbf{B}_2)_t;\\
&\rho_1 u_{n1}\mathbf{u}_{t1} - \frac{\mathbf{B}_{t1}B_{n1}}{\mu_0} = \rho_2 u_{n2}\mathbf{u}_{t2} - \frac{\mathbf{B}_{t2}B_{n2}}{\mu_0}.
\end{aligned} \tag{3.30}$$

Here the subscript n denotes the component of $\mathbf{u}$ or $\mathbf{B}$ normal to the shock front (i.e. parallel to the shock velocity) and t denotes the component tangential to the shock front (note that the tangential "component" is actually a two-dimensional vector: the tangential component of $\mathbf{B}$ is not in general parallel to the tangential component of $\mathbf{u}$). This is the most general form of oblique shock, in which neither the initial gas velocity nor the magnetic field is parallel to the shock velocity.

The first three of the above equations are recognisably the same as the Rankine-Hugoniot conditions without magnetic field, though with extra terms to account for magnetic forces. The fourth is just a consequence of Maxwell's equations, $\nabla \cdot \mathbf{B} = 0$, the fifth balances tangential forces, and the sixth is the extension of equation (3.18) for the case where u_1 has a non-zero tangential component, with the addition of the magnetic forces.

Unlike equations (3.17), (3.18) and (3.21), where the solution either is trivial (all quantities constant between region 1 and region 2) or involves a shock, equations (3.30) can describe discontinuities in which there is no shock: in the rest frame of the discontinuity $u_{n1} = u_{n2}$. These are classified as *contact discontinuities*, in which only ρ and T change discontinuously, with all other quantities continuous, and *tangential discontinuities*, in which $B_n = 0$, ρ and B_t change discontinuously, but the total pressure $p + (B^2/2\mu_0)$ is conserved. These cases are of interest as regards the magnetohydrodynamics of the situation, but clearly will not be associated with particle acceleration since there is no particle transport across the discontinuity. We are only interested in the case where $u_{n1} \neq u_{n2}$, which corresponds to a shock front with particle transport across the shock.

In magnetised shocks, the Mach number as defined in equation(3.25) is replaced by a *magnetosonic* Mach number

$$\mathcal{M} = \frac{V}{\sqrt{c_s^2 + V_A^2}}, \tag{3.31}$$

where V is the speed of the shock, c_s is the sound speed in the gas, $c_s = \sqrt{\gamma_g p/\rho}$, and V_A is the *Alfvén velocity*, $V_A = B/\sqrt{\mu_0 N_e m_i}$ where N_e is the electron number density and m_i is the mass of the ions (assuming an ionised plasma).

In astrophysics, it is most often the case that the Alfvén velocity dominates over the sound speed, $V_A > c_s$.

Oblique shocks can be divided into *slow mode shocks*, in which B_t decreases, and *fast mode shocks*, in which B_t increases. Both types are observed in astrophysical contexts. As equations (3.30) become easier to handle if one of the components of $\mathbf{B}$ is zero, shocks are often described as *quasi-parallel* ($B_t \simeq 0$) or *quasi-perpendicular* ($B_n \simeq 0$) if the angle between the magnetic field and the shock normal, θ_{Bn}, is respectively close to 0 or close to 90°. Because shock fronts are rarely planar, the angle between the magnetic field and the shock can vary considerably depending on position: for example, a planetary bow shock is roughly hemispherical, and its relation to the direction of the interplanetary magnetic field (itself a complicated object, wound into a spiral by the Sun's rotation and distorted by the effects of coronal mass ejections) therefore varies from quasi-parallel to quasi-perpendicular over its surface.

If the shock is fast enough (such that $\mathcal{M} > 10$ or so), the u-terms are much more important than the B-terms in equations (3.30), and we can essentially ignore the effect of the magnetic field on the shock dynamics. For shocks with $\mathcal{M} < 10$ and $\mathbf{B}_t \neq 0$, the magnetic field does have to be considered, and we have a magnetohydrodynamic shock.

Mach numbers in the solar system are fairly small, and therefore many solar system shocks do have to be analysed using magnetohydrodynamics. Other circumstances in which the magnetic field is probably important in the shock dynamics include the vicinity of pulsars, because pulsars are characterised by very strong magnetic fields. Also, because of the effect of the $(\mathbf{u} \times \mathbf{B})$ force, both the acceleration times and the injection efficiency depend on the obliquity of the shock. This is a tricky calculation which has produced conflicting results in the past[312] and appears to require careful high-resolution simulation to achieve reliable results. The challenge, as always, is that the shocks in the solar system, for which we have the best data, do not closely resemble the much stronger and faster shocks likely to be encountered in the vicinity of supernova remnants and active galactic nuclei.

3.4.4 Observations of astrophysical shocks

If shocks play a role in the acceleration of charged particles to high energies, they must occur in conjunction with the evidence of such acceleration, i.e. synchrotron radiation, high-energy photons, and (in future) high energy neutrinos. As noted above, the only shocks we can directly confirm and measure, by sending instruments through the shock front itself, are those in the solar system (see, for example, figure 3.2), which are unlikely to be typical of those that accelerate particles to the energies of the highest-energy cosmic rays. Extrasolar shocks can only be observed indirectly, through sharp discontinuities in emission, and/or inferred theoretically, by modelling the source.

Supernova remnants, which are characterised by synchrotron emission (implying the presence of relativistic electrons), tend to have a well-defined sharp edge where the supernova blast wave collides with the interstellar medium. As an example, figures 3.4 and 3.5[313] show the remnant of SN 1006, an extremely bright Type Ia supernova observed by the Chinese. Most of the edge of this SNR has the observational properties one would associate with a strong shock, and is marked by non-thermal X-ray emission which is interpreted as synchrotron radiation. Just outside this is Hα emission, which indicates that there is neutral gas just downstream of the shock front. In SN 1006, the outer shock does not

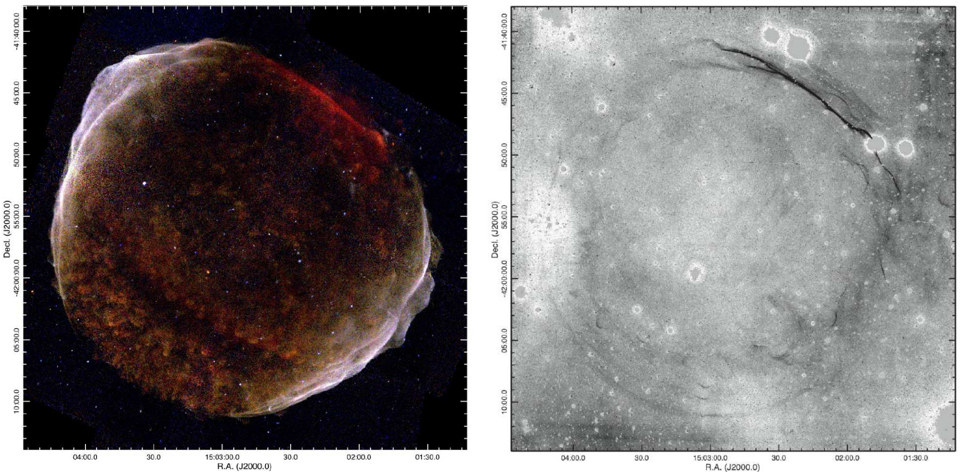


Figure 3.4: Images of the remnant of the Type Ia supernova SN 1006 (the brightest of the historical naked-eye supernovae), in X-rays (left) and Hα (right)[313]. The X-ray image from Chandra/ACIS is colour coded by X-ray energy: red = soft (0.5–1.2 keV), green = medium (1.2–2.0 keV), blue = hard (2.0–7.0 keV); note the harder spectrum associated with synchrotron X-rays at the rim of the supernova compared to the softer thermal X-rays from the interior. The right panel is a difference image between 24 stacked 10-minute exposures through an Hα filter (656.3 nm) and 22 stacked continuum (665.0 nm) images taken with the 4 m Blanco telescope at CTIO. Sharp-edged features in both images suggest the presence of a strong shock at the outside edge of the SNR, as one might expect given the high speed of supernova ejecta. The association of this shock with high-energy synchrotron emission indicates that electrons are being accelerated to high energies in the vicinity of this boundary shock—the lifetimes of electrons that can emit synchrotron radiation up to X-ray energies are too short to assume that they have been transported in from elsewhere.

seem to extend all the way round the remnant: the SE rim (lower left in figure 3.4) does not have a sharp edge feature and appears to consist of ejecta streaming unimpeded into a region of low ISM density[313]. The association of the shock front with X-ray synchrotron emission is strong evidence that electrons are being accelerated to high energies in the vicinity of the shock. In further support of this, TeV γ-rays are detected from the regions of the SNR which are bright in synchrotron X-rays[314].

In terms of extragalactic objects, the class of sources most clearly associated with particle acceleration is radio-loud active galactic nuclei; recall that nearly all identified extragalactic sources of TeV photons are blazars. Radio-loud AGN are characterised by relativistic jets, so, at least in the case of AGN in clusters of galaxies, the presence of a bow shock where the jet is slowed down by the intracluster medium is practically inevitable.

An example of a shock front in Centaurus A, a nearby (3.8 Mpc) FR I radio galaxy, is shown in figure 3.6. The morphology of the shock is similar to that of SN 1006, and the X-ray emission is consistent with synchrotron radiation. Like SN 1006, Cen A is detected in TeV γ-rays[316] and must be accelerating particles to at least TeV energies. However, in this case it is not at all clear that the shock is directly implicated in the particle acceleration; Wykes et al.[317] suggest that the acceleration takes place in the body of the lobe as a result of high levels of magnetohydrodynamic turbulence, which makes second-order (stochastic) Fermi acceleration a viable prospect.

In summary, although the association between shocks and acceleration is probably not exclusive, there is no doubt that (1) solar system shocks do ac-

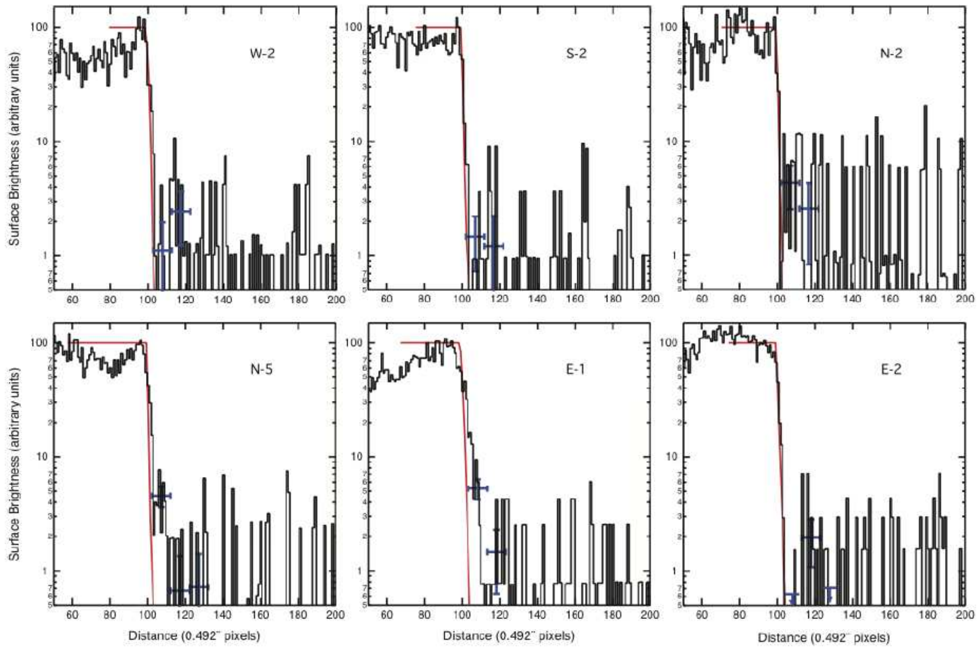


Figure 3.5: Surface brightness profiles of SN 1006 across the shock front[313], in various positions on the synchrotron-dominated NE and SW rims. All plots are oriented such that the post-shock region is to the left, and normalised to 100 for the immediate post-shock peak. The red line represents the effect of a step function (sharp edge) convolved with the Chandra point spread function.

celerate particles (to modest energies, but then they are rather modest shocks) and (2) shocks are found in classes of astrophysical object that are likely sites of particle acceleration. It is therefore reasonable to consider acceleration models in which shock crossings play a key role. Such a model, currently the most popular theory of astrophysical particle acceleration, is diffusive shock acceleration (DSA), also known as first-order Fermi acceleration.

3.5 Diffusive shock acceleration

Diffusive shock acceleration is, as its name suggests, the acceleration of fast particles diffusing through shock fronts. Its key advantage over stochastic acceleration is that the presence of the shock modifies the distribution of velocities, such that the particles diffusing through the shock will always encounter favourable collision geometries and will thus gain energy much more rapidly.

3.5.1 Test particle approach

As a first approximation, consider fast particles diffusing across a strong shock, and assume that the motion of thre fast particles does not affect the gas dynamics (in other words, treat the fast particles as massless test particles). The conditions on either side of the shock are set by the strong shock jump conditions, equations (3.28) and (3.29), and therefore the pre- and post-shock gas velocities are respectively $-V$ and $-\frac{1}{4}V$ in the shock rest frame, where $V \ll c$ is the speed of the shock.

Because of scattering, the population of fast particles in a volume of gas will quickly become isotropic in the frame of reference in which the gas is at rest. This means that the particles in the gas in front of the shock have $\langle \mathbf{v} \rangle = 0$, and the particles in the shocked gas have $\langle \mathbf{v} \rangle = \mathbf{u}_2 = \frac{3}{4}\mathbf{V}$ (note that these are vector averages—each individual particle is whizzing about with $v \sim c$, but the fast particle *population as a whole* is at rest relative to the gas in which it is embedded). Therefore, a fast particle from the upstream (pre-shock) gas diffusing across the shock will, on average, see the post-shock gas approaching it at mean speed $\frac{3}{4}V$, *and a particle diffusing in the other direction will see the same thing*, because in the frame of the post-shock gas, the upstream gas is approaching at a speed of $\frac{3}{4}V$. Both directions ensure head-on collisions, leading to an energy increase $\propto V/c$ instead of $\propto V^2/c^2$ as in the stochastic case.

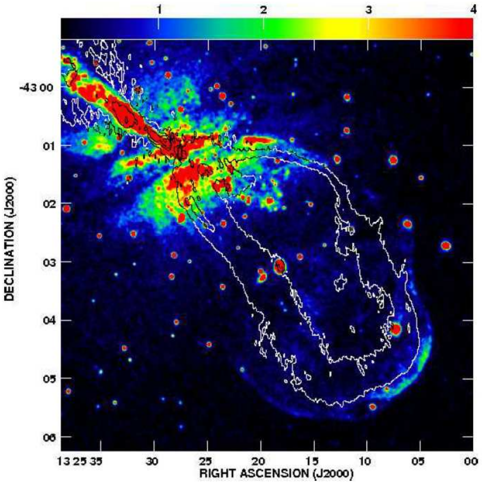


Figure 3.6: The southwest inner lobe of the nearby active galaxy Centaurus A (NGC 5128) seen in X-rays from Chandra (colour scale) and in radio at 1.4 GHz (contours)[315]. The bright X-ray emission just ahead of the end of the radio lobe is interpreted as a strong shock. The X-ray properties of the proposed shock are similar to those of the shock front of SN 1006 (see figure 3.4) and consistent with a synchrotron-radiation origin for the emission.

Starting with a particle of momentum $\mathbf{p}$ in the stationary gas upstream of the shock, we Lorentz transform into the rest frame of the downstream gas:

$$E' = \gamma_U(E + p_x U), \tag{3.32}$$

where $U = \frac{3}{4}V$ is the speed of the downstream gas. Although the gas is non-relativistic, $U \ll c$, we assume that some acceleration has already taken place and that therefore the fast particles are ultra-relativistic, $E \simeq cp$. Hence $p_x = (E/c)\cos\theta$, where θ is the angle between the particle trajectory and the shock normal. Assuming that the fast particles are isotropically distributed, the probability that any given particle crosses the shock with an incident angle between θ and $\theta + \mathrm{d}\theta$ is $\propto \sin\theta\,\mathrm{d}\theta$, and the number of such particles crossing the shock per unit time is proportional to $v_x \simeq c\cos\theta$. Therefore, the probability of a given particle crossing the shock in a given time interval is proportional to $\sin\theta\cos\theta\,\mathrm{d}\theta$. Integrating this over the appropriate range, $0 < \theta \leq \pi/2$ (since particles heading away from the shock clearly aren't going to cross it!), gives $\int_0^{\pi/2} \frac{1}{2}\sin(2\theta)\mathrm{d}\theta = 0.5$, so the properly normalised probability is

$$P(\theta)\mathrm{d}\theta = 2\sin\theta\cos\theta\mathrm{d}\theta,$$

and therefore the average energy gain is

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{U}{c}\int\limits_0^{\pi/2} 2\cos^2\theta\sin\theta\,\mathrm{d}\theta = \frac{2}{3}\frac{U}{c} = \frac{1}{2}\frac{V}{c}. \tag{3.33}$$

Particles crossing the shock in the opposite direction, from downstream to upstream, see *exactly the same kinematics*, and therefore have the same mean energy gain. Hence, the mean energy gain in one round trip, from upstream to downstream and back, is

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\frac{U}{c} = \frac{V}{c} \tag{3.34}$$

(note that, although in principle the fractional energy gain in the second shock crossing is $\Delta E'/(E+\Delta E)$, we can approximate this to $\Delta E/E$ since $V \ll c$).

Thus, as long as a fast particle remains in the vicinity of the shock, its energy will increase exponentially. However, as scattering makes the fast particle population isotropic in the rest frame of the post-shock gas, which is moving more slowly than the shock, fast particles will gradually be swept out of the shock region, a process known as *advection.* In order to derive the expected energy spectrum of the fast particles, we need to calculate the probability that a particle will be advected away from the shock.

By the same probability calculation as above, the rate at which fast particles cross the shock is $\frac{1}{4}Nc$ where N is their number density and we are assuming that their speed $v \simeq c$. Since the downstream gas is moving at speed $-\frac{1}{4}V$ in the rest frame of the shock, the rate at which fast particles are advected away by the bulk motion of the downstream gas is $\frac{1}{4}NV$. Therefore, the fraction of particles lost per unit time is $\frac{1}{4}NV/\frac{1}{4}Nc = V/c$. As we have assumed $V \ll c$, thius is a very small fraction: most of the fast particle population will undergo many shock crossings.

Putting the energy gain and the escape probability together, we have

- after each return trip, the particle energy has increased from E to fE where $f = 1 + (V/c)$;
- the probability that the particle then remains in the vicinity of the shock to undergo more shock crossings is $P = 1 - (V/c)$.

After k shock crossings, a fast particle population with initial energy E_0 and initial number density N_0 will have number density $N_k = N_0 P^k$ and energy $E_k = E_0 f^k$. Taking logs,

$$\ln\left(N_k/N_0\right) = k \ln P;$$
$$\ln\left(E_k/E_0\right) = k \ln f.$$

Dividing the first of these equations by the second to eliminate k, and then taking antilogs, we have

$$\frac{N(E \geq E_k)}{N_0} = \left(\frac{E_k}{E_0}\right)^{\ln P/\ln f};$$

this is the number of particles with $E \geq E_k$, because these particles are still in the vicinity of the shock after k return trips and can therefore undergo more shock crossings. To get the differential spectrum $N(E)\,\mathrm{d}E$, we just need to differentiate this with respect to E:

$$N(E)\,\mathrm{d}E \propto E^{(\ln P/\ln f)-1}\,\mathrm{d}E. \tag{3.35}$$

Now, because $V \ll c$, $\ln P = \ln[1-(V/c)] \simeq -V/c$, and similarly $\ln f \simeq +V/c$. Therefore $\ln P/\ln f \simeq -1$, and our predicted energy spectrum for diffusive shock acceleration is

$$N(E)\,\mathrm{d}E \propto E^{-2}\mathrm{d}E. \tag{3.36}$$

Notice that this result is completely independent of the speed of the shock. This mechanism therefore provides a natural explanation of why the same cosmic ray spectral index is produced by a range of different sources with different physical properties. Admittedly, it's not the *right* spectral index—as we have seen, the observed spectral index is rather higher than this, somewhere in the region of 2.5 to 3.0—but it is of the right order of magnitude (it's not 0.2 or 20, for example), and this was a very approximate treatment. Given the evidence that strong shocks are common in the classes of astrophysical objects believed to be acting as particle accelerators (see the previous section), this result does tend to support the idea that diffusive shock acceleration may be the dominant mechanism in these objects.

A more general test-particle result

The result derived above was for the limit of strong shocks (compression ratio $r = \rho_2/\rho_1 = 4$) and ultrarelativistic test particles $E \simeq pc$. We can repeat the above calculation with these requirements relaxed; the result is a power law in momentum:

$$f(p) \propto p^{-3r/(r-1)}, \tag{3.37}$$

where $f(p)$ is the phase space density and r is the compression ratio defined above. This still has no dependence on the diffusion coefficient or the shock obliquity. Assuming an isotropic distribution, the volume element in momentum space $\mathrm{d}^3\mathbf{p}$ can be written as $4\pi p^2 \mathrm{d}p$, leading to a power law in momentum $p^{-\Gamma}$ where

$$\Gamma = \frac{3r}{r-1} - 2 = \frac{r+2}{r-1}.$$

Changing variables from momentum to energy gives

$$N(E)\,\mathrm{d}E \propto p^{-\Gamma}\frac{\mathrm{d}p}{\mathrm{d}E}\mathrm{d}E.$$

For relativistic particles, $E \simeq cp$ and the kinetic energy distribution $N(E) \propto E^{-\Gamma}$, which gives us the E^{-2} law for $r = 4$ as previously derived. For non-relativistic particles, the kinetic energy $E \simeq p^2/2m$, which means that $\mathrm{d}p/\mathrm{d}E \propto 1/p$ and the power law becomes $N(E) \propto E^{-(\Gamma+1)/2}$; for $r = 4$ this is $E^{-3/2}$.

This explains why cosmic rays are predominantly protons and heavier ions, rather than electrons: the electrons will become ultrarelativistic at lower energies, so the change in spectral index occurs earlier, suppressing the high-energy tail. Therefore, even if protons and electrons are injected into the shock region at the same rate—as we might expect—we should see fewer electrons at GeV energies and above. The observed ratio of about 1% electrons can be accounted for by this effect.

We also expect that there will be a high-energy cut-off in the power law spectrum, since at high energies the gyroradius of the fast particles will be such that the magnetic field cannot contain them within the shock region. This is usually modelled[318] as an exponential, $\exp(-E/E_0)$ where E_0 is some characteristic maximum energy.

The maximum energy can be estimated (see Longair[171] section 17.4) by considering Maxwell's equation

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.$$

For an order of magnitude estimate, the derivatives can be replaced by simple divisions,

$$\frac{E}{L} \sim \frac{B}{L/V},$$

where V is the speed of the shock and L is the size of the region in which the acceleration is taking place. If we consider that a particle of charge ze is accelerated by this induced electric field $E = BV$, then the maximum energy is of the order of

$$E_{\mathrm{max}} \sim zeEL \sim zeBVL. \tag{3.38}$$

For a young supernova remnant with $B \simeq 10^{-10}$ T, $V \simeq 10^4$ km s^{-1} and an age $\tau \simeq 10^3$ years, the length scale is of order 3×10^{17} m and hence the maximum energy per nucleon is around 10^{14} to 10^{15} eV. This is comparable to, though somewhat lower than, the maximum energy that we inferred from the composition of cosmic rays, see section 2.2.3.

3.5.2 Beyond the test-particle approach

The test-particle approach to diffusive shock acceleration relies on the basic assumption that the fast particle population does not affect the shock—that's what we mean by "test particles". Unfortunately, both theoretical and observational considerations imply that this assumption is not even approximately true. As we saw earlier, supercritical shocks *must* reflect particles back across the shock front in order to generate enough entropy to conserve energy and momentum: this implies a *significant* reflected flux, not a negligible population of test particles. Confirming this, observations of solar system shocks such as the Earth's bow shock indicate that the shock accelerates ions *efficiently*, with about 25% of the solar wind energy flux converted into superthermal ions[309]—the kinetic energy of these ions is very modest, only a few tens of keV per nucleon, but their flux is well above the Maxwellian tail.

We also need a turbulent magnetic field, as discussed in section 3.1. Interstellar magnetic fields always have some degree of turbulence, but this is not enough to produce efficient particle acceleration over a wide energy range. Fortunately, it turns out that a consequence of efficient acceleration of particles in strong shocks is that magnetic turbulence is generated via streaming instabilities[309, 319], as shown in figure 3.7.

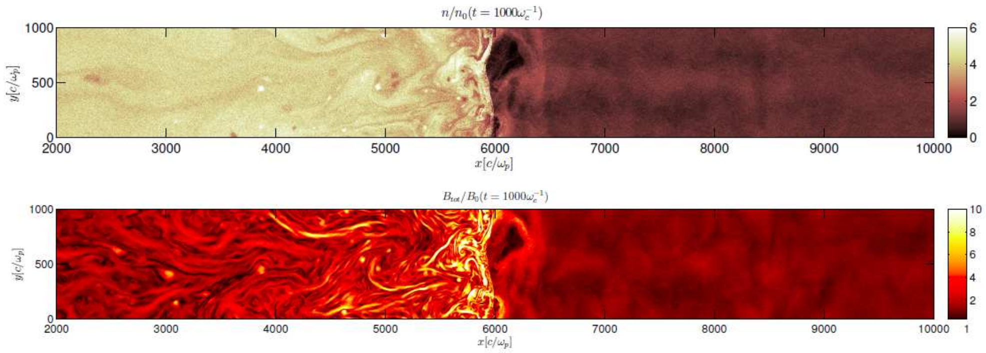


Figure 3.7: Magnetic field turbulence in the wake of a strong shock. The figure shows the output of a large *hybrid simulation* of a shock with magnetosonic Mach number 20 at time $t = 1000/\omega_c$, where $\omega_c = eB/mc$ is the cyclotron frequency[319]. The top panel shows the ion density—note the compression ratio of ∼4 between pre-shock on the right and post-shock on the left—and the magnitude of the magnetic field. The scale is in terms of the initial values. A hybrid simulation is one in which the ions are treated kinetically as particles, while the electrons are considered as a continuous fluid. Figure from Caprioli and Spitkovsky[319], paper II.

A realistic analysis of diffusive shock acceleration therefore involves:

1. injection of a seed population of particles to be accelerated;

2. acceleration of these particles by the Fermi first-order mechanism;

3. effect of accelerated particles on shock and magnetic field.

If the seed population is the thermal tail of the gas, which seems logical, then this requires a more careful analysis than we have presented so far (we have assumed that the particle velocity $v \gg V$, the shock speed; this would not be true for a thermal population). The second and third stages form a feedback loop—the shock properties determine the response of the particles, and in turn the particle response modifies the shock. Models taking all this into account are *non-linear*, and difficult to address analytically, so this work is normally done using hydrodynamic simulations. Furthermore, the simulations are highly non-trivial[309]:

- reducing dimensionality of simulations from 3D to 1D or 2D is not valid, because particles are then trapped on magnetic field lines and cannot diffuse across the field;
- particle distributions are highly anisotropic;
- the feedback loop needs time to settle down into a stable state, so simulations must be run long enough to achieve this;
- turbulence, which is required to produce acceleration, is notoriously difficult to model.

The effect of the first three items is to increase the required computer power substantially; the effect of the last is to increase the required sophistication of the modelling, which tends to increase the CPU requirements still further. Consequently, although acceleration in collisionless shocks has been studied intensively since the late 1970s, new results are still being produced, often contradicting older studies conducted using lower resolution or reduced dimensionality.

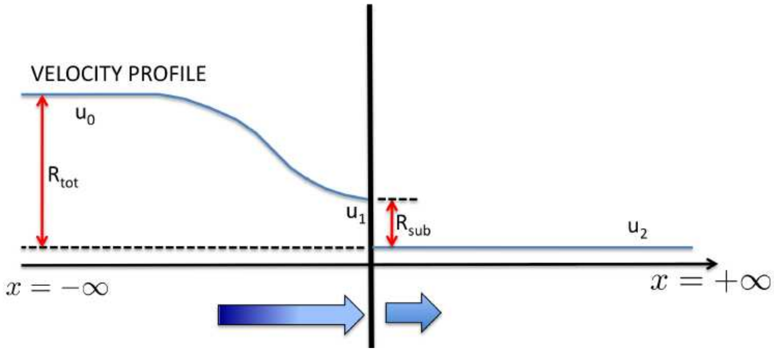


Figure 3.8: Modification of the shock front by pressure from accelerated particles. The figure shows the shock rest frame, with unshocked gas approaching from the left (the shock is therefore moving from right to left). Instead of a single discontinuous transition from u_0 to u_2 at the shock front, we have a gradual slowing down of the incoming gas from u_0 to u_1, followed by a much smaller discontinuous transition or subshock. Figure from Blasi[320].

Schematically, the effect of the back pressure in accelerated particles is to soften the sharp edge in the velocity distribution by its effect on the gas just ahead of the shock. In the rest frame of the shock, the gas approaching the shock slows down as it nears the shock front, producing a *shock precursor*, while the discontinuity at the shock front itself is correspondingly less pronounced and is referred to as a *subshock*[309, 320] (see figure 3.8). In the lab frame, the pressure from the fast particles is exerting a force on the gas immediately in front of the shock, with the result that this gas is no longer stationary when the shock front hits it.

This has a few potential consequences. First, the shock jump conditions we derived earlier apply to the *subshock*, not to the whole transition including the precursor, and as a result the overall compression ratio can in principle be greater than 4 (only the subshock compression ratio is restricted to $r_{\rm sub} \leq 4$). In the extreme gas where the effective value of γ_g goes from 5/3 (monatomic ideal gas) to 4/3 (relativistic gas), the overall compression ratio could become as high as 7, although simulations indicate[320, 319] that it actually tends to remain quite close to 4. Second, as the higher-energy fast particles naturally tend to diffuse further into the shock precursor, and therefore experience a larger effective compression ratio as shown in figure 3.8, the resulting energy or momentum spectrum is no longer a simple power law.

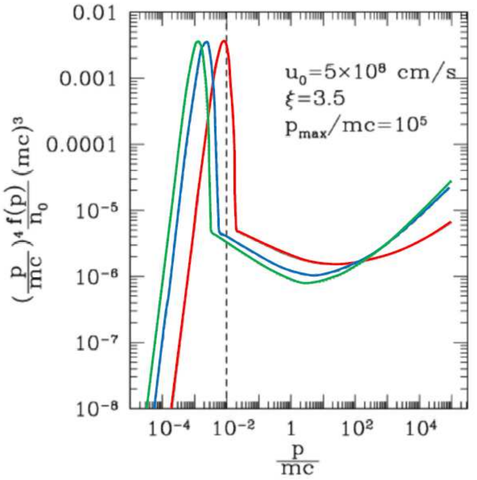


Figure 3.9: Particle phase space density $f(p)$, scaled by p^4, in a modified shock with Mach number $\mathcal{M} = 10$ (red), 50 (blue) and 100 (green). The sharp peak is the thermal distribution of unaccelerated gas, and the long tail at high p/mc is the accelerated cosmic rays. The vertical dashed line shows where the peak of the thermal distribution would be in a pure test-particle shock; because of the scaling, the high-momentum tail of the test-particle distribution would be a horizontal line. Figure adapted from Blasi[320]. The speed of the shock (or, in the shock rest frame, of the unshocked gas) is fixed at 5×10^6 m s^{-1}; the Mach number is changed by adjusting the temperature of the unshocked gas. The injection parameter $\xi = 1 - (u_1/u_0)$.

In the model of Blasi et al.[309, 320], this results in a power spectrum which is concave when scaled by p^4, as shown in figure 3.9: the spectrum is steeper than the canonical value at low momenta and shallower at high momenta. This effect does not seem to be seen in the hybrid simulation of Caprioli and Spitkovsky[319], where the tail is nearly flat, with an exponential cut-off that moves to higher energies as the simulation is allowed to run for longer times. Unfortunately the `dHybrid` code used by Caprioli and Spitkovsky is non-relativistic and cannot probe the high-energy regime of interest for cosmic rays; the upturn seen in figure 3.9 occurs at relativistic energies, $p > mc$, and is not inconsistent with the later work.

Last but not least, the presence of a significant fast-particle population results in amplification of the magnetic field and generation of magnetic turbulence[320]. Field amplification is essential in order to understand the synchrotron radiation from young supernova remnants, which requires magnetic fields of the order of a few hundred μG (a few tens of nT) rather than the few μG typically found in the interstellar medium[320, 321]. As noted above, turbulence is required to provide the necessary "magnetic mirrors" for first-order Fermi acceleration. There are various mechanisms by which this might occur[308, 320]; the most promising appear to be due to *streaming instabilities* induced by particles moving through a plasma at more than the plasma Alfvén speed. This can be resonant, exciting waves with wavenumber $k = 1/r_g$, where r_g is the gyroradius of the fast particles, or non-resonant, exciting small-scale waves with $k > 1/r_g$. Somewhat counterintuitively, it appears[320] that the non-resonant mode is more important, as the resonant mode is self-limiting (the growth of the magnetic fluctuations destroys the resonance). Caprioli and Spitkovsky[319] find that resonant modes dominate for shocks with Mach numbers up to around 30, and non-resonant modes for stronger shocks.

Detailed simulations[319] show that the small-scale non-resonant instabilities themselves induce modes with larger spatial scales and the formation of filamentary structures. In addition, there is a larger-scale *firehose instability* (so named because it is analogous to the way that a garden hose sometimes thrashes about sideways in response to a fast water flow), but this may not grow fast enough to be relevant, at least for young supernova remnants. It is clear that this is a complex problem, and further work with sophisticated numerical simulations is probably necessary.

In summary, particle-in-cell (PIC) simulations of non-linear diffusive shock acceleration indicate that efficient acceleration of cosmic rays is possible (Caprioli and Spitkovsky[319] find that quasi-parallel shocks can accelerate ions very efficiently, with 10–20% of the thermal energy converted into kinetic energy of fast (though, in their code, still non-relativistic) particles; quasi-perpendicular shocks are much less successful, because the accelerated particles escape too quickly). Overall, there is strong circumstantial evidence that DSA is the mechanism responsible for the acceleration of cosmic rays in supernova remnants; the case for other putative CR sources, such as AGN, is less clear-cut.

3.5.3 Shock drift acceleration

As noted above, diffusive shock acceleration takes place in the context of quasi-parallel shocks: in quasi-perpendicular shocks, the direction of the magnetic field does not facilitate repeated shock crossings. An acceleration mechanism that does occur in, and indeed requires, quasi-perpendicular shocks is *shock drift acceleration* (with the somewhat confusing acronym of SDA, which is not a misprint for DSA!). In shock drift acceleration, particles entrained in the magnetic field drift *along* the shock front for some distance before being either reflected from it or transmitted through it.

Shock drift acceleration is analysed by Ball and Melrose[322]. The result is that the maximum kinetic energy ratio for particles eventually reflected from the shock is

$$\left(\frac{E_r}{E_i}\right)_{\max} = \frac{1 + \sqrt{1 - (B_1/B_2)}}{1 - \sqrt{1 - (B_1/B_2)}}, \tag{3.39}$$

where B_1 and B_2 are the magnetic fields on the upstream and downstream sides of the shock (the downstream magnetic field B_2 is stronger than B_1). The minimum value of B_1/B_2 for a non-relativistic shock is $\frac{1}{4}$, which corresponds to a strong perpendicular shock (with compression ratio 4 and angle between $\mathbf{B}$ and the shock normal close to 90°); this gives $E_r/E_i \simeq 14$. Particles transmitted through the shock do less well, with a maximum ratio of 7.46 for particles transmitted from upstream to downstream, and only 1.87 for those transmitted from downstream to upstream.

Although a factor of 14 is much larger than the factor of $1+\frac{V}{c}$ obtained from a single shock crossing in diffusive shock acceleration, this maximum value is attained only for particles with a particularly favourable initial geometry[322]. Furthermore, particles involved in SDA tend to escape from the neighbourhood of the shock rather quickly, after only one or a few shock crossings or reflections, thus limiting the total increase in energy. For this reason, shock drift acceleration is not believed to play a major role in accelerating cosmic rays. It is, however, known to occur in the solar system: it has been studied at planetary bow shocks and may be responsible for Type II solar radio bursts[323]. It is also possible that it might provide the initial "seed" population for diffusive shock acceleration: as particles can be accelerated by a factor of 10 or so in a single reflection, the problem that small energy gains are easily wiped out by ionisation losses at low energies is less acute.

3.6 Relativistic shocks

The analyses above all considered non-relativistic shocks, in which it is reasonable to assume that the relativistic particles, $v \sim c$, are isotropic from the point of view of the shock, $V \ll c$. However, many astrophysical collisionless shocks, particularly those in gamma-ray bursts and in the jets of active galactic nuclei, are likely to be moving at relativistic speeds, such that $V \sim c$ and $\gamma_V \gg 1$. In this case, the velocities of the particles are not isotropic with respect to the shock, and the idea of *diffusive* shock acceleration is no longer valid—the shock will catch up to particles in the unshocked gas before they can scatter often enough to become isotropic, and the relevant angles will be subject to relativistic beaming effects.

In the rest frame of the shock, the shock jump conditions for a relativistic shock are

$$\begin{aligned} \gamma_1 \rho_1 \beta_1 &= \gamma_2 \rho_2 \beta_2; \\ \gamma_1^2 w_1 \beta_1^2 + p_1 &= \gamma_2^2 w_2 \beta_2^2 + p_2; \\ \gamma_1^2 w_1 \beta_1 &= \gamma_2^2 w_2 \beta_2; \end{aligned} \tag{3.40}$$

where ρ is the mass density, the *enthalpy* $w = \mathcal{E}+p$, $\mathcal{E}$ is the energy density, p the pressure, β the velocity in units of c, $\gamma = (1-\beta^2)^{-1/2}$, and the subscripts 1 and 2 refer to the upstream (unshocked) and downstream (shocked) gas respectively. These are just the relativistic forms of the Rankine-Hugoniot conditions that we derived in section 3.4.1.

We can also define[324] an effective equation of state

$$p = (\hat{\gamma} - 1)(\mathcal{E} - \rho c^2), \tag{3.41}$$

where $\hat{\gamma}$ in the non-relativistic (or fully relativistic) case would be the ratio of specific heats, γ_g; in more complicated situations such as where the gas is only mildly relativistic or where it consists of multiple components only some of which are relativistic (e.g. a plasma of relativistic electrons and non-relativistic ions), the interpretation of $\hat{\gamma}$ is less straightforward, but its value does generally lie between the $\frac{4}{3}$ expected for a fully relativistic monatomic gas and the $\frac{5}{3}$ of a non-relativistic monatomic gas[325].

Equations (3.40) must be solved numerically in general. However, in the simple case where the shock is sufficiently highly relativistic that the unshocked pressure p_1 is negligible, and the downstream particles are highly relativistic so that $\rho_2 c^2$ can be neglected in equation (3.41), we have

$$p_2 \simeq (\hat{\gamma} - 1)\mathcal{E}_2$$

and (combining the second and third shock jump conditions)

$$\gamma_1^2 w_1 \beta_1^2 = \gamma_2^2 w_2 \beta_2 \beta_1 \simeq \gamma_2^2 w_2 \beta_2^2 + (\hat{\gamma} - 1)\mathcal{E}_2.$$

Now $w_2 = \mathcal{E}_2 + p_2 \simeq \hat{\gamma}\mathcal{E}_2$, so this can be rearranged to give

$$\gamma_2^2\beta_2(\beta_1 - \beta_2)\hat{\gamma}\mathcal{E}_2 = (\hat{\gamma} - 1)\mathcal{E}_2.$$

Cancelling off the $\mathcal{E}_2$ and writing $\beta_1 \simeq 1$ since in the shock rest frame the unshocked gas is travelling at close to the speed of light, we have $\gamma_2^2(\beta_1 - \beta_2) \simeq 1/(1+\beta_2)$ and hence

$$\frac{\beta_2}{1+\beta_2} = \frac{\hat{\gamma} - 1}{\hat{\gamma}},$$

which gives

$$\beta_2 = \hat{\gamma} - 1, \tag{3.42}$$

or $\beta_2 = \frac{1}{3}$ assuming $\hat{\gamma} = \frac{4}{3}$ as appropriate for a fully relativistic gas.

If we write $\beta_1 = 1 - \epsilon$ where $\epsilon \ll 1$, then $\gamma_1 = \left(1 - (1-\epsilon)^2\right)^{-1/2} \simeq 1/\sqrt{2\epsilon}$. By the relativistic formula for addition of velocities, the relative velocity between the shocked and unshocked gas is

$$\beta_{\rm rel} = \frac{\beta_1 - \beta_2}{1 - \beta_1\beta_2} \simeq \frac{\frac{2}{3} - \epsilon}{1 - \frac{1}{3}(1-\epsilon)} \simeq 1 - 2\epsilon \tag{3.43}$$

using the binomial expansion. Therefore

$$\gamma_{\rm rel} \simeq \left(1 - (1-2\epsilon)^2\right)^{-1/2} \simeq \frac{1}{2\sqrt{\epsilon}} = \frac{\gamma_1}{\sqrt{2}}. \tag{3.44}$$

We can now use these relations to investigate the behaviour of particles crossing the shock. Following [326], we define

$$\mu \equiv \cos\theta \simeq \beta_{\|}$$

for relativistic particles with $\beta \simeq 1$, where θ is the angle between the particle velocity and the shock normal in the upstream (unshocked) rest frame. Note that, as the motion of both the shock and the accelerated particles is relativistic, the angles will have to be transformed appropriately in other rest frames.

Consider a particle with initial energy E_i which crosses the shock from upstream to downstream with a cosine to the shock normal of μ_1, scatters elastically in the downstream gas, and then recrosses the shock with direction cosine μ_2. In the upstream rest frame, the ratio between the final energy E_f and the initial energy is

$$\frac{E_f}{E_i} = \frac{1 - \beta_{\rm rel}\mu_1}{1 - \beta_{\rm rel}\mu_2}, \tag{3.45}$$

or, equivalently,

$$\frac{E_f}{E_i} = \frac{1}{2}\gamma_s^2\left(1 - \beta_{\rm rel}\mu_1\right)\left(1 + \beta_{\rm rel}\bar{\mu}_2\right), \tag{3.46}$$

where $\bar{\mu}_2$ is the direction cosine measured by a *downstream* observer (following [326] in denoting quantities in the downstream rest frame by an overline), and Lorentz transforming the angles gives

$$\mu_2 = \frac{\bar{\mu}_2 + \beta_{\rm rel}}{1 + \beta_{\rm rel}\bar{\mu}_2};$$

we have used equation (3.44) to substitute $\frac{1}{2}\gamma_s^2$ for $\gamma_{\rm rel}^2$.

If the shock propagates with speed β_s, all upstream particles with velocities such that $-1 \leq \beta_1\mu_1 < \beta_s$ will be overtaken by the shock front and will therefore cross the shock. We can consider two extreme cases:

1. The particle that crosses the shock is simply a particle of the upstream gas, almost at rest in the upstream rest frame. In this case, $\beta_1 \ll 1$ and $E_i \simeq mc^2$ where m is the particle mass. In the downstream rest frame, the energy of the particle is $\bar{E} = \gamma_{\rm rel} mc^2$. After scattering (and Lorentz transforming) back into the upstream gas, the energy of this particle is

$$E_f = \frac{1}{2}\gamma_s^2(1 + \beta_{\rm rel}\bar{\mu}_2)mc^2, \tag{3.47}$$

   using equation (3.46). Since we know that $\beta_2 = \frac{1}{3}$, all particles that get back across the shock must have $\frac{1}{3} < \bar{\mu}_2 \leq 1$; this gives

$$\frac{2}{3}\gamma_s^2 < \frac{E_f}{E_i} \leq \gamma_s^2. \tag{3.48}$$

2. The particle that crosses the shock belongs to a relativistic population whose directions are isotropic in the rest frame of the upstream gas. In this case the initial angle must be in the range $-1 \leq \mu_1 < \beta_s$, but because of relativistic beaming the particle flux for a given μ_1 is proportional to $\beta_s - \mu_1$. Taking $\beta \simeq 1$ and averaging over μ_1 gives $\langle 1 - \beta_{\rm rel}\mu_1 \rangle \simeq \frac{4}{3}$, and substituting this into equation (3.46) gives

$$\frac{E_f}{E_i} = \frac{2}{3}\gamma_s^2(1 + \beta_{\rm rel}\bar{\mu}_2), \tag{3.49}$$

   from which

$$\frac{8}{9}\gamma_s^2 < \frac{E_f}{E_i} \leq \frac{4}{3}\gamma_s^2, \tag{3.50}$$

   assuming that the angles μ_1 and $\bar{\mu}_2$ are uncorrelated so that we can average them independently.

Assuming that intermediate cases will lie between these two extremes, it is safe to conclude that the first return crossing produces a fractional energy gain of order γ_s^2, with a numerical coefficient within ∼30% of unity.

Subsequent shock crossings have a rather different outcome. The key point is that the direction cosine required to recross the shock, $\bar{\mu}_2 > 1/3$, transforms into the upstream rest frame as $\mu_2 > \beta_s = 1 - (1/\gamma_s)^2$. As $\gamma_s \gg 1$ (since we are assuming this is a highly relativistic shock), this defines a critical angle $\theta_{\rm c} \simeq \sin\theta_{\rm c} = 1/\gamma_s$. In order to recross the shock for a new cycle, the particle must be scattered out of this cone so that it can be overtaken by the shock. However, as a consequence of relativistic beaming, it can be shown[326] that the maximum angle through which the particle can be scattered is itself of order $1/\gamma_s$, so the particles after scattering are still confined to a narrow cone, this time of half-angle $2/\gamma_s$. Hence, for a subsequent shock crossing,

$$\theta_2 < \frac{1}{\gamma_s} < \theta_1 < \frac{2}{\gamma_s}$$

(we must have $\theta_1 > 1/\gamma_s$ in order that the particle is overtaken by the shock). Substituting the relations $\beta_{\rm rel} \simeq 1 - (1/\gamma_s^2)$ and $\mu = 1 - \frac{1}{2}\theta^2$ into equation (3.45) gives

$$\frac{E_f}{E_i} \simeq \frac{1 + \frac{1}{2}\gamma_s^2\theta_1^2}{1 + \frac{1}{2}\gamma_s^2\theta_2^2}. \tag{3.51}$$

Since most of the solid angle of a cone is near the edge, we can estimate $\theta_1 \simeq 2/\gamma_s$ and $\theta_2 \simeq 1/\gamma_s$; this gives $E_f/E_i \simeq 2$. Therefore, in contrast to the first shock crossing, subsequent crossings of a relativistic shock will typically only double the energy. This is still much better than the performance of a non-relativistic shock, where the energy gain per shock crossing is of order $\Delta E/E \sim \beta_s \ll 1$, but it is nowhere near as large as the gain in the first shock crossing, which is $\sim 10^4$ if the Lorentz factor of the shock is ~ 100.

Also, the requirement that $\bar{\mu}_2 > \frac{1}{3}$ in order to recross the shock implies that the escape probability is rather large, $\sim \frac{1}{3}$. The combination of these two factors results in a somewhat steeper slope for the energy spectrum, with a spectral index of order 2.2–2.3 [326] instead of the index of 2 predicted by the test particle calculation for non-relativistic shocks. This is good, because the combination of the observed cosmic ray spectral index of 2.7 and the effects of propagation through the Galaxy does in fact suggest a spectral index steeper than 2, see section 3.8 below.

To summarise, acceleration in relativistic shocks is potentially a much faster process than diffusive shock acceleration in non-relativistic shocks, with a single shock crossing able to accelerate particles to γ-factors of order 10^4 given a sufficiently relativistic shock. There are issues that we have not discussed, principally as regards the turbulence necessary to provide appropriate scattering[327], which may present problems in highly magnetised plasmas (where magnetic reconnection, see below, may be the preferred mechanism), but overall particle acceleration in relativistic shocks encountering unshocked gas with low magnetisation appears to be a viable process.

3.7 Particle acceleration by magnetic reconnection

Diffusive shock acceleration is the best studied and most popular mechanism for efficiently accelerating cosmic rays. However, there are other possibilities, and at least one of these—*magnetic reconnection*—is well attested in the solar system, specifically in solar flares and coronal mass ejections.

Magnetic reconnection (see figure 3.10) occurs when magnetic field lines of opposite polarities are forced close together. Eventually the original field lines will "snap" and reform in an orthogonal direction[328]. The reconnection event leads to a lower-energy configuration of the magnetic field, and the energy thus released can be converted into particle acceleration and/or bulk flow of

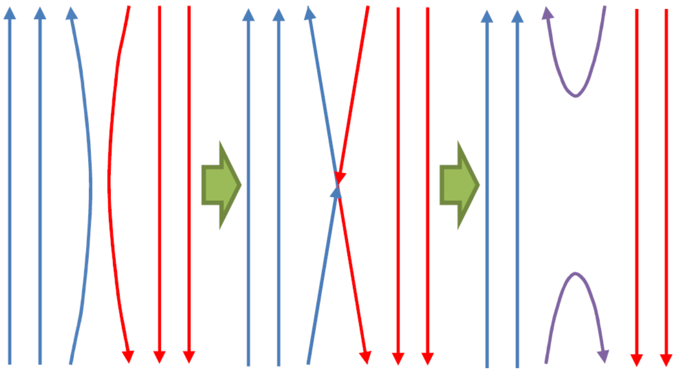


Figure 3.10: Schematic diagram of magnetic reconnection. Magnetic field lines in plasma flows may be forced closer together (left panel). At some critical point (middle panel), they may merge to form so-called X-lines, which then reconnect (right panel) to produce orthogonal field lines. The plasma flows inwards from left and right, and vertically outwards after the reconnection event.

the plasma: the latter is believed to be responsible for launching coronal mass ejections[329].

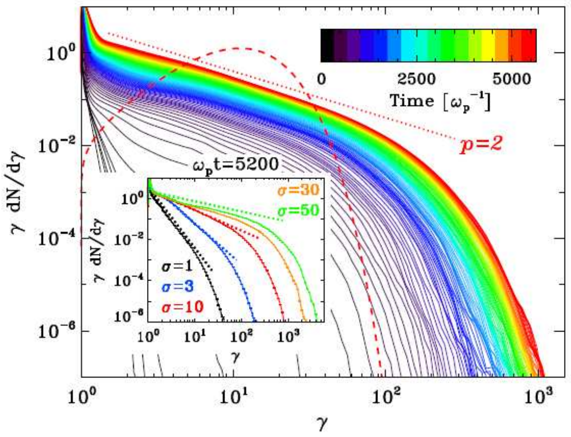


Figure 3.11: Energy spectrum of particles accelerated by magnetic reconnection, according to a 2D PIC simulation[332]. The main plot shows the evolution of the energy spectrum for a plasma with magnetisation parameter $\sigma = 10$; the inset shows the dependence on σ. The dotted red line is a power law of spectral index 2, while the dashed red line shows the Maxwellian energy distribution that would be obtained if the magnetic field energy were dissipated thermally.

There are various routes by which magnetic reconnection could lead to particle acceleration. The geometry shown in figure 3.10, extended to three dimensions, can result in a *reconnection layer* in which multiple reconnection events take place, producing isolated "magnetic islands" as the purple lines in the right-hand panel of figure 3.10 join up with equivalent lines from a neighbouring reconnection event to form closed loops. This magnetic geometry provides a suitable environment for first-order Fermi acceleration, as particles trapped in the reconnection layer scatter repeatedly off the surrounding magnetic fields[330]. The reconnection events also induce a strong transient electric field, which can accelerate particles directly. Cerutti et al.[331], using this phenomenon to model fast gamma-ray flares in the Crab Nebula, argue that "the reconnection layer acts almost as a pure linear accelerator", tending to accelerate electrons and positrons from an e^+e^- pair plasma up to the maximum possible energy, eEL where L is the length of the layer. Particle-in-cell (PIC) simulations by Sironi and Spitkovsky[332] support the contention that relativistic reconnection accelerates particles to high energies, rather than simply raising the temperature of the plasma (see figure 3.11).

The magnetisation of the plasma is described by the parameter σ, defined in SI units as

$$\sigma = \left(\frac{\omega_c}{\omega_p}\right)^2 = \frac{\varepsilon_0 B^2}{nmc^2}, \tag{3.52}$$

where $\omega_c = eB/mc$ is the gyrofrequency, $\omega_p = \sqrt{ne^2/\varepsilon_0 m}$ is the plasma frequency, B is the magnetic field and n is the electron number density (note that most papers in astrophysics will quote ω_p in cgs units as $\sqrt{4\pi ne^2/m}$). As shown in figure 3.11, Sironi and Spitkovsky[332] find that the electron energy spectrum produced by magnetic reconnection depends quite strongly on this parameter, becoming harder as the magnetisation increases. The particle acceleration in this simulation takes place primarily in the magnetic islands.

A key point of acceleration by magnetic reconnection is that it can be *fast*. This is important because some sources show sudden gamma-ray flares suggestive of rapid acceleration of electrons—that is, the flare indicates the presence of a population of relativistic electrons that is presumably not present in the quiescent phase of the source. For example, the Crab Nebula produces short flares of medium-energy (>100 MeV) γ-rays, lasting only a day or so, which appear to be produced by synchrotron emission off a population of PeV-energy particles[331]. The sudden onset of these flares strongly suggests that this population is produced rapidly by some transient event, instead of building up over time as would be the case with diffusive shock acceleration. A large-scale magnetic reconnection event involving a reconnection layer with a length of a few light-days and a reconnecting magnetic field of a few milligauss (a few tenths of a microtesla) produces results broadly consistent with observations[331], as shown in figure 3.12, though it should be noted that this simulation used a test-particle approach and did not consider the response of the field to the accelerated particles.

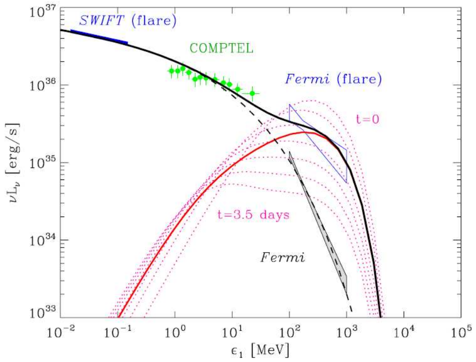


Figure 3.12: Spectral modelling of a γ-ray flare of the Crab Nebula, observed in September 2010. The black dashed line is the quiescent emission, and the red solid line is the emission from the reconnection layer model, averaged over the 4-day duration of the flare. The red dotted lines show the time evolution of the emission. The model[331] assumes that the quiescent emission is caused by a population of $e^{\pm}$ with a power-law spectral index of 2.2 and a high-energy cut-off of $\gamma = 2 \times 10^9$ (1 PeV) in a magnetic field of 200 μG (20 nT). The flare is modelled by introducing a population of $e^{\pm}$ accelerated in a 4-light-day (10^{14} m, 700 AU) long reconnection layer. Figure from Cerutti, Uzdensky and Begelman[331].

Many suspected sites of particle acceleration, particularly blazars and pulsar wind nebulae like the Crab, show rapid flares Magnetic reconnection events may be a good way of explaining these sudden transient episodes.

Magnetic-reconnection-driven particle acceleration may also occur in accretion discs, such as those believed to feed the supermassive black holes at the cores of active galactic nuclei. Differentially rotating accretion discs of magnetised plasma are subject to the *magnetorotational instability*[333], which generates turbulence in the accretion flow and is considered a possible mechanism for transporting angular momentum in the disc. This will induce magnetic reconnection events which can accelerate particles. Hoshino[334] analysed magnetorotational instability using a PIC simulation, and found that a hard spectrum ($\propto E^{-1}$) can be generated, at least up to γ fators of 100 or so. This could explain particle acceleration in the vicinity of massive black holes.

3.8 Propagation of cosmic rays through the Galaxy

Studies of acceleration mechanisms, whether analytical or by simulations, tell us the expected energy spectrum of accelerated particles produced by a source with given properties (such as magnetic field, shock speed, etc.). This can be used to deduce the expected spectrum of synchrotron radiation from the source, which can then be compared with observation. However, the spectrum of accelerated particles produced by the source *cannot* be directly compared with the observed spectrum of cosmic rays in the solar system, because the cosmic rays first have to travel through the Galaxy to reach us. The Galaxy is pervaded by a gaseous interstellar medium and by magnetic fields, so this propagation can affect the slope of the observed spectrum.

The speeds of cosmic rays clearly exceed the escape velocity of the Galaxy. Galactic cosmic rays are confined within the Galaxy by magnetic fields, not by gravity. A simple model of cosmic-ray propagation and escape[335] treats the magnetised volume of the Galaxy as a leaky cylindrical box with radius $r_\mathrm{d} \simeq 15$ kpc, the radius of the Galactic disc, and height $H \simeq 3$ kpc, the extent to which the Galactic magnetic field extends onto the halo (as estimated from radio-frequency synchrotron emission). Cosmic rays will diffuse out of the box on a characteristic timescale

$$\tau_\mathrm{esc} \simeq \frac{H^2}{D(E)},$$

where $D(E)$ is the diffusion coefficient in the Galaxy. We expect the diffusion coefficient to increase with energy: for a given magnetic field, higher-energy particles have larger gyroradii and will therefore escape more easily. Assuming a power law, $D(E) = D_0 E^\delta$, for the diffusion coefficient and another power law, $N_s(E) \propto E^{-\alpha}$, for the cosmic-ray spectrum produced by the sources and injected into the Galaxy, we have for the spectrum of observed *primary* cosmic rays, i.e. those actually produced by the sources,

$$N(E) \simeq \frac{N_s(E)\mathcal{R}_s}{2\pi r_\mathrm{d}^2 H}\tau_\mathrm{esc} \propto E^{-(\alpha+\delta)}, \tag{3.53}$$

where $\mathcal{R}_s$ is the source rate—if, as is generally believed, Galactic cosmic rays are accelerated by young supernova remnants, then $\mathcal{R}_s$ is the rate of supernovae in the Galaxy.

By itself, this does not allow us to distinguish the contributions of α, the source spectral index, and δ, the effect of diffusion. However, as we saw in the previous chapter, some cosmic-ray nuclei are not produced in the sources themselves, but are created by spallation. These nuclei should have a different spectrum,

$$N_\mathrm{sec}(E) \simeq N(E)\mathcal{R}_\mathrm{spall}\tau_\mathrm{esc} \propto E^{-(\alpha+2\delta)}, \tag{3.54}$$

where $\mathcal{R}_\mathrm{spall}$ is the rate of spallation reactions.

Comparing these two spectra, we see that the ratio of secondary to primary particles $N_\mathrm{sec}/N \propto E^{-\delta}$, so by comparing the observed energy spectra of secondary and primary nuclei we should be able to deduce δ and hence infer α.

A suitable pair of secondary and primary nuclei is boron (a secondary nucleus, produced only by spallation) and carbon (a primary nucleus of similar atomic mass). The observed boron-to-carbon ratio is shown in figure 2.17; as noted in the caption, the data below 1 GeV are affected by solar modulation and are not relevant to this question. In principle, fitting the data at higher energies by a power law should give a value for δ; in practice, the different datasets are not completely consistent with each other, and several have very large error bars, so the resulting fit is not very constraining—Amato[335] quotes a range $0.3 < \delta < 0.7$. Comparing this to the spectrum for primary cosmic rays, which has a spectral index of 2.7, gives $2.0 < \alpha < 2.4$.

It should be noted that the "universal" power law for test-particle diffusive shock acceleration, E^{-2}, is consistent with this result, albeit only at the limit of the range. However, it must be said that more sophisticated analyses tend to produce, if anything, energy spectra that are flatter than E^{-2}, whereas studies of emission from supernova remnants tend to prefer spectra that are steeper than E^{-2}[335]. The latter would agree with the observations (for δ somewhat larger than 0.3), but is difficult to reproduce using simulations of diffusive shock acceleration.

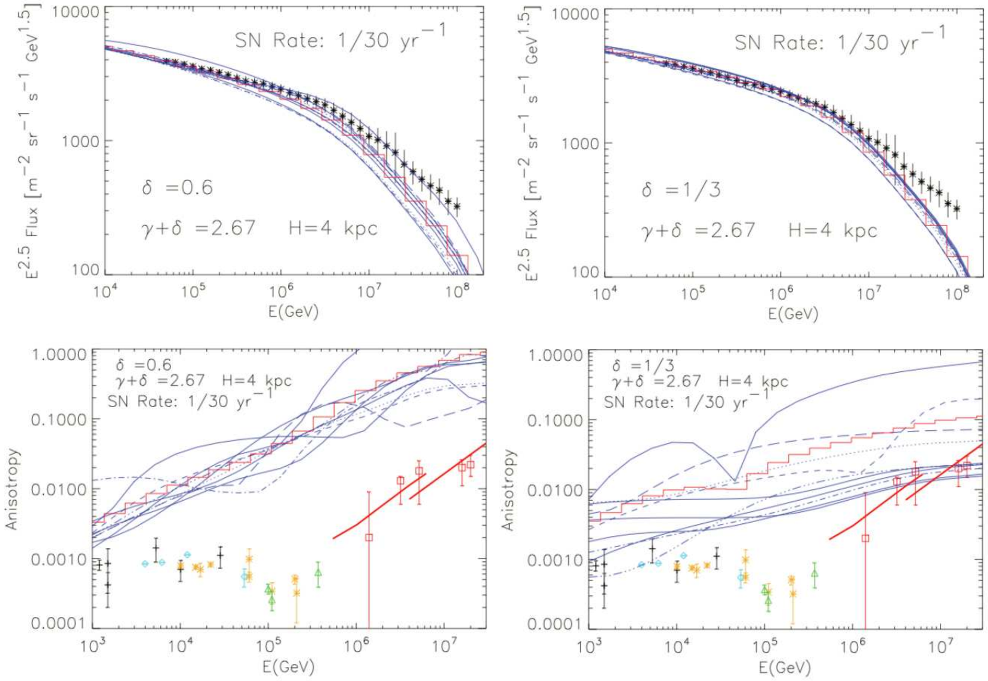


Figure 3.13: Top, fits to the all-particle cosmic-ray spectrum. The points with error bars are an average over the various datasets and are the same in both panels. Both panels assume a supernova rate of one per 30 years and an overall spectral index of 2.67; in the left panel, the injection spectral index is $\alpha = 2.07$ and the propagation index is $\delta = 0.60$, while in the right panel the values are 2.34 and 0.33 respectively. The different curves represent different spatial distributions of the supernovae; the red histogram is the average. Bottom panel, the predictions of the same fits for the anisotropy of the cosmic ray flux compared to a range of different datasets. Figures from Amato[335] (note she uses γ for the injection index, but we have too many variables called γ already).

Figure 3.13 shows the result of modelling Galactic cosmic ray spectra according to two different assumptions about the injection and propagation spectral indices—respectively 2.07 and 0.60 on the left, and 2.34 and 0.33 on the right. As might be expected from equation (3.53), as both hypotheses have the same sum, both fit the energy spectrum quite well up to the "knee" region of 10^6 to 10^7 GeV. However, both overpredict the observed anisotropy of cosmic rays, with the first hypothesis being significantly worse than the second. Amato[335] argues that the overprediction in the bottom right panel can be dealt with in more sophisticated models which take account of the true distribution of young supernova remnants, whereas the discrepancy in the left panel, which is about an order of magnitude worse, cannot be explained away so simply. This therefore favours an injection spectrum somewhat steeper than E^{-2} and a propagation spectral index close to $\frac{1}{3}$. The complexities of the non-linear theory of diffusive shock acceleration are such that it is not in fact too difficult to find a plausible mechanism for generating steeper spectra: the amplification of the magnetic field can lead to a significant difference between the bulk motion of the gas and the motion of the Alfvén waves that generate the magnetic turbulence. As the fast particles actually scatter off the magnetic turbulence and not the gas this affects the resulting spectrum[335, 336], and can perhaps lead to a significant increase in the spectral index.

3.9 Summary

Terrestrial accelerators use electric fields to accelerate particles and magnetic fields to steer them. Large-scale electric fields cannot be maintained over long periods in astrophysical sources, because of the high conductivity of ionised gases, so this model is not directly applicable to astrophysical accelerators. Instead, acceleration is achieved using electric fields induced by time-varying magnetic fields, according to Maxwell's equation $\nabla \times \mathbf{E} = -\partial\mathbf{B}/\partial t$.

The original theory of cosmic-ray acceleration by magnetic fields, Fermi's second-order mechanism[304], shows that random scattering off magnetic irregularities causes a gradual increase in particle energy, $\Delta E/E \propto V^2/c^2$ where $V \ll c$ is the speed at which the magnetic "mirror" is moving. As these speeds are typically rather slow, $\sim$20 km s^{-1} relative to the rotating frame of the Galactic disc (the *Local Standard of Rest*) or $\sim$200 km s^{-1} relative to a non-rotating frame such as the Galactic halo, this rate of energy increase is too slow to be useful in the context of the interstellar medium, though it is possible that it might play a significant role in regions of strong, turbulent magnetic fields.

A more promising variant of Fermi acceleration takes place in the vicinity of strong collisionless shocks. Here, the reflection of the particle from the magnetic turnulence always takes place at favourable geometry, so that the rate of acceleration is $\propto V/c$ instead of V^2/c^2; this mechanism is therefore sometimes called *Fermi first-order acceleration*, although the more descriptive term "diffusive shock acceleration" (DSA) is preferred in more recent literature. A simple test-particle approach to acceleration by repeated shock crossings predicts an energy spectrum $N(E) \propto E^{-2}$ independent of the details of the shock: this so-called "universal power law" is a key finding, because the evidence of the observed cosmic-ray power spectrum suggests that different sources—specifically, Galactic sources below $\sim 10^{15}$ eV and extragalactic sources above that energy—do indeed generate very similar power laws. Given the near-ubiquity of collisionless shocks in regions where particle acceleration is expected (because of the presence of synchrotron radiation) and the promising universality of the test-particle spectral index, DSA is generally regarded as the most likely acceleration mechanism for the majority of cosmic-ray sources. Collisionless shocks of various kinds occur in the solar system, where they can be studied directly by spacecraft, and all such shocks are found to accelerate particles.

In reality, the simple test-particle approach cannot be justified as a good approximation. The inclusion of magnetic fields into the shock jump conditions complicates the analysis significantly, but the principal issue is that DSA is actually a fairly efficient acceleration mechanism ($\sim$10–20% of the thermal energy converted into energy of a population of accelerated non-thermal particles), so the effect of the fast particles on the shock cannot be neglected. The resulting feedback loop—the shock affects the particles; the particles in turn affect the shock—means that the equations describing the response of the system are non-linear and must be solved numerically. Detailed 3D particle-in-cell simulations are needed to model the system satisfactorily, and the improvement in such models as the available CPU power increases is leading to a better understanding of their behaviour.

An alternative model of particle acceleration, which occurs in the solar system in association with solar flares and coronal mass ejections, is the phenomenon of magnetic reconnection. Here, adjacent magnetic field lines of opposing polarity break and re-form, creating large transient electric fields which can accelerate particles rapidly to high energies. Magnetic reconnection is still not particularly well understood, but studies suggest that this mechanism might be important in very highly magnetised environments, particularly pulsar wind nebulae and the jets of radio-loud active galactic nuclei. The rapidity with which particles can be accelerated by this mechanism makes it especially attractive when attempting to model the sudden γ-ray flares observed in such objects, which appear to require the acceleration of $e^{\pm}$ to PeV energies over a comparatively short period.

Finally, in considering the *observed* spectrum of cosmic rays, the effect of propagation through the Galaxy cannot be neglected. Comparisons of secondary nuclei produced through spallation with neighbouring primary nuclei accelerated by the sources suggests that the effect of propagation is to steepen the spectrum by $\sim$0.3–0.7. This may account for the difference between the test-particle power law E^{-2} and the observed cosmic-ray powr law $E^{-2.7}$, though there is evidence that the spectrum produced by the sources is already somewhat steeper than E^{-2}.

In summary, our exploration of the observed properties of high-energy astroparticles, both directly (cosmic rays, γ-rays, and high-energy neutrinos) and indirectly (radio-frequency synchrotron radiation produced by high-energy electrons), together with the study of possible acceleration mechanisms in this chapter, leads to a reasonably self-consistent picture of high-energy astrophysics.

- Charged particles of all types (electrons, protons and heavier nuclei) are accelerated in certain classes of astrophysical objects; some of these escape from the source regions to be observed as cosmic rays.

- Most of these particles probably acquire their high energy through diffusive shock acceleration, although magnetic reconnection may play an important role in some sources.

- High-energy photons and neutrinos are secondary products of high-energy cosmic rays, generated either by the decay of pions produced by interactions of high-energy protons with ambient radiation or gas or—in the case of photons—by inverse Compton scattering of low-energy photons off high-energy electrons.

- Direct evidence from the chemical composition of cosmic rays (see figure 2.14), and theoretical indications from calculations of the escape time, both indicate that the "knee" in the cosmic-ray energy spectrum is probably associated with a shift from Galactic to extragalactic sources.

- Young supernova remnants and pulsar wind nebulae are the favoured candidates for the sources of Galactic cosmic rays, as they are known to be sources of TeV photons and appear to provide appropriate conditions for diffusive shock acceleration.

- The likeliest candidate sources for extragalactic cosmic rays are radio-loud active galactic nuclei, which are the dominant extragalactic sources of TeV photons, and/or gamma-ray bursts.

- Owing to energy degradation through production and decay of the $\Delta(1232)$ resonance, ultra-high-energy cosmic rays ($E > 10^{19}$ eV or so) cannot be produced at cosmological distances, but must originate from the fairly local universe (within $\mathcal{O}(100)$ Mpc). Likewise, the range of TeV photons is limited by pair-production, $\gamma\gamma \rightarrow e^+e^-$, off ambient low-energy photons, so the failure to observe TeV photons associated with gamma-ray bursts is not evidence that such photons are not produced.

This model of high-energy astrophysics gives us a clear short-list of candidate sources to consider. In the next chapter, we shall investigate these source classes in more detail.

3.10 Questions and Problems

1. The mean free path for cosmic rays in the interstellar medium is of order 0.1 pc, and the typical velocity of an interstellar cloud relative to the Local Standard of Rest is of order 20 km s^{-1}. Estimate how how it would take to accelerate a proton from an initial kinetic energy of 200 MeV to the "knee" of the cosmic ray spectrum at about 3×10^{10} MeV.

2. The diffusion coefficients for a relativistic particle ($v \simeq c$) in a magnetic field B with fluctuations of size δB are given by[335]

$$D_{||}(p) = \frac{4}{3\pi}\left(\frac{B}{\delta B}\right)^2 cr_g; \qquad D_{\perp}(p) = \frac{\pi}{12}\left(\frac{\delta B}{B}\right)^2 cr_g, \tag{3.55}$$

   where r_g is the gyroradius p/eB. In the Galactic magnetic field, we expect that the size of the fluctuations is given by[335]

$$(\delta B)^2 = B^2 (r_g/L)^{2/3}, \tag{3.56}$$

   where $D_{||}$ is the diffusion coefficient parallel to the direction of the underlying magnetic field, $D_{\perp}$ is the coefficient perpendicular to the field, the characteristic size L is of the order of 50–100 pc, and the magnetic field of the Galaxy B is of order 5 μG (0.5 nT). Use these values and the discussion in section 3.8 to calculate the escape time for cosmic ray protons at a few energies from 10 GeV to 10 PeV. Plot your results and comment.

3. By considering equations (3.30) in the case where $\mathbf{B}_t = 0$, check that the magnetic field does not contribute to the shock jump conditions in this case. Compare your result to equations (3.17), (3.18) and (3.21), and explain any differences.

4. Use equation (3.11) to calculate $\langle(\Delta E)^2\rangle$.

5. One of the seminal papers on diffusive shock acceleration is AR Bell (1978)[337]. In this question we work through Bell's derivation of the energy spectrum.

   Defining terms as in figure 3.3, and working in the rest frame of the upstream gas, Bell argues that the energy of a particle after $k+1$ shock crossings is, in terms of its energy after k shock crossings,

$$E_{k+1} = E_k \frac{1 + (v_{k1}\cos\theta_{k1})(u_1 - u_2)/c^2}{1 + (v_{k2}\cos\theta_{k2})(u_1 - u_2)/c^2}, \tag{3.57}$$

   where v_{k1} is the velocity with which the particle crosses from upstream to downstream, θ_{k1} is the angle this velocity makes with the shock normal, and v_{k2} and θ_{k2} are the equivalent quantities for the return crossing. (Note that $cos\theta_{k2} < 0$ in this frame.)

(a) Justify equation (3.57).

(b) By taking logs of this equation, show that the energy of the particle after ℓ shock crossings, compared to its initial energy E_0, is given by

$$\ln\frac{E_\ell}{E_0} = \sum_{k=0}^{\ell-1} \ln\left[1+\frac{u_1-u_2}{c}\cos\theta_{k1}\right] - \sum_{k=0}^{\ell-1} \ln\left[1+\frac{u_1-u_2}{c}\cos\theta_{k2}\right], \tag{3.58}$$

assuming that $v_{k1} \simeq v_{k2} \simeq c$.

(c) Assuming that it is reasonable to replace the sums in equation (3.58) by ℓ times the average, show that this equation can be written as

$$\ln\frac{E_\ell}{E_0} = \frac{4}{3}\ell\,\frac{u_1-u_2}{c}. \tag{3.59}$$

(d) By the argument given in section 3.5.1, the escape probability per crossing is $4u_2/c$. Hence show that equation (3.59) leads to an energy spectrum

$$N(E)\,\mathrm{d}E = \frac{\alpha-1}{E_0}\left(\frac{E}{E_0}\right)^{-\alpha}\mathrm{d}E \tag{3.60}$$

where $\alpha = (2u_2+u_1)/(u_1-u_2)$.

(e) Show that $\alpha = (r+2)/(r-1)$ where r is the compression ratio, $r = \rho_2/\rho_1$.

6. Explain why pulsar wind nebulae are considered to be likely sites for acceleration by magnetic reconnection, whereas shell supernova remnants are not.

Chapter 4

Astrophysical Accelerators: Sources

4.1 Introduction

So far, we have considered the observational evidence for the presence of relativistic particles in some astrophysical sources, and discussed possible mechanisms by which electrons, protons and heavier ions could be accelerated to these high energies. In this chapter, we shall consider the problem from the other side, and investigate how the mechanisms we have discussed might be implemented in astrophysical objects.

In general, the ability of a given type of astrophysical source to accelerate charged particles to some energy E is limited by several factors:

1. for mechanisms such as diffusive shock acceleration which depend on repeated passage of the particle through the accelerating region, the particle must remain confined in the source at least until it reaches energy E;

2. the time taken to accelerate particles to energy E must not exceed the lifetime of the source;

3. the rate at which the particle gains energy from the acceleration mechanism must exceed the rate at which it loses energy from processes such as synchrotron radiation, bremsstrahlung, etc.

For diffusive shock acceleration, satisfying the first criterion is primarily a matter of ensuring that the gyroradius of the particle does not take it outside the source. It is therefore proportional to the magnetic field B and the source size R; numerically[339]

$$E_{\max} \lesssim 10^{24} v Z \left(\frac{B}{\mathrm{G}}\right)\left(\frac{R}{\mathrm{kpc}}\right)\ \mathrm{eV}, \tag{4.1}$$

where Z is the atomic number and v is the speed of the shock measured in units of c. A log-log plot of magnetic field versus characteristic source size is known as a *Hillas plot*, after Prof. Michael Hillas of Leeds University who first presented it as a diagnostic[338]; on such a plot, the condition for a source's being able to reach a given energy E is that it lies above the diagonal line $B = 10^{-24}E/(vR)$. A current version of the Hillas plot[339] is shown in figure 4.1.

The second criterion is particularly significant for transient events such as gamma-ray bursts—clearly the energy attained during an explosive event which is over in seconds is likely to be limited by the available time—but may also be relevant in more long-lasting temporary phenomena such as supernova remnants if the rate at which particles gain energy is slow.

Finally, the energy loss is typically dominated by synchrotron emission. This imposes a maximum energy which can be derived by integrating the energy loss from synchrotron emission, equation (2.40) over a time corresponding to the characteristic size of the source R. The result is[339]

$$E_{\max} \lesssim 3 \times 10^{16} \frac{A^4}{Z^4} \left(\frac{B}{\mathrm{G}}\right)^{-2} \left(\frac{R}{\mathrm{kpc}}\right)^{-1} \ \mathrm{eV}, \tag{4.2}$$

where A is the atomic mass of the ion in question and Z is its atomic number. For protons, this gives an upper limit $B = 1.7 \times 10^8/\sqrt{R}$ on the Hillas plot. Combining this with the lower limit given by the gyroradius criterion in equation (4.1) yields the blue triangular areas in figure 4.1; the red area is the same calculation for iron nuclei ($Z = 26$, $A = 56$).

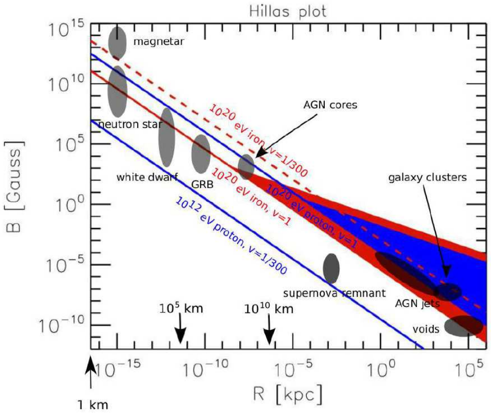


Figure 4.1: Hillas plot showing astrophysical objects which are potential sources of cosmic rays[339]. The diagonal lines represent the minimum requirement for the given particle type, energy and shock speed v (expressed in units of c); the blue and red wedges represent the allowed regions for 10^{20} eV protons and iron nuclei, respectively, when synchrotron radiation losses are taken into account. From this plot, it appears that non-relativistic shocks, e.g. in supernova remnants, are capable of accelerating most cosmic rays, with energies $< 10^{16}$ eV or so, but the highest-energy cosmic rays require relativistic shocks.

It can be seen from figure 4.1 that many Galactic sources are in principle capable of accelerating protons to energies in excess of 10^{12} eV in the vicinity of non-relativistic shocks. Essentially no Galactic sources can accelerate the very highest-energy cosmic rays, with energies $\sim 10^{20}$ eV, because the gyroradius is too large, and even extragalactic sources probably cannot accelerate protons to such energies without the involvement of relativistic shocks. Iron nuclei, and heavy nuclei in general, have less stringent constraints because of their greater charge, which makes them easier to confine.

The Hillas plot provides constraints on possible sources. As we have seen in previous chapters, positive indications of the presence of accelerated particles in astrophysical objects are provided by the detection from such objects of electromagnetic emission implying their presence—chiefly synchrotron radiation, which is produced by relativistic electrons, and γ-ray emission, whose production by inverse Compton scattering or π^0 decay requires the presence of relativistic electrons or hadrons respectively. This diagnostic focuses our attention on supernova remnants and pulsar wind nebulae among Galactic sources, and gamma-ray bursts and radio-loud active galactic nuclei beyond the Galaxy. In addition, it is worth looking more closely at the solar system: although particle acceleration in the solar system is on a very modest scale, it is the only place where collisionless shocks and the associated processes can be observed directly *in situ*, rather than indirectly via their electromagnetic signatures.

4.2 Particle acceleration in the solar system

The Sun constantly emits a stream of charged particles, the solar wind, from its upper atmosphere. The solar wind has two components, the slow solar wind and the fast solar wind, but both are supersonic and therefore likely to induce shocks. In fact, many different types of collisionless shocks are seen in the solar system (see figure 4.2), and all of them seem to be associated with particle acceleration.

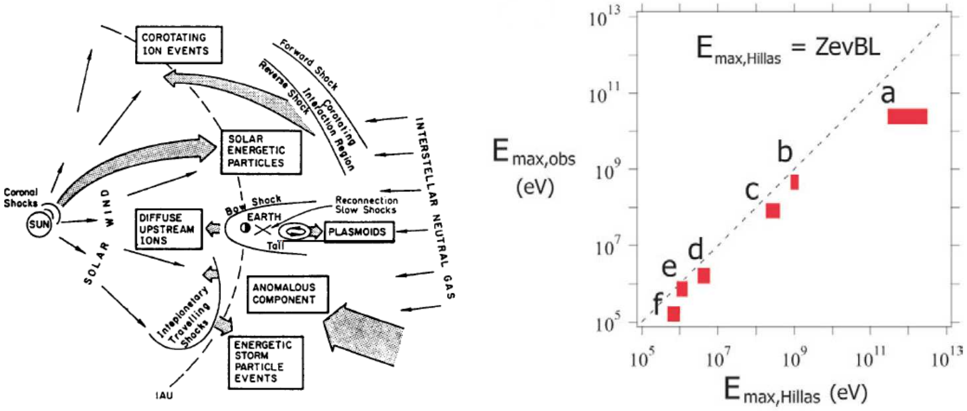


Figure 4.2: Left panel, sketch of shocks and particle acceleration in the inner solar system, from Scholer (1984)[340]. This does not show the termination shock where the solar wind hits the interstellar medium, which has been observed by the Voyager spacecraft[307] and is the likely source of the anomalous cosmic rays. Right panel, Hillas plot for shocks in the solar system[341]. The plot shows the observed and predicted maximum particle energies for (a) solar flares and CME-drive shocks, (b) the Earth's Van Allen belt[343], (c) heliospheric shocks (producing energetic storm particles, ESPs, and anomalous cosmic rays), (d) the Earth's bow shock, (e) the Earth's magnetotail and (f) the Earth's foreshock.

The speed of the solar wind is quite modest—400 km s^{-1} for the slow component and 750 km s^{-1} for the fast—and the energies to which particles are accelerated are consequently quite low, typically tens of MeV per nucleon, though higher energies, up to tens of GeV, are seen in association with coronal mass ejections[341]. The key advantage of solar-system shocks is that the shock front can be observed directly by spacecraft, instead of relying on indirect indications such as the presence of synchrotron radiation. The Earth's bow shock and shocks driven by coronal mass ejections are particularly well studied, because of their potential implications for terrestrial electromagnetic equipment ("space weather"[342]), but other planetary bow shocks and the solar wind termination shock have also been investigated.

The properties of the energetic particles associated with solar system collisionless shocks are broadly consistent with expectations from diffusive shock acceleration[341, 344]. Figure 4.2 shows a comparison of the Hillas criterion, $E_{\rm max} \simeq ZevBL$ where v is the speed of the shock and L its characteristic size, with the observed maximum particle energies associated with various solar-system shocks. With the exception of CME shocks, which do not achieve the expected particle energies (perhaps they are too short-lived), the agreement between the order-of-magnitude Hillas estimate and the observations is surprisingly good.

4.2.1 Solar flares and coronal mass ejections

The association between solar flares and accelerated particles (so-called *solar energetic particles* or SEPs) was first noticed in the 1940s[345]. These events were observed from the ground, as changes in the atmospheric ionisation produced by cosmic rays, and were initially interpreted as "charged particles actually being emitted by the Sun"[345]. More detailed investigation, however, pointed to a picture where the particles were produced in magnetohydrodynamic shock waves accompanying solar flares, rather than directly by the Sun in the flare event.

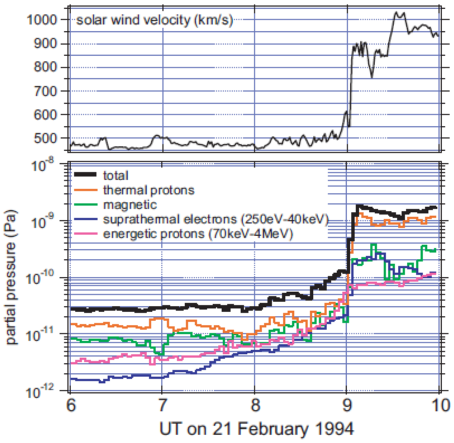


Figure 4.3: Solar wind velocity (top) and partial pressure for different components (bottom) for an interplanetary shock detected at 09:03 UT on 21 February 1994 [341]. The precursor plus subshock structure expected if the accelerated particles modify the shock can be seen in the distribution of suprathermal electrons (blue) and energetic protons (magenta) in the bottom panel. Compare this with figure 3.8.

Solar flares, coronal mass ejections and similar phenomena are typically associated with bursts of radio noise from the Sun. These bursts are classified into four types[346, 323] depending on frequency range and duration. Proton-rich SEP events typically accompany[347] Type II and some Type IV radio bursts, both of which (but not the other types) are associated with magnetohydrodynamic shock waves. Further work, summarised by Desai and Burgess[347], distinguished between electron-only events (associated with Type III radio bursts and solar flares) and "mixed" events containing protons and nuclei as well as electrons, associated with Type II and Type IV bursts. A more modern classification divides SEP events into "impulsive" events, lasting hours, restricted to a longitude range of < 30°, dominated by electrons (electron/proton ratio 10^2–10^4 and associated with a short (<1 hr) X-ray flare, and "gradual" events, lasting days, covering ∼180° of longitude, less electron-dominated (ratio of 50–100) and associated with a long X-ray flare (>1 hr). Impulsive events are associated with solar flares and are common (∼1000/year); gradual events are associated with coronal mass ejections and interplanetary shocks and are rare (∼10/year). It is believed that different acceleration mechanisms operate in these two different event types: diffusive shock acceleration at CME-driven interplanetary or coronal shocks for gradual events, and acceleration by magnetic reconnection in the flare for impulsive events.

The compositions of impulsive and gradual SEP events differ by more than simply the electron fraction. Gradual events have an elemental composition broadly similar to the solar wind in general, consistent with a seed population consisting of the ambient gas—coronal material for shocks within the solar corona proper, solar wind for interplanetary shocks—whereas impulsive events are enriched in unusual species, most notably ^{3}He (which is normally *extremely* rare, but can actually be more abundant than ^{4}He in impulsive SEP events[347]), but also heavy ions such as iron: the Fe/O ratio is also much higher in impulsive SEP events than in the solar wind. Neither type of event accelerates heavy ions enough to strip them completely, but the ionisation state of iron in impulsive events (up to +20) is higher than in gradual events (up to +14)—though the latter does increase to +20 as the energy of the iron ion increases. This is all consistent with the idea that the seed material for impulsive events is not typical coronal or solar-wind material, but has been pre-heated to very high temperatures (10^7 K) before acceleration. This would be understandable in the context of magnetic reconnection within a solar flare.

The modification of the original shock front caused by the accelerated particles has been observed in interplanetary shocks; an example is shown in figure 4.3 [341]. This shows that observations of shocks in the solar system can potentially be used to test details of non-linear DSA simulations that cannot be easily studied at interstellar distances.

4.2.2 Planetary bow shocks

Planetary bow shocks occur when the supersonic solar wind encounters the obstacle presented by the planet. Most planets—Mercury, Earth and the gas giants—have substantial magnetic fields, so the "obstacle" in question is the planetary magnetosphere rather than the planet proper. Venus and Mars, which do not have strong magnetic fields, still have bow shocks, because the planetary ionosphere is conductive and can deflect the flow. The Moon, which has neither magnetic field nor ionosphere, does not have a bow shock[348]: the solar wind hits the lunar surface and is absorbed by it.

The most extensively studied planetary bow shock is of course the Earth's (see figure 4.4). This has been observed over long periods by a multiplicity of spacecraft, giving high resolution data, often from multiple viewpoints since more than one spacecraft is observing at any given time. A disadvantage of this is that the bow shock is approximately a standing shock from the Earth's point of view[1] and the spacecraft are generally in Earth orbit or at the L1 point, so they have rather slow speeds relative to the shock, meaning that one does not get the clean cross-section of properties across the shock seen when a probe such as *Voyager* crosses a bow shock at high speed. Nevertheless, the long-term detailed observations of the Earth's bow shock by various spacecraft provide an invaluable resource for the study of

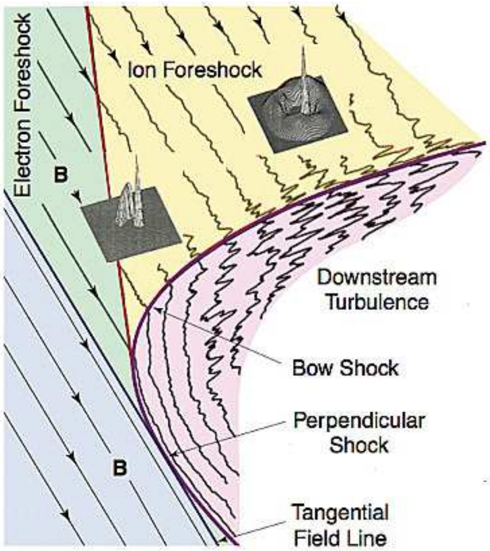


Figure 4.4: The Earth's bow shock, from Desai and Burgess[347]. Note how the shock orientation changes from quasi-parallel at the top of the diagram to quasi-perpendicular on the left. Note the magnetic field turbulence that provides the scattering for diffusive shock acceleration; this extends into the foreshock as expected for realistic (not test-particle) shocks. The small 3D plots show the ion velocity distributions in the DSA region (top) and close to the field-aligned boundary (left); the spike in these plots is the unaltered solar wind speed.

[1] It is not entirely stationary, because of the variability of the solar wind.

collisionless shocks.

The Earth's bow shock is a supercritical shock, with Mach number of order 10 (varying from 6–12 depending on solar wind conditions[349]), so we expect it to accelerate particles. It is a strong shock (compression ratio ∼4), and also clearly a magnetised shock, which means that the magnetohydrodynamic shock jump conditions of equations (3.30) must be applied; furthermore, because the shock front is curved, its orientation relative to the ambient magnetic field varies from quasi-parallel to quasi-perpendicular along the shock front, as shown in figure 4.4. Spacecraft that have made observations relevant to studies of the Earth's bow shock include ACE[83], CLUSTER[350], IBEX[141], STEREO[351] and *Wind*[352], among many others.

The Earth's bow shock is associated with a number of distinct populations of non-thermal ions and electrons[347]. The energies involved are not large, ranging from about 1 keV to 1 or 2 MeV, but are high enough to implicate the bow shock in particle acceleration. These suprathermal particles are backscattered upstream of the shock, producing both a diffuse population upstream of the quasi-parallel shock region and low-energy collimated beams aliogned along the field lines (hence known as *field-aligned beams*) originating from the quasi-perpendicular part of the shock.

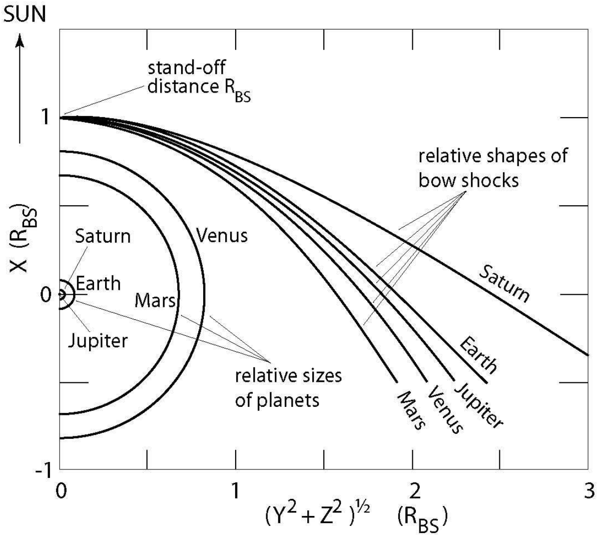

SUN
stand-off distance R_{BS}
1
relative shapes of bow shocks
X (R_{BS})
Saturn
Venus
Saturn
0
Earth
Mars
Jupiter
Earth
relative sizes of planets
Jupiter
Mars
Venus
-1
0
1
2
3
$(Y^2+Z^2)^{½}$ (R_{BS})

Figure 4.5: Schematic of planetary bow shocks, scaled to the same stand-off distance $R_{\rm BS}$[349]. Note the contrast between Mars and Venus, which have small or non-existent magnetospheres, and Earth, Saturn and Jupiter (with this scaling, Jupiter shrinks to the dot at the origin).

Diffusive shock acceleration in the quasi-parallel part of the shock appears to account for the properties of the diffuse population of upstream ions[347]. Observations of the field-aligned beams by CLUSTER suggest that they are produced by reflection off the shock front in the quasi-perpendicular shock, i.e. shock drift acceleration. The Earth's bow shock is an interesting laboratory for studying the transition from quasi-parallel to quasi-perpendicular shocks, although Desai and Burgess[347] comment that its small size means that different regions are not well separated, and its closeness to the Earth introduces the possibility that some energetic ions do not originate in the shock at all but have escaped from the Earth's magnetosphere.

The outer giant planets all have strong magnetic fields, and the solar wind is comparatively weaker owing to their greater distance from the Sun. Therefore they all have very large magnetospheres and strong bow shocks at a large stand-off distance[348, 349]. The bow shocks are qualitatively similar to the Earth's, though much larger, but more of the shock is quasi-perpendicular because the solar magnetic field becomes less radial and more azimuthal as we move further from the Sun. All the outer planet bow shocks have been crossed by spacecraft, although in the cases of Uranus and Neptune our information is limited to one fast flyby by *Voyager* 2.

The bow shocks of the non-magnetic terrestrial planets Mars and Venus have also been extensively studied. As can be seen from figure 4.5, these shocks are much closer to the planet than those associated with magnetospheres, being generated by the conductivity of the ionosphere. The stand-off distance for both Mars and Venus is about 1.5 times the planet's radius[349], nearly a factor of 10 closer than the Earth's bow shock at 12–14 Earth radii. The interaction of the solar wind with the Martian atmosphere has been suggested as the reason for the loss of the dense early atmosphere implied by evidence of past surface water; however, in view of the similarity of the bow shocks of Mars and Venus (where the atmosphere clearly has not been stripped!), other factors must also be acting: a combination of severe early bombardment and later gradual losses seems likely[354].

Of the minor bodies of the solar system, asteroids are similar to the Moon in having neither atmosphere nor magnetic field, and are not likely to maintain significant bow shocks, but comets develop large gaseous envelopes which will interact with the solar wind and generate a bow shock not unlike that of Venus. The bow shock of comet P/Halley was observed by the *Giotto* probe in 1986.

4.2.3 The termination shock

The solar wind termination shock is where the solar wind slows to subsonic speeds under the pressure of the interstellar medium. Both *Voyager* spacecraft have crossed it—interestingly, at quite different heliocentric distances (94 AU for *Voyager* 1, only 84 for *Voyager* 2), which suggests that the termination shock is noticeably asymmetric[355]. though the two spacecraft crossed three years apart (in December 2004 and August 2007 respectively) so variation with the solar cycle may also have contributed.

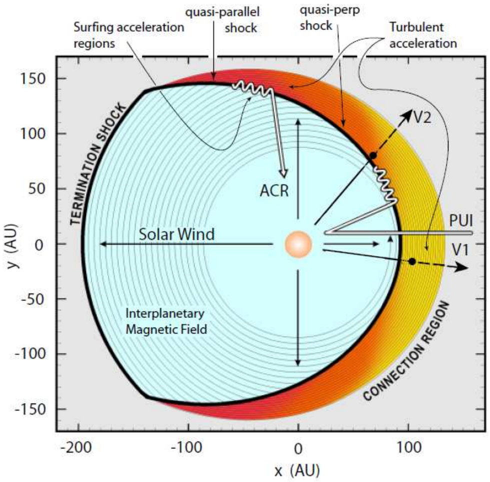


Figure 4.6: A "blunt nose" model of the termination shock[356], showing the exit paths of *Voyagers* 1 (V1) and 2 (V2) In this model the anomalous cosmic rays (ACRs) are produced by shock surfing—a variant of shock drift acceleration—in certain regions of the shock where the magnetic field geometry is favourable, and were not seen by *Voyager* 1 which crossed at the wrong place. Pickup ions (PUI) drift in from the local interstellar medium as neutral atoms before being ionised and picked up by the solar wind; it is these ions that are believed to be accelerated to form the anomalous cosmic rays.

Because of the spiral shape of the solar wind, the termination shock is quasi-perpendicular with respect to the solar wind magnetic field if it is spherical, though in "blunt-nose" models such as that in figure 4.6 there are quasi-parallel regions. Comparison of the termination shock with Neptune's bow shock[307] shows that the termination shock is weaker than the bow shock, which may be the result of the transfer of energy to *pickup ions*. Pickup ions are formed when neutral atoms drift into the heliosphere from the local interstellar medium and are subsequently ionised and picked up by the solar wind. It is believed that the *anomalous cosmic rays* are produced by the acceleration of these pickup ions at the termination shock, but it is unclear exactly hoe this is achieved. Giacolone and Decker[357] conclude from hybrid simulations (in which the ions are treated as particules and the electrons as a fluid) that shock drift acceleration can account for the low-energy anomalous cosmic rays in the energy range 40 keV to 5 MeV; the distribution of ions in this energy range was observed to peak as the *Voyagers* crossed the termination shock. The distribution of higher-energy ACRs did not peak at the shock crossing[355], but continued to rise: if these higher-energy particles are also accelerated by the termination shock, this acceleration is not taking place in the regions that the *Voyagers* crossed. A possible explanation for this may be that diffusive shock acceleration of anomalous cosmic rays does take place and does extend up to the highest observed energies, but does not do so near the nose of the shock because the structure of the magnetic field ensures that they are swept away before they have time to reach the higher energies[358]. Alternatively, acceleration to higher energies may take place away from the shock nose because of a more favourable magnetic field geometry[359]. Other explanations include acceleration by magnetic reconnection in the heliosheath (i.e. outside the termination shock)[360]. It is clear that this question is far from resolved: plausible mechanisms exist, but the evidence does not at present allow us to decide which is or are correct.

4.3 Galactic sources

By calculating the gyroradius associated with the Galactic magnetic field of a few microgauss (a few tenths of nanotesla), we can see that the vast majority of charged cosmic rays are likely to originate within the Galaxy. It is believed that the principal sources of Galactic cosmic rays are supernova remnants (SNRs): although direct proof is lacking because of the non-directional nature of charged cosmic rays, the circumstantial evidence in favour of this is extremely strong; Blasi[320] refers to it as “the supernova remnant paradigm”, the implication of the word ‘paradigm’ being a generally accepted assumption in the field. Two types of supernova remnant are implicated: the standard “shell” supernova remnant, consisting of a roughly spherical shell of gas representing the ejected envelope of the former star, and *pulsar wind nebulae*, “filled” supernova remnants where the interior of the remnant is energised by a central pulsar. Some SNRs have both an outer spherical shell and a filled interior: these are known as “composite” supernova remnants.

4.3.1 Supernovae and supernova remnants

Supernovae are exploding stars, in contrast to ordinary novae which are thermonuclear events on the surface of a white dwarf in an accreting binary, and which leave the original white dwarf essentially unchanged. This distinction was first made explicitly by Baade and Zwicky in 1934[361], although awareness that some “novae” were much brighter than typical examples had been gradually dawning over the preceding two decades, fuelled by the contrast between the (super)nova S Andromedae in M31, which reached a peak visual magnitude of 5.85 in August 1885[362], and faint photographic novae in M31, which only reached magnitudes of 16 or 17[363]. Baade and Zwicky presciently argued both that “super-novae” might be the result of a star collapsing to a neutron star (recall that the neutron had only been discovered two years earlier!) and that they might be the sources of cosmic rays.

Classification of supernovae						
No H			Early H	H always present		
Si II	No Si II		↓	Light curve		Narrow
↓	He	No He	↓	Plateau	Linear	lines
Ia	**Ib**	**Ic**	**IIb**	**II-P**	**II-L**	**IIn**

Table 4.1: Principal classes of supernovae[364]. Although the main *observational* distinction is between those that do not display hydrogen lines (Type I) and those that do (Type II), the *physical* distinction is between Type Ia (exploding white dwarf) and all the rest (massive star core collapse).

As summarised in table 4.1, supernovae are classified principally according to their spectral features, with some input from the shape of the light curve. Astronomical nomenclature is notorious for its lack of logic—caused by the fact that the names are established before the physics is understood—and supernova classification is no exception: the physical division is not between Type I and Type II, but between Type Ia and everything else. It is now known that SNe Ia are the result of explosive carbon burning in a carbon/oxygen white dwarf that has exceeded the Chandrasekhar mass limit of $1.4M_\odot$, whereas all the other types represent the core collapse of an evolved massive star. The lack of hydrogen in Types Ib and Ic, and the very small amount present in Type IIb (where hydrogen lines are present early on but rapidly disappear, leaving a spectrum similar to Type Ib) are understood as resulting from the star's having lost its hydrogen envelope prior to the explosion, either by stellar winds as in Wolf-Rayet stars or by mass loss to a binary companion.

Both types of supernova will leave behind an expanding shell of gas, but only core-collapse supernovae leave a compact object (usually a neutron star, sometimes a black hole); in Type Ia supernovae, the white dwarf is completely disrupted by the explosion, leaving no compact remnant.

The first supernova remnant to be identified as such was the Crab Nebula, which was associated with the "guest star" observed by the Chinese in 1054, tentatively by Hubble[365] in 1928, and securely by Mayall and Oort[366] in 1942; Mayall and Oort also present strong arguments for identifying the 1054 object as a supernova[2]. In modern terminology, the Crab is a pulsar wind nebula, not a classic shell-type supernova remnant, but other historically-attested Galactic supernovae, such as Tycho's (SN 1572) and Kepler's (SN 1604), have shell SNRs.

Estimates using the unstable isotope ^{26}Al[367] suggest a rate of 1.9 ± 1.1 core-collapse supernovae (CCSNe) per century in the Milky Way. This is not inconsistent with the rate of historically observed supernovae: the last two observed Galactic supernovae were Tycho's and Kepler's, in 1572 and 1604, but there have been at least two since then that were not seen (Cas A in about 1680, and G1.9+0.3 around the turn of the 20th century), and because massive star supernovae occur in the spiral arms, which are heavily obscured by dust from our viewpoint, we are not observing the whole Galaxy. Adding in SNe Ia, which do not make ^{26}Al, suggests an overall rate of about 2.5 supernovae per century, or one every 40 years on average. This would supply enough energy to maintain the Galactic cosmic ray flux, provided that supernovae convert about

[2] This was not obvious at the time, although the subsequent discovery of a young pulsar in the Crab proves it beyond doubt, and also establishes that the supernova in question was a massive star core collapse.

10–20% of their available energy into cosmic rays[368]; this level of efficiency is consistent with simulations of diffusive shock acceleration.

4.3.2 Evolution of a supernova remnant

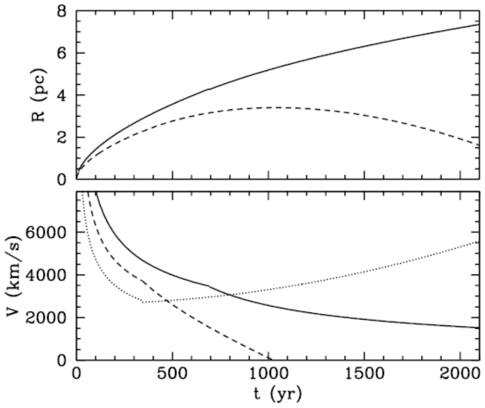


Figure 4.7: Evolution of shock radii and velocities in a shell-type supernova remnant[369]. The solid line is the forward shock and the dashed line the reverse shock; in the lower panel, the dotted line is the reverse shock velocity in the frame of the ejecta. The model is from Truelove and McKee[371], adjusted by Vink to match Kepler's supernova.

With the important exception of long-soft γ-ray bursts (see later), the actual supernova explosion is not crucial to the production of high-energy particles: instead, we are concerned with the aftermath of the explosion. The ejected material—the stellar envelope in the case of core-collapse supernovae, the reprocessed remains of the entire star for SNe Ia—streams away from the star at highly supersonic speeds, creating a shock front, the *forward shock*. This will collide with and sweep up the surrounding interstellar medium, decelerating as it does so. The effect of this is to set up a second shock—the *reverse shock*—which moves back into the ejecta. Despite its name, the "reverse" shock initially moves outwards in the observer's reference frame, though inwards with respect to the expanding ejecta[369]; it will, however, reverse direction after $\sim$1000 years as the pressure of the shock-heated ejecta behind it drives it inwards (see figure 4.7). Both forward and reverse shocks can in principle accelerate particles, although the absence of electron-capture isotopes like ^{59}Ni from observed cosmic rays suggests, as discussed in section 2.2.3, that the material accelerated is predominantly swept-up ISM rather than newly synthesised elements from the SN proper.

The subsequent evolution of the supernova remnant is divided into three phases: the *free expansion* phase, the *Sedov* or *Sedov-Taylor* phase and the *radiative* phase[368]. Free expansion is the period when the forward shock has accumulated relatively little ambient interstellar medium and is still travelling at approximately constant speed. When the forward shock has swept up a mass of interstellar material comparable to the ejecta mass, the remnant enters the Sedov phase, during which the shock is decelerating significantly. Finally, when the forward shock velocity slows to about 200 km s^{-1}, the post-shock temperature has decreased to the point at which nuclei are no longer fully ionised and spectral line emission becomes a significant means of energy loss: this is the radiative phase. Particle acceleration is likely to be confined to the first two phases, since in the radiative phase the shock velocity is low and much of its energy is being dissipated by radiative losses.

The timescale for these phases can be estimated for Type Ia supernovae by assuming that the surrounding ISM typically has a number density of $n_{\mathrm{ISM}} = 1$ cm^{-3} and a temperature of 1 eV ($\sim 10^4$ K)[368]—these numbers are of the right order of magnitude for the "warm interstellar medium" or "intercloud medium"[370]—and that the supernova ejects around $1M_\odot$ of material with a total energy of 10^{44} J (10^{51} ergs; the equivalent of converting 0.06% of a solar mass into energy). Then the initial speed of the supernova blast wave is

$$V_0 = \sqrt{\frac{2E_{\rm SN}}{M_{\rm ej}}} = 10^7 \text{ m s}^{-1} \left(\frac{E_{\rm SN}}{10^{44}\text{ J}}\right)^{1/2} \left(\frac{M_{\rm ej}}{1M_\odot}\right)^{-1/2} \tag{4.3}$$

and the sound speed in the ISM is

$$c_s = \sqrt{\frac{{\rm d}P}{{\rm d}\rho}} \simeq 10^4\sqrt{T/1\text{ eV}}\text{ m s}^{-1} \tag{4.4}$$

assuming that the ISM is an ideal gas of neutral atomic hydrogen. Therefore the ejecta are highly supersonic, with a Mach number of order 1000.

The Sedov phase starts when the swept-up ISM becomes comparable in mass to the ejecta, i.e. when $\frac{4}{3}\pi\rho_{\rm ISM}R^3 = M_{\rm ej}$. This gives

$$\begin{aligned} R_{\rm Sedov} &= 2.1\text{ pc}\left(\frac{M_{\rm ej}}{1M_\odot}\right)^{1/3}\left(\frac{n_{\rm ISM}}{1\text{ cm}^{-3}}\right)^{-1/3} \\ t_{\rm Sedov} &= 210\text{ yr}\left(\frac{M_{\rm ej}}{1M_\odot}\right)^{5/6}\left(\frac{E_{\rm SN}}{10^{44}\text{ J}}\right)^{-1/2}\left(\frac{n_{\rm ISM}}{1\text{ cm}^{-3}}\right)^{-1/3} \end{aligned} \tag{4.5}$$

where again we have assumed that the ISM is neutral atomic hydrogen.

During the Sedov phase, radiative losses are assumed to be negligible. If we therefore assume that the total kinetic energy is conserved, $\frac{1}{2}MV^2$ = constant, and that the velocity of the post-shock gas is proportional to the shock velocity (as we deduced when deriving the shock jump conditions), we find that for the forward shock

$$R_{\rm fs}^3 V_{\rm fs}^2 = \text{constant}$$

and therefore, since $R_{\rm fs} \propto V_{\rm fs}t$,

$$\begin{aligned} R_{\rm fs} &= R_{\rm Sedov}\left(\frac{t}{t_{\rm Sedov}}\right)^{2/5}; \\ V_{\rm fs} &= \frac{{\rm d}R_{\rm fs}}{{\rm d}t} = \frac{2R_{\rm Sedov}}{5t_{\rm Sedov}}\left(\frac{t}{t_{\rm Sedov}}\right)^{-3/5}. \end{aligned} \tag{4.6}$$

The end of the Sedov phase comes when the radiative cooling time becomes comparable to the age of the SNR. This is given by[368]

$$t_{\rm tr} = 2.8\times 10^4\text{ yr}\left(\frac{E_{\rm SN}}{10^{44}\text{ J}}\right)^{4/17}\left(\frac{n_{\rm ISM}}{1\text{ cm}^{-3}}\right)^{-9/17}. \tag{4.7}$$

Thus, the remnants of the historical Type Ia supernovae, such as Tycho's (securely identified as a Type Ia by observing an "echo" of its spectrum, see [286]), are currently in either the free-expansion or the early Sedov phase, depending on the exact value of $n_{\rm ISM}$ and the amount of material ejected. The same is probably true of historical core-collapse supernovae such as Cas A, but the time estimates for the latter case are made much more difficult by the likelihood of pre-supernova mass loss from the massive star, which means that the supernova ejecta are expanding into a pre-existing stellar wind rather than undisturbed ISM.

Meanwhile, the reverse shock propagates inwards towards the centre of the explosion. Simulations indicate[371] that as it encounters dense, slowly-moving ejecta near the centre it will be reflected, causing a secondary outward-propagating shock, which may in turn generate a secondary reverse shock. The shock structure within the SNR has the potential to become very complex, even in the idealised situation of a spherically symmetric explosion—and modern 3D simulations indicate that most supernova explosions are far from spherically symmetric. Fortunately, most of the observational signatures of particle acceleration, such as X-ray synchrotron emission and TeV photons, seem to come from the limb of the SNR and hence to be associated with the relatively uncomplicated primary forward shock.

4.3.3 Observational evidence of particle acceleration

The standard catalogue of Galactic supernova remnants is maintained by David Green[372] and currently (May 2014 edition) contains 294 SNRs. Of these, 274 (93%) are reliably detected at radio wavelengths. In contrast, only ∼40% of the catalogued remnants are detected in the X-ray band, and only ∼30% in the optical. This is at least partly because core-collapse supernovae (CCSNe) involve very massive stars, which are confined to the Galactic plane and therefore often obscured by dust from our viewpoint.

Of the 294 remnants, 234 are classified as shell-type or probably shell-type, 36 as definitely or probably composite, and only 9 as filled-centre (pulsar wind nebulae). The remaining 15 are either insufficiently well observed for a type to be assigned, or do not match either morphology; some of the latter may possibly be misidentified objects (for example, G16.8–1.1, which was in previous editions of the catalogue, has been removed from the May 2014 edition because it is now believed to be an HII region rather than an SNR[372]).

Radio observations

As noted above, most supernova remnants are identified by their radio emission, which is confidently identified as synchrotron radiation based on its power-law spectrum and on the fact that it is polarised[373]. The level of polarisation is, however, less than the ∼70% that would be naïvely calculated for a synchrotron source, and is lower (∼10–15%) in young remnants than in older ones (up to 40%)[373]. This suggests that the magnetic fields in the remnants are quite disordered, so that the polarisation partially cancels. As turbulent magnetic fields are essential for diffusive shock acceleration, this is an encouraging finding. A recent study of the remnant of SN 1006 by Reynoso, Hughes and Moffett[374] (see figure 4.8) finds that polarisation is low (17%) in the regions where there is bright radio and X-ray emission suggesting efficient acceleration, and significantly higher ($60 \pm 20\%$) in the southeastern sector where there is little evidence of acceleration. By considering the direction of polarisation, they show that the efficient acceleration occurs in the region of the shock that most probably has quasi-parallel geometry, and conversely the southeastern sector appears to have quasi-perpendicular orientation. This is highly consistent with expectations from DSA.

Just over half (159) of the catalogued SNRs have a definite spectral index α, where flux $S \propto \nu^{-\alpha}$, quoted for their radio emission, and a further 64 have a spectral index of questionable validity; Green[372] points out that the data on which this information is based are of very variable quality. (The remaining 71 remnants either have spectra which are not a pure power law (14 SNRs) or have radio spectra than cannot be fitted owing to poor quality or thermal contamination.) Whether the "questionable" values are included or not, the mean is 0.48 ± 0.01; using the relation $\alpha = \frac{1}{2}(\delta - 1)$ where δ is the electron spectral index, this is a good match to the test-particle DSA prediction of

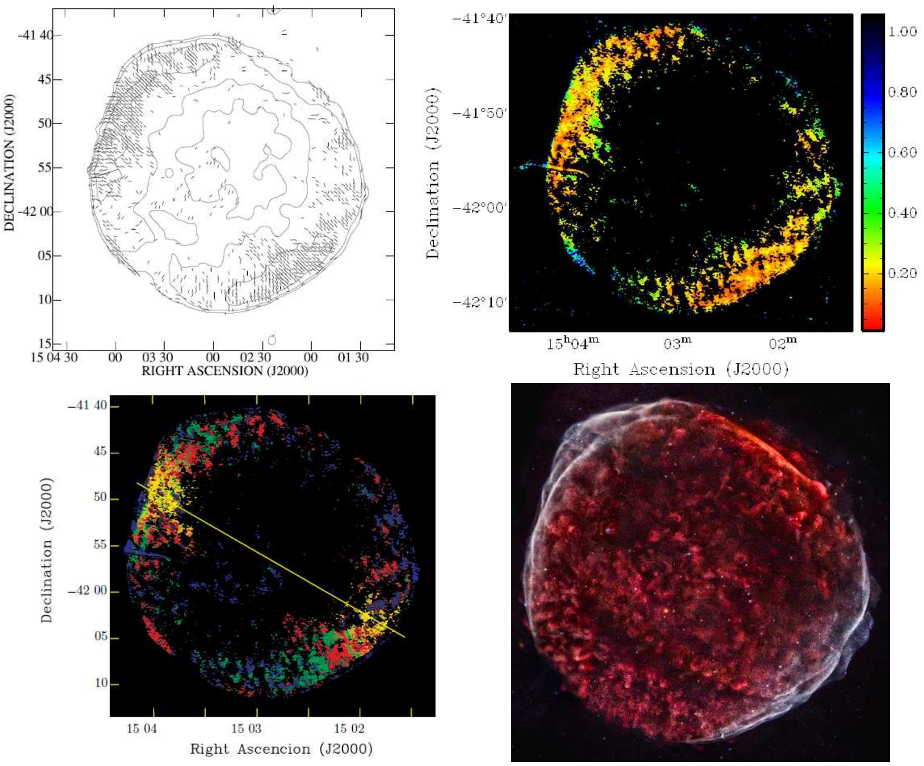


Figure 4.8: Polarisation of the radio signal in SN 1006[374]. Top row, direction (left) and intensity (right) of polarisation. Bottom row, left panel: direction of polarisation relative to Galactic plane (yellow line). Red pixels are for vectors at a fixed angle of 60° (the direction of the Galactic plane), while green indicates vectors that are locally radial. In both cases, a tolerance of ±14° is adopted with the intensity of pixels fading as they approach the limit. Pixels that do not fall in either of these two groups are plotted in blue, such that the fainter the blue, the closer to 60°. Bottom right, X-ray image from Chandra[375] for comparison: note that the bright X-ray rims correspond to the regions of low polarisation and quasi-parallel magnetic geometry.

$\delta = 2$. However, the spread in values is quite large, as shown in figure 4.9: the standard deviation is 0.14 without, and 0.15 with, the questionable values. Younger objects have higher spectral indices, typically 0.6–0.8[373, 376], as seen for Cas A in figure 4.9. Völk[378] explains this in terms of the modification of the shock front by reflected ions: the low-energy electrons responsible for the radio synchrotron emission (recall that the characteristic synchrotron frequency $\propto E^2$) respond only to the discontinuity at the subshock (see figure 3.8) and not to the shock precursor; therefore the compression ratio r is less and the spectral index $\Gamma = (r + 2)/(r - 1)$ increases (see page 157).

Pulsar wind nebulae (see below) have much flatter spectra, typically 0.0–0.3[282], and indeed the average for composite and filled morphologies in Green's catalogue is 0.36 ± 0.04, with a standard deviation of 0.17 (the numbers are essentially the same whether questionable index values are included or not). Such a flat spectrum is not consistent with DSA, but pulsar wind nebulae have other possible acceleration mechanisms available, such as magnetic reconnection.

X-ray emission

X-ray emission from SNRs is often thermal bremsstrahlung and line emission from the shock-heated plasma inside the SNR[369]. However, many SNRs also have a featureless power-law continuum in the X-ray region of the spectrum

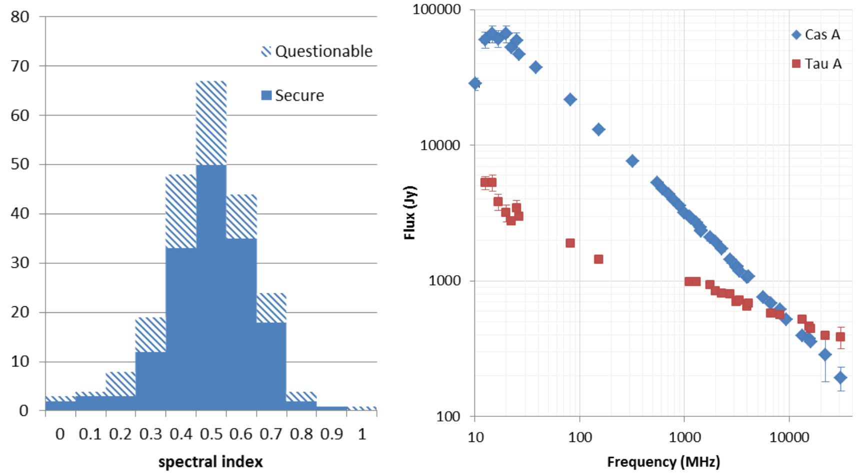


Figure 4.9: Left panel, radio spectral indices of SNRs in Green's catalogue[372]. Right panel, well-measured radio spectrum of the young shell-type SNR Cas A (remnant of a supernova from ∼1680) and the pulsar wind nebula Tau A (the Crab Nebula, remnant of SN 1054), from Baars et al.[377]. Cas A has a steep spectrum with a spectral index of 0.77; note the much flatter spectrum of the Crab. The radio flux of Cas A decreases with time[368, 377]; the points shown here have been corrected to epoch 1965.

as well as in the radio. This is identified as synchrotron radiation because, to quote Reynolds[373], "nothing else works." Bremsstrahlung would be accompanied by strong atomic line emission, and inverse Compton emission would have the same slope as the radio synchrotron emission ($\alpha \sim 0.6$), whereas the X-ray emission has a much steeper slope, ∼2.3 [373]. Because of the γ^2 factor in the characteristic frequency of synchrotron radiation (see section 2.3.5), synchrotron radiation in the X-ray region implies much higher-energy electrons than radio synchrotron radiation.

X-ray synchrotron emission is a common feature of pulsar wind nebulae, but much less common in shell-type SNRs. Shell SNRs with strong X-ray synchrotron emission are generally young objects[373, 369] such as SN 1006, Tycho (SN 1572) and Cas A (SN ∼1680): the X-ray synchrotron emission is coming from a "thin rim" at the edge of the SN and presumably delineates the forward shock of the supernova blast wave (these young SNRs are in either the free expansion or the very early Sedov phase of their evolution).

A key feature of these thin rims is that they *are* thin, which implies that either the magnetic field or the electron population—or both—changes rather rapidly downsteam of the shock. One possibility would be that synchrotron radiation losses deplete the electron population fast enough that the X-ray synchrotron radiation can only be sustained in the immediate vicinity of the shock where the electrons are accelerated. If this is the case, then we can deduce[373] a limit on the magnetic field

$$B > 200 \left(\frac{V_s}{1000 \text{ km s}^{-1}} \frac{0.01 \text{ pc}}{w} \right)^{2/3} \mu\text{G}, \tag{4.8}$$

where V_s is the speed of the shock and w is the width of the X-ray rim. This implies a very substantial amplification of the ambient magnetic field, to $\mathcal{O}(100)\,\mu$G as opposed to the typical interstellar magnetic field of 3–5 μG. On the other hand, the thin rim morphology is also observed at radio wavelengths, where synchrotron losses cannot dominate[373]. This feature is clearly seen in figure 4.10, which shows SN 1006 at 1.5 GHz[379].

Pohl, Yan and Lazarian[380] argue that the magnetic turbulence induced by DSA may be strongly damped on length scales of order 0.01 pc, so that the fall-off in synchrotron intensity is a consequence of a reduction in magnetic field rather than electron number density. This might account for the similarity of the rim morphology in both radio and X-ray synchrotron emission, as the electron lifetime is not involved in this model.

Cassam-Chenaï et al.[381], modelling Tycho's SNR (which has a very similar morphology to SN 1006), found that both the synchrotron-loss model and the magnetic-damping model would fit the X-ray data, but neither gave a good description of the radio profile—though they suggested that a combined model, with the magnetic field damped to an intermediate value of 50–100 μG, might do better.

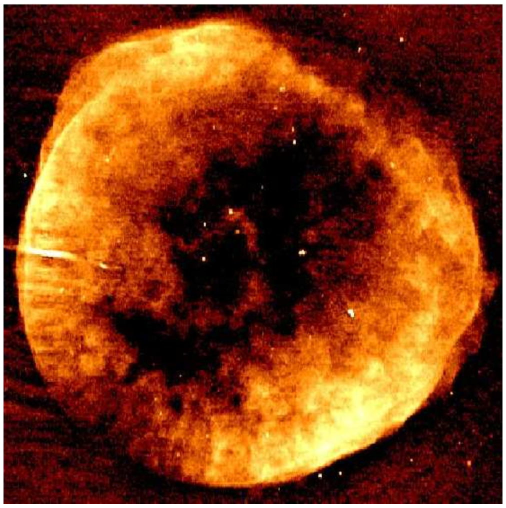


Figure 4.10: Radio image of SN 1006 at 1.5 GHz[379], using data from the Very Large Array (VLA) and the Australia Telescope Compact Array (ATCA). Compare with the Chandra image at bottom right of figure 4.8, noting the "thin rim" morphology in both wavebands. The near-horizontal streak at the left edge, seen in the radio but not the X-ray, is a background radio galaxy.

Whichever model is correct, the magnetic field in the vicinity of the shock is amplified considerably more than the factor of r expected from simple diffusive shock acceleration (though the field can be lower in the magnetic damping case than it is in the synchrotron loss case). This is important, as a higher magnetic field generally increases the maximum energy to which particles can be accelerated. If SNRs are to be the source of Galactic cosmic rays up to the "knee" in the CR spectrum, magnetic field amplification is probably required in order to speed up the acceleration process[373, 382].

In the context of magnetic field strength, a particularly interesting observation is the presence of "stripes" of high-intensity X-ray emission in Tycho's SNR[383] (see figure 4.11). Considered as a consequence of diffusive shock acceleration, the striped pattern is most naturally interpreted as a reflection of the gyroradius of accelerated protons; the spacing is much too broad to be caused by gyrating electrons. The energy of the protons is then given by[383]

$$E_{\rm CR} = 9\left(\frac{\ell_{\rm gap}}{1''}\right)\left(\frac{D}{4.0\ {\rm kpc}}\right)\left(\frac{B}{\mu{\rm G}}\right)\times 10^{12}\ {\rm eV}. \tag{4.9}$$

This yields $\sim 2\times 10^{14}$ eV if the magnetic field has a typical interstellar value of 3 μG, but $\sim 2\times 10^{15}$ eV if a value of 30 μG, consistent with the models of Cassam-Chenaï[381], is adopted. The latter value is a good match to the "knee" of the cosmic ray spectrum.

GeV and TeV photon emission

Emission at GeV and TeV energies is much more common in pulsar wind nebulae than in shell-type SNRs. TeVCat[257] lists 62 TeV γ-ray sources that might broadly be classed as supernova remnants: 37 of them are pulsar wind nebulae (which may or may not be properly described as supernova remnants, see below), 1 is described as a composite SNR, 11 are supernova remnants interacting with molecular clouds, and 13 (21%) are shell-type SNRs. This contrasts strongly with Green's catalogue, in which 80% of the 294 catalogued SNRs are shell-type or probably shell-type.

A systematic study of shell-type SNRs (i.e., excluding PWNe) by *Fermi*–LAT[384] identifies 19 SNRs seen at GeV energies, with 25 additional candidates (positive identification requires evidence of spatial extension or association with a TeV source; candidates are GeV sources spatially coincident with SNRs in Green's catalogue but not displaying extended structure). The positively identified GeV sources fall into two distinct categories: SNRs interacting with molecular clouds, which are typically brighter at GeV energies than they are as TeV sources, and young SNRs, which have harder spectra and are typically strong TeV sources.

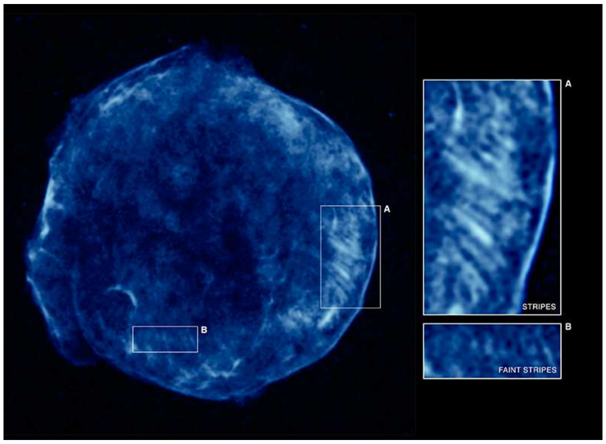


Figure 4.11: X-ray "stripes" in a deep *Chandra* image of Tycho's supernova remnant[383]. Although the stripes are seen in the interior of the SNR, Eriksen et al.[383] believe them to be features of the forward shock thin rim projected on to the main body of the remnant. The stripes have a spacing of 8.6″ on average (varying from 4.4″ to 13.3″.

GeV and TeV emission may be produced by inverse Compton scattering or π^0 decay, as discussed in section 2.4.2. The spectral energy distributions (SEDs) of SNRs vary quite significantly[385], and examples of both IC-dominated and apparently π^0 dominated spectra can be identified, as shown in figure 2.68. Both Tycho's SNR and RX J1713.7–3946 are young supernova remnants. *Fermi*–LAT has also found evidence[386] that the older, interacting SNRs W44 and IC443 (both belonging to the category of SNRs interacting with molecular clouds) also produce γ-rays through π^0 decay (see figure 4.12) rather than bremsstrahlung (inverse Compton is ruled out by the shape of the spectrum).

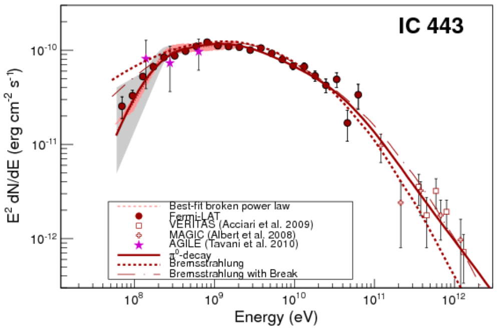


Figure 4.12: Spectral energy distribution of the SNR IC443[386] in the γ-ray region, with fits to π^0 decay (solid line), bremsstrahlung (short dashes) and bremsstrahlung with a broken power law (long dashes).

SNRs producing high-energy γ-rays through π^0 decays must necessarily be accelerating hadrons to high energies—there is no practicable way of generating enough π^0s through leptonic reactions—so these observations are, to quote [386], "direct evidence that cosmic-ray protons are accelerated in SNRs." Those SNRs whose spectral energy distributions are better fitted by leptonic models are not ruled out as cosmic-ray sources either: Ellison et al.[288], while demonstrating that the GeV–TeV emission of RX J1713.7–3946 is dominated by inverse Compton emission, explicitly state that "even though CR electrons dominate the GeV-TeV emission, the efficient production of CR ions is an essential part of our leptonic model", and explanations of the steep radio spectrum of young SNRs[378] in terms of a modified shock profile also rely on hadron acceleration (accelerating electrons alone would not provide enough back pressure to modify the shock front).

4.3.4 Pulsar wind nebulae

Most core-collapse supernovae result in the formation of a pulsar (a few will produce black holes instead). Conservation of angular momentum and trapping of magnetic field lines imply that young pulsars have very large magnetic fields and spin extremely rapidly—the Crab pulsar, for example, has a spin period of 33 ms. The combination of high magnetic field and high spin rate induces a very large electric field which effectively rips charged particles off the surface of the neutron star and accelerates them to high energies[387]. In the dense environment surrounding the pulsar, these particles initiate an electromagnetic shower, producing large numbers of e^+e^- pairs: the pulsar magnetosphere becomes filled with a *pair plasma* (dominated by e^+ and e^-, rather than ions and electrons as in a normal plasma). Such a plasma will be forced by the pulsar magnetic field to rotate with the pulsar, but this becomes impossible beyond a radius $r_L = cP/2\pi$ from the pulsar's rotation axis, where P is the pulsar period, since beyond this the plasma would have to move faster than light. On reaching r_L, the plasma escapes along open magnetic field lines, creating a magnetised relativistic wind away from the pulsar. This wind creates the pulsar wind nebula[3]. (The radius r_L defines the *light cylinder* around the pulsar.) The observational signature of pulsar wind nebulae is sufficiently clear-cut that a number of X-ray and/or TeV γ-ray sources have been classified as such despite the fact that no central pulsar has (yet) been detected[388]; of course, as pulsar pulses are beamed, it is entirely possible that the pulsar powering a nebula is not visible as such in our line of sight.

The energy radiated in the pulsar wind is powered by a loss of rotational kinetic energy: pulsars *spin down* over time (for example, the Crab pulsar period is increasing by 36 ns per day[282]). Therefore, unlike shell supernova remnants, pulsar wind nebulae (PWNe) are not powered by the energy of the supernova explosion, and so are not strictly speaking "supernova remnants" at all. However, the Crab Nebula is universally referred to as a supernova remnant, despite the surprising lack of any evidence of a surrounding shell[389].

The Crab is the archetypal PWN. It was one of the first radio sources to be identified with an optical counterpart[390] and is observed at all wavelengths from radio to TeV γ-rays (for which it has become the *de facto* standard calibration source—for example, TeVCat[257] quotes the fluxes of catalogued sources in "Crab units"). It is known beyond reasonable doubt to be associated with the supernova seen by multiple Chinese and Japanese sources in 1054[391], and its central pulsar was also one of the first to be identified, in 1968[392]. The spin-down rate of the pulsar provides a power of about 5×10^{31} W, which is enough to account for the energy radiated by the nebula[282].

Since the kinetic energy of a pulsar is $\frac{1}{2}I\omega^2$, where I is the moment of inertia of the pulsar and $\omega = 2\pi/P$ is its angular velocity, the radiated energy and the

[3] In older literature, pulsar wind nebulae around young pulsars are sometimes referred to as *plerions*, from the Greek for "full" This word has largely disappeared from modern usage.

spin-down rate are related by

$$\frac{\mathrm{d}E}{\mathrm{d}t} = 4\pi^2 I \frac{\mathrm{d}P/\mathrm{d}t}{P^3}; \tag{4.10}$$

therefore faster pulsars will generally radiate more energy and be more likely to power a pulsar wind nebula. PWNe are therefore normally associated with either young pulsars such as the Crab or "recycled" pulsars which have been spun up by accretion from a binary companion.

For young pulsars, the age of the pulsar can in principle be inferred from its spin-down rate. On the assumption that[282]

$$\dot{\omega} = \frac{\mathrm{d}\omega}{\mathrm{d}t} = -k\omega^n,$$

where k is a constant and n is the *braking index*, the age of the pulsar is given by

$$\tau = \frac{1}{n-1}\frac{P}{\mathrm{d}P/\mathrm{d}t}\left(1 - \left(\frac{P_0}{P}\right)^{n-1}\right), \tag{4.11}$$

where P_0 is the initial period of the pulsar and we assume that $n \neq 1$. Of course, we do not know P_0, but it is usually reasonable to assume that $P_0 \ll P$, and therefore the second term in the brackets is a small correction. A somewhat more significant problem is that we don't generally know n either: the need to measure the second derivative of the period means that it can be determined only for very young pulsars that have been monitored for a significant amount of time. The prediction for magnetic dipole radiation is $n = 3$, and this is used to define the *characteristic age*

$$\tau_\mathrm{c} = \frac{P}{2\dot{P}}, \tag{4.12}$$

which is the result of setting $P_0 \ll P$ and $n = 3$ in equation (4.11). The problems with this are twofold: where the true age of the pulsar is known from historical records, τ_c is often found to be a considerable overestimate (for example, the Crab comes out as 1240 years old instead of 960), probably because for young pulsars it is not reasonable to assume $P_0 \ll P$; furthermore, where n has been measured it is invariably less than 3 [393] (which would actually move the characteristic age in the wrong direction, presumably because neglecting P_0 is a larger effect). The reason why pulsars do not behave like a constant magnetic dipole ($n = 3$) is not currently well understood—there are several possibilities[393], but little evidence for any of them.

A parameter of importance in understanding pulsar wind nebulae is the ratio of magnetic energy to particle energy, which in SI units is

$$\sigma = \frac{B_1^2}{\mu_0 n_1 \gamma_1 m c^2} \tag{4.13}$$

(in cgs units, as it is usually quoted in the astronomical literature, the factor of μ_0 is replaced by 4π). Here B_1 is the upstream magnetic field, n_1 is the upstream particle density, and γ_1 is the Lorentz factor of the flow. The parameter σ is usually called the magnetisation of the wind.

Within the pulsar magnetosphere, most of the energy is carried by the magnetic field: the calculated value of the magnetisation is $\sigma > 10^4$ [282] at the light cylinder. In contrast, models of pulsar wind nebulae (in particular the

Crab, which is the best studied) at the termination shock require a particle-dominated wind, $\sigma \ll 1$ (typically a few $\times 10^{-3}$). There must be a dramatic change in the nature of the pulsar wind between these two boundaries, with the magnetic energy largely transferred to particles. This transformation is still not well understood.

4.3.5 Evolution of pulsar wind nebulae

The pulsar wind is clearly highly supersonic, and so a termination shock will be generated when it collides with the surrounding material, either the supernova ejecta or the interstellar medium. The radius r_{TS} of the termination shock is given by[395]

$$r_{\mathrm{TS}}^2 = \frac{\mathrm{d}E/\mathrm{d}t}{4\pi\eta c p}, \tag{4.14}$$

where $\mathrm{d}E/\mathrm{d}t$ is the rate at which the pulsar injects energy into the wind, η is the fraction of total solid angle over which the wind is emitted, and p is the outside pressure. This represents the location at which the outside pressure and the wind ram pressure are balanced. Chevalier[396] divides the evolution of a PWN into six distinct phases:

I. $t < 10^2$ yr: the synchrotron radiation from the PWN is absorbed by the supernova ejecta;

II. $10^2 < t < 10^3$ yr: the SNR has become transparent to the PWN synchrotron emission, and the PWN expands within the freely expanding SNR;

III. $10^3 < t < 10^4$ yr: the power ejected by the pulsar drops steeply as it ages, and the PWN expands adiabatically within the SNR;

IV. $10^4 < t < 10^5$ yr: the reverse shock from the SNR shell hits the PWN, probably asymmetrically since most pulsars are born with a significant sideways “kick” from the asymmetric SN explosion, and there is a complex interaction between the edge of the PWN and the reverse shock, possibly generating oscillations, filamentary substructure and an amplified magnetic field[282];

V. $t \sim 10^5$ yr: the pulsar passes through and interacts with the surrounding SNR shell;

VI. $t > 10^5$ yr: the pulsar leaves the SNR and interacts with the ambient interstellar medium.

The timings of the different phases are extremely approximate, and depend on the properties of the supernova, the pulsar and the ambient ISM; Gaensler and Slane[282] prefer a somewhat shorter timescale, quoting $t \sim 4 \times 10^4$ yr for phase V. Phase I has yet to be observed, as there are no pulsars young enough and close enough to offer an opportunity to do so: the most promising candidate is SN 1987A in the Large Magellanic Cloud, but so far there is no confirmed detection of a pulsar, compact central X-ray source or pulsar wind nebula in this object, though recent high-resolution microwave and sub-mm observations may hint at one[397]. Observed PWNe around young pulsars such as the Crab typically correspond to phase II or III, but there are older examples such as the Vela pulsar ($\tau_{\mathrm{c}} = 11$ kyr) which are in phase IV, and “bow-shock” PWNe interacting with interstellar gas, such as PSR B1957+20 [282].

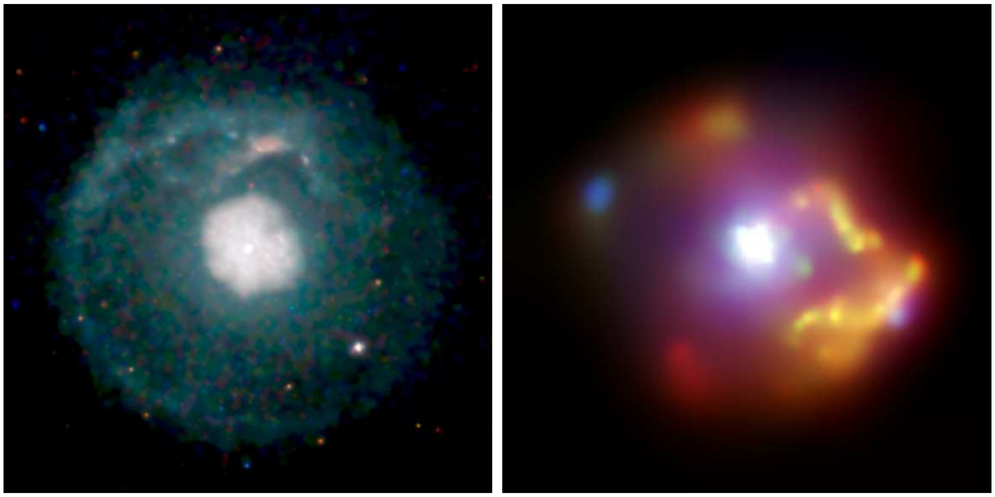


Figure 4.13: The young supernova remnants G21.5–0.9 (left) and SNR 0540–69.3 (right), as imaged in X-rays by *Chandra* [398, 399]. Both remnants show a central pulsar wind nebula surrounded by a shell SNR.

Morphologically, a typical young PWN would be a reasonably symmetrical nebula in the centre of a shell SNR. Examples of this, shown in figure 4.13, include the Galactic supernova remnant G21.5–0.9, which in X-rays consists of a bright symmetrical PWN surrounded by a faint shell, and the LMC supernova remnant SNR 0540–69.3, which has a bright shell enclosing a jet-torus PWN remarkably similar to the Crab. As mentioned earlier, the Crab itself is atypical in having no detectable shell.

In older SNRs, we expect to see the pulsar moving away from the centre of the shell and developing a bow shock. Examples of this are seen in figure 4.14: the nearby Vela supernova remnant, which is believed to be about 10000 years old, and W44, at a distance of about 3 kpc and an estimated age of 20 kyr. Finally, when the pulsar leaves its associated supernova remnant and travels through the interstellar medium, it is still moving at supersonic speed and may therefore still have a bow-shock PWN if its spin-down rate is high enough to power one. This is true of some "recycled" old pulsars which have been spun up by binary companions, such as the "Black Widow" pulsar B1957+20 [401].

Pulsar wind nebulae are not an inevitable accompaniment to pulsars: spin-down rates reduce as the pulsar ages, and most old isolated pulsars are not radiating enough energy to energise a pulsar wind nebula, though a conventional bow shock may form if they are travelling at supersonic speed through the ISM. Conversely, pulsar wind nebulae are sufficiently characteristic that a number of X-ray sources have been categorised as pulsar wind nebulae despite the absence of any detected pulsar[388]. (Note that a pulsar may not be detected as such if its magnetic axis is not close enough to our line of sight. Some radio-quiet neutron stars associated with supernova remnants have been detected purely as (non-pulsing) soft X-ray sources, known as Compact Central Objects or CCOs—this includes the central neutron stars of Cas A and Vela Jr[400].)

4.3.6 Particle acceleration in PWNe

The striped wind

Pulsars are oblique rotators: that is, the magnetic axis is not parallel to the rotation axis. As a result, the rapid pulsar rotation creates a current sheet

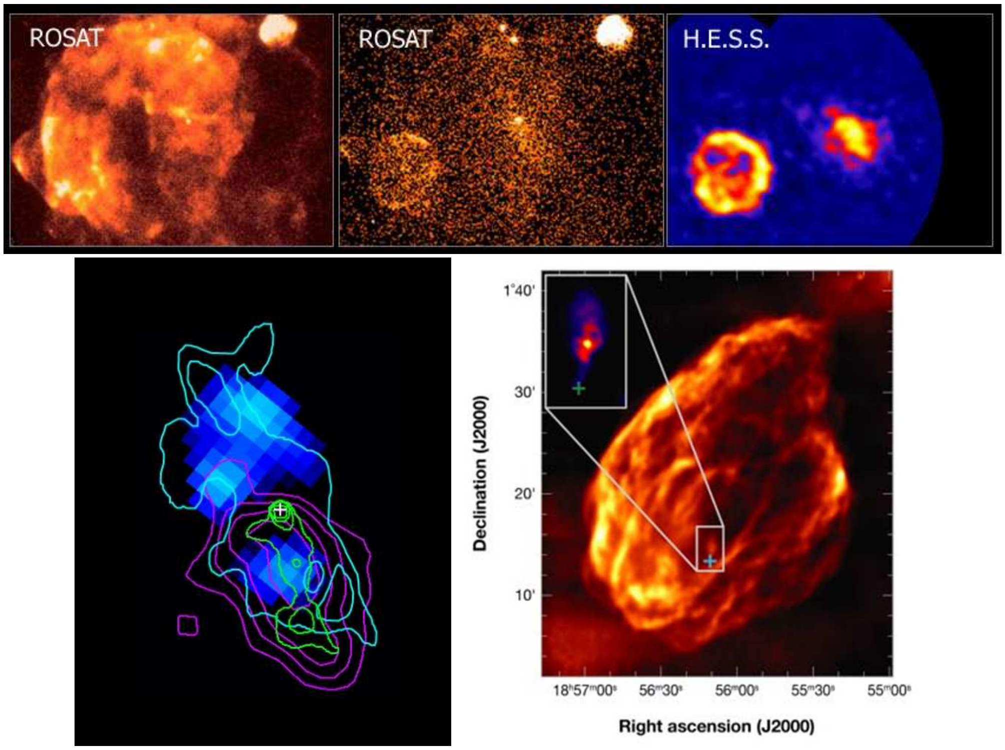


Figure 4.14: The "middle-aged" supernova remnants Vela and W44. Top panel: the Vela SNR in X-rays by ROSAT (left), ROSAT higher-energy (> 1.3 keV) X-rays (middle) and TeV γ-rays by HESS. The left panel shows the very large shell SNR: at only 250 pc, Vela is one of the closest supernova remnants to us, and covers 8° on the sky. At higher energies (middle), the shell is much fainter, and the pulsar can be seen near the centre of the image. Two other SNRs are visible: the bright blob in the top right is Puppis A, a background object about 1 kpc away, and the circle at lower left is the younger SNR nicknamed "Vela Jr", which is probably only about 1000 years old and, on the basis of X-ray absorption, somewhat more distant than Vela[400]. In TeV γ-rays (right), the Vela shell does not radiate, though the younger Vela Jr does, but the Vela pulsar wind nebula can be seen extending downwards from the pulsar. The bottom left panel shows the Vela PWN in X-rays and γ rays: the green contours are ROSAT data (0.5–2 keV), cyan the Birmingham telescope on Spacelab2 (2.5–12 keV), blue pixels INTEGRAL/IBIS (18–14 keV) and purple contours HESS (> 1 TeV). Note that the pulsar itself, which is clearly seen by ROSAT, has been subtracted from the INTEGRAL data; at TeV energies, in contrast, the pulsar is not seen. Upper images from HESS webpage[402]; lower from INTEGRAL webpage[403]. Lower right panel, SNR W44, imaged with the VLA at 1.4 GHz, with inset showing a close-up of the pulsar (marked with a +) and its small PWN. Image taken from [282].

between two different magnetic polarities that oscillates as the position of the magnetic equator changes. This phenomenon leads to two interleaved spirals of magnetic flux, of opposite polarities, known as the *striped wind* (see figure 4.15).

The importance of the striped wind is suggested by X-ray observations. Many pulsar wind nebulae (see figure 4.16) have a jet-torus morphology in X-rays, suggesting that the accelerated particles creating the X-ray synchrotron emission emanate from this equatorial belt. Estimates of the likely radius of the pulsar termination shock[282] yield values of order 0.1 pc, which is consistent with the observed radius of the Crab X-ray torus and the associated optical "wisps". The large ratio between the radii of the termination shock and the light cylinder ($r_{\rm L} = 1600$ km for the Crab, a ratio of 2×10^9) implies that the plasma of the pulsar wind expands by an enormous factor as it traverses this space, and is thus extremely cold (in a thermodynamic sense) when it reaches the termination shock. Its motion is thus very highly ordered, which makes it unlikely to produce much radiation[387]. As the motion is not only ordered, but also highly relativistic (the γ factor of the wind at the termination shock is estimated to be $\sim 10^6$ [394]), relativistic beaming further reduces the chance of observing emission from the wind zone between the light cylinder and the termination shock. As with synchrotron radiation, the effect of beaming is to reduce the visibility of emission to a very narrow viewing angle $\sim 1/\gamma$; the result of this is that any emission from the wind zone will appear as a central point source. It is in fact true that in well-resolved jet-torus PWNe such as Vela and the Crab, the region inside the X-ray ring does appear dark, with a central point source identified with the pulsar proper: as Kirk et al.[387] point out, because of beaming effects we *cannot* conclude that no radiation of energy takes place in this region. (An unfortunate consequence of this is that it is quite difficult to test the theory of the striped wind by observation.)

The magnetic geometry of the striped wind has important implications for acceleration mechanisms. In particular, it is clear from figure 4.15 that the termination shock must be quasi-perpendicular, which—as we saw when considering SN 1006—is not a favoured geometry for diffusive shock acceleration, since particles entrained on the magnetic field lines will not cross the shock.

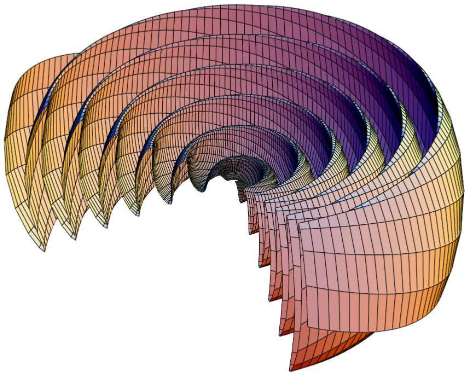


Figure 4.15: The surface traced out by a pulsar's magnetic equator, carried outwards by a radial wind[387]. The magnetic fields above and below the sheet have opposite polarities, leading to a "striped wind" of alternating polarity in the equatorial region. The high-latitude wind is not striped, but the observational evidence suggests that the X-ray emission from PWNe does not come from high latitudes.

The termination shock

As noted above, it is generally assumed that the particle acceleration needed to account for the synchrotron and TeV emission of pulsar wind nebulae takes place at or around the termination shock, as suggested by the similarity in scale between the observed torus and the calculated radius of the termination shock. However, the mechanism by which this acceleration takes place is not well understood. The spectral energy distributions of

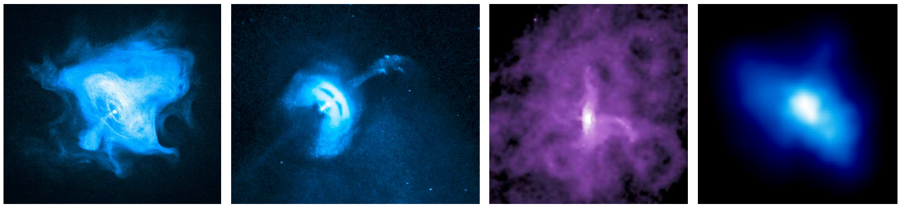


Figure 4.16: Pulsar wind nebulae with jet-torus morphology: from left to right, the Crab Nebula (SN 1054), the Vela pulsar, 3C 58 (SN 1181) and SNR 0540–69.3 in the Large Magellanic Cloud. All images from the Chandra photo album, `http://chandra.harvard.edu/photo/`.

pulsar wind nebulae are more complicated than those of shell supernova remnants, with several spectral breaks, but do not in general require high-energy hadrons: the TeV emission can be modelled as inverse Compton emission. A noteworthy feature of "bow shock" PWNe is that the X-ray synchrotron emission is usually very close to the pulsar, but the radio synchrotron emission forms an extended "tail" behind it: this is a consequence of the fact that high-energy electrons lose energy through synchrotron radiation much more quickly than lower-energy electrons do, and therefore the electron population quickly becomes depleted at the high-energy end. Hence, X-ray synchrotron radiation requires continued injection of high-energy electrons, whereas radio emission can continue for some time after the energy source has moved on.

Spectral energy distributions of PWNe are characterised by rather flat spectra (spectral index ∼0.0–0.3) in the radio, but much steeper power laws in X-rays. As discussed in section 2.3.5, steepening of synchrotron spectra at high energies is generally regarded as a consequence of the depletion of high-energy electrons mentioned above. However, the change in spectrum between radio and X-ray is rather greater than one would expect: the reduction in effective injection rate causes a steepening of the electron spectron by one power of E, and therefore, since the frequency spectral index is given by $\alpha = \frac{1}{2}(\delta - 1)$ where δ is the electron spectral index, we would expect a change $\Delta\alpha = \frac{1}{2}$, whereas the actual change is generally larger than this[282]. Other factors may also come into play: as the pulsar ages, the rate at which it injects energy into the nebula declines, and this may affect the electron spectral index, for example. Such changes would be seen first in the higher-energy synchrotron emission and subsequently propagate to lower frequencies.

Lifetime against synchrotron losses should also produce a change in the spectral index with distance from the centre of the PWN, since the effective lifetime of high-energy electrons is shorter than the travel time to the edge of large PWNe such as the Crab. This steepening is indeed observed for several PWNe[282], although the details of exactly how the spectral index changes with radius are not well modelled.

Assuming that particle acceleration does take place at or near the termination shock, as suggested by the presence of bright X-ray synchrotron emission in this region, there are several possible mechanisms for accomplishing this:

- *Diffusive shock acceleration* is disfavoured by the quasi-perpendicular magnetic geometry of the shock and by the spectral index of the radio synchrotron emission, which is much too flat for DSA. It is possible, however, that DSA may account for the high-energy tail of the electron spectrum responsible for the X-ray synchrotron emission: Kirk et al.[387] suggest that perhaps the ordered magnetic field of the striped wind might be sufficiently disrupted by the termination shock to produce small-scale turbulence around the shock front that could support DSA. The spectral index of ∼2.2 produced by acceleration at relativistic shocks would yield the correct spectral index for the Crab's X-ray emission.

- *Magnetic reconnection* is a natural candidate in view of the magnetic geometry of the striped wind, which offers the parallel flows with opposing polarity needed for reconnection events. If we assume that the magnetic field continues to carry most of the energy (i.e. $\sigma \gg 1$) right up to the termination shock, the sudden compression at the shock could cause massive reconnection and transfer this energy to the particles in the wind[387, 404, 332]. 3D particle-in-cell simulations indicate that magnetic reconnection at relativistic shocks can produce a non-thermal power-law distribution of accelerated particles[332], generally with a harder (i.e. flatter) spectrum than diffusive shock acceleration. The radio synchrotron radiation, with typical spectral index 0–0.3, requires an electron energy spectral index of 1–1.6, which is not at all consistent with relativistic DSA, but which might be achievable by magnetic reconnection if the magnetisation parameter is fairly large ($\gtrsim 50$).

Even if magnetic reconnection is not the dominant acceleration mechanism, it may still be responsible for the fast γ-ray flares produced by the Crab Nebula[331, 332]: sudden large-scale reconnection events could in principle result in large transient increases in the number of high-energy electrons, generating a sudden increase in the photon flux.

- *Resonant cyclotron absorption*[285, 387] is an acceleration mechanism that relies on a thermodynamically cold, ion-loaded plasma wind. Because of their highly ordered motion, the ions gyrate *collectively* in the magnetic field, emitting strong cyclotron waves which are then resonantly absorbed by the e^+e^- pairs. This can produce a suitably flat spectrum ($\delta < 2$) if the ions dominate the energy of the wind ($U_\mathrm{i}/U_\mathrm{tot} \sim 80\%$). We did not consider this mechanism in chapter 3, because it is not a candidate for cosmic-ray acceleration in general—it only works for e^+e^- (in fact, energy is transferred *from* the ions *to* the electrons). As an explanation for acceleration in PWNe, where there is no evidence for high-energy hadrons, it has some nice features: it can account for the inferred maximum energy and spectrum of the accelerated e^+e^-, and is highly efficient provided that the ions are sufficiently dominant. The main drawback is that the required density of ions in the pulsar magnetosphere is much higher than the maximum expected theoretically[387]; on the other hand, an ion-loaded wind is certainly not excluded by observation. This is a case where neutrino telescopes could usefully contribute: observation of neutrinos from the Crab or another nearby PWN would confirm the presence of protons.

On the whole, the best supported of these mechanisms appears to be magnetic reconnection, although resonant cyclotron acceleration is an attractive option if the required level of ion loading can be achieved.

4.4 Extragalactic sources

The gyroradius of a proton in a magnetic field B is given by

$$r_g\ (\mathrm{pc}) \sim 10^{-7}\frac{E\ (\mathrm{GeV})}{B\ (\mathrm{nT})},$$

assuming $v \simeq c$. If we assume that protons with gyroradii of order 1 kpc will random walk out of the Galaxy in a time short compared to the Hubble time, then since the Galactic magnetic field is of order 0.1 nT we conclude that cosmic rays with energies exceeding 10^9 GeV are likely to originate from outside the Galaxy. This coincides approximately with the "ankle" in the cosmic ray spectrum, see figure 2.13; the change in power law index observed at the ankle supports the idea that there is some change in the nature of cosmic ray soruces at this energy.

Observations of high-energy photons (see section 2.4) offer two clear candidates for extragalactic astrophysical accelerators: radio-loud active galactic nuclei, particularly blazars, and gamma-ray bursts. Nearby blazars are by far the most numerous extragalactic sources of TeV γ-rays (see section 2.4.6), which implies that they must accelerate electrons to at least these energies; GRBs have not been seen to emit γ-rays above a few tens of GeV, but appear to produce synchrotron radiation at X-ray energies, again diagnostic of very high energy electrons. There is little direct evidence for the acceleration of hadrons, but GRBs must contain relativistic collisionless shocks and are therefore very likely to accelerate hadrons, and most models of particle acceleration in AGN jets also presuppose hadron acceleration.

4.4.1 Gamma-ray bursts

As summarised in sections 2.4.4 and 2.6.3, GRBs are extremely intense transient sources of soft ($\sim$1 MeV) γ-rays, typically located at high redshift. The γ-ray emission lasts from a fraction of a second to a few minutes, but is often followed by an "afterglow" ranging from X-rays to radio and lasting for a few days. The discovery of the first GRB afterglows in 1997 revolutionised the study of GRBs, because the X-ray, optical and radio afterglow emission can be localised with much greater precision than the prompt γ-ray burst itself. This enabled observers to locate GRBs within external galaxies[405], establishing that they lie at cosmological distances and allowing the determination of redshifts, association (or lack thereof) with star formation and other essential information. Initially, all afterglow observations referred to long GRBs: short GRBs have much fainter afterglows, so that the first short GRB afterglow was not detected until 2005, following the launch of the purpose-built *Swift* satellite[191]. It is believed[406] that afterglows arise from the interaction of the GRB blast wave with the surrounding medium: the faintness of short GRB afterglows compared with those of long GRBs is explained by the much lower density of the circumburst medium in the latter case.

GRB afterglows

Somewhat ironically, GRB afterglows are now better understood theoretically than the prompt γ-ray emission of the burst itself. The afterglow is caused by the GRB blast wave—a relativistic forward shock propagating into the surrounding interstellar medium. As we saw in section 3.6, if the shock propagates with Lorentz factor Γ_s into a stationary interstellar medium, the bulk Lorentz factor of the shocked plasma will be $\Gamma = \Gamma_s/\sqrt{2}$. We expect particle acceleration to take place at the shock front, with associated synchrotron radiation assuming that a suitable magnetic field is present.

As discussed in section 3.6 and the review article by Kumar and Zhang[406], the energy of a proton of the unshocked gas in the rest frame of the shocked gas (the upstream rest frame or URF) is $\Gamma m_p c^2$ (the thermal energy of the unshocked gas is negligible, so we are just seeing the effect of the Lorentz boost). The effect of crossing the shock is essentially to randomise the directions of the particles without changing their energy (as measured in the URF). As viewed in the lab frame (equivalent to the rest frame of the unshocked gas), the average energy of protons in the shocked gas is $\Gamma^2 m_p c^2$, consistent with our finding in section 3.6 that the first return shock crossing increases the particle energy by about a factor of Γ_s^2 (which is $2\Gamma^2$, but here we are only considering a one-way crossing).

If we assume for the moment that the blastwave is isotropic (see below for evidence that it is not), and that the number density of the interstellar medium at a distance R from the burst is given by $n(R) = n_0 R^{-k}$ where k is a constant, then the total energy in the shocked plasma is given by

$$E_{\text{iso}} = \Gamma^2 m_p c^2 \int_0^R 4\pi r^2 n_0 r^{-k} \mathrm{d}r = 4\pi n_0 \Gamma^2 m_p c^2 \frac{R^{3-k}}{3-k}. \tag{4.15}$$

If E is constant (adiabatic expansion), it follows that

$$\Gamma \propto R^{(k-3)/2}. \tag{4.16}$$

The reason for introducing an R-dependent density is that long GRBs are convincingly associated with Type Ibc supernovae, which are interpreted as the core collapse of a massive star that has lost its hydrogen envelope (and, in the case of Type Ic, the helium layer as well), either by mass transfer to a close binary companion or by a stellar wind. If the circumburst medium is not dominated by the ambient interstellar medium but by a stellar wind from the progenitor star, then we would expect its density to decrease with distance from the burst. In the simplest case of constant mass loss rate $\dot{M}$ and constant wind velocity v_w, the number density at distance R from the star is given by $\dot{M} = 4\pi R^2 n(R) m_p v_w$, which for constant $\dot{M}$ and v_w implies $n(R) \propto R^{-2}$, i.e. $k = 2$ in the above equations.

If the emission is collimated into two back-to-back jets of half-angle $\theta_J \ll 1$, then the solid angle covered is given by $2\pi\theta_J^2$ instead of 4π, and the energy is therefore $E = \frac{1}{2}\theta_J^2 E_{\text{iso}}$. This does not, at least initially, change the scaling relations, as E and E_{iso} are directly proportional.

For a relativistic blastwave, relativistic aberration means that we will only see emission from particles moving more or less directly towards us. The observed time taken for the shock front to expand by an amount ΔR is therefore

$$\Delta t_{\text{obs}} = \frac{\Delta R}{v} - \frac{\Delta R}{c} = \frac{\Delta R}{c}\left(\frac{1}{\beta} - c\right) \simeq \frac{\Delta R}{2c\Gamma_s^2}, \tag{4.17}$$

using the fact that $1 - \beta \simeq 1/2\Gamma_s^2$ for $1 - \beta \ll 1$.

To obtain the time t_{obs} taken for the blastwave to expand from radius 0 to radius R, we should in principle integrate this (since Γ_s is not constant): this will introduce a numerical factor but will not change the functional dependence, which is

$$t_{\text{obs}} \propto R^{4-k} \propto \Gamma^{\frac{8-2k}{k-3}}. \tag{4.18}$$

For a constant-density circumburst medium, we expect $t_{\text{obs}} \propto R^4 \propto \Gamma^{-8/3}$; for a wind-dominated medium with $k = 2$, we get $t_{\text{obs}} \propto R^2 \propto \Gamma^{-4}$.

This does not exhaust the possibilities: as discussed by Kumar and Zhang[406], for some types of central engine the blastwave energy may increase with time, if energy is injected into it continuously over some finite time instead of in a single explosive impulse. For the case in which the luminosity of the central engine is given by

$$L(t) = L_0 \left(\frac{t_{\text{obs}}}{t_0}\right)^{-q}$$

with $q < 1$ ($q > 1$, luminosity falling off rapidly with time, is essentially equivalent to the adiabatic case), an argument similar to the above gives

$$t_{\text{obs}} \propto R^{\frac{4-k}{2-q}} \propto \Gamma^{-\frac{8-2k}{2+q-k}},$$

where as before $k = 0$ for a constant-density circumburst medium and 2 for a wind-dominated medium.

The afterglow spectrum

The afterglow is generally assumed to be generated by synchrotron radiation at the external shock. As we saw in section 2.3.5, synchrotron radiation from an electron of energy E is emitted at frequencies close to $\nu_{\rm syn} = \frac{3}{2}\gamma^2\nu_g$, where $\gamma = E/m_e c^2$ is the electron Lorentz factor and $\nu_g = eB/2\pi m_e$ is the cyclotron frequency, and a power-law electron spectrum, $N(E) \propto E^{-\delta}$, leads to a power-law synchrotron spectrum, $f_\nu \propto \nu^{-(\delta-1)/2}$.

The synchrotron energy loss from an electron of Lorentz factor γ is

$$\left|\frac{\mathrm{d}E}{\mathrm{d}t}\right| = 2\sigma_{\rm T} U_{\rm mag}\beta^2\gamma^2,$$

see equation (2.39), where the magnetic field energy density $U_{\rm mag} = B^2/2\mu_0$. If such an electron emits synchrotron radiation for some time t_0 (corresponding, in the case of GRBs, to the time since the burst), then it will lose a significant fraction of its initial energy if

$$2\sigma_{\rm T} U_{\rm mag}\beta^2\gamma^2 \geq \frac{\gamma m_e c^2}{t_0},$$

i.e.

$$\gamma \geq \gamma_c = \frac{\mu_0 m_e c}{\sigma_{\rm T} B^2 t_0}. \tag{4.19}$$

The Lorentz factor γ_c corresponds to a synchrotron frequency

$$\nu_c = \gamma_c^2 \frac{3eB}{4\pi m_e} = \frac{3e\mu_0^2 c^2}{4\pi\sigma_{\rm T}^2 B^3 t_0^2}. \tag{4.20}$$

This is known as the **synchrotron cooling frequency**. Above this frequency, the synchrotron radiation spectrum is steeper by one factor of E as a result of the progressive loss of the high-energy tail of the electron spectrum.

In addition to this modification at *high* frequency, the *low* frequency end of the synchrotron spectrum is modified by synchrotron self-absorption (see page 87), where the medium is opaque to its own radiation. As a consequence, there are three critical frequencies governing the spectrum of the GRB afterglow[406]:

- ν_a, the upper threshold for self-absorption (the radiation is self-absorbed for $\nu < \nu_a$);
- ν_m, the minimum Lorentz factor for the accelerated electrons (determined by the shock Lorentz factor Γ_s);
- ν_c, the synchrotron cooling frequency.

This leads to a broken power law with three slope breaks. The exact form of the power law depends on whether $\nu_m < \nu_c$ ("slow cooling") or vice versa ("fast cooling"): for slow cooling[406]

$$f_\nu = \begin{cases} f_0\left(\frac{\nu_a}{\nu_m}\right)^{1/3}\left(\frac{\nu}{\nu_a}\right)^2 & \nu < \nu_a; \\ f_0\left(\frac{\nu}{\nu_m}\right)^{1/3} & \nu_a \leq \nu < \nu_m; \\ f_0\left(\frac{\nu}{\nu_m}\right)^{-(\delta-1)/2} & \nu_m \leq \nu < \nu_c; \\ f_0\left(\frac{\nu_c}{\nu_m}\right)^{-(\delta-1)/2}\left(\frac{\nu}{\nu_c}\right)^{-\delta/2} & \nu \geq \nu_c; \end{cases} \tag{4.21}$$

while for fast cooling

$$f_\nu = \begin{cases} f_0 \left(\frac{\nu_a}{\nu_m}\right)^{1/3} \left(\frac{\nu}{\nu_a}\right)^2 & \nu < \nu_a; \\ f_0 \left(\frac{\nu}{\nu_m}\right)^{1/3} & \nu_a \le \nu < \nu_c; \\ f_0 \left(\frac{\nu}{\nu_m}\right)^{-1/2} & \nu_c \le \nu < \nu_m; \\ f_0 \left(\frac{\nu_m}{\nu_c}\right)^{-1/2} \left(\frac{\nu}{\nu_c}\right)^{-\delta/2} & \nu \ge \nu_m; \end{cases} \tag{4.22}$$

where f_0 is the maximum flux density ($f_\nu(\nu_m)$ for slow cooling and $f_\nu(\nu_c)$ for fast cooling).

The power $\frac{1}{3}$ segments in the above are caused by the fact that the spectrum of synchrotron radiation from a single electron below the peak frequency has slope $\frac{1}{3}$—see the top left panel of figure 2.38 on page 84. In the fast cooling scenario, the "usual" synchrotron spectral index $(\delta-1)/2$ is never seen, because even the lowest-energy electrons from the original burst of acceleration have undergone significant energy loss.

The self-absorbed spectrum is $\propto \nu^2$ instead of $\propto \nu^{5/2}$ as on page 87 because the power spectrum for accelerated electrons by definition does not extend below ν_m, and $\nu_a < \nu_m$ for months after the burst[406] because of the low density of the circumburst medium. At ν_a the electron energy distribution is probably roughly thermal, which would result in a ν^2 (Rayleigh-Jeans) spectrum in the self-absorbed region.

The "jet break": evidence for collimated emission

We saw on page 102 that GRB emission must be relativistically beamed, because the energy requirements imposed by the observed luminosity are such that if the radiation were not beamed, the source would absorb its own γ-rays through e^+e^- pair production. Relativistic aberration avoids this problem, because photons emitted from a relativistic source are confined to a cone with half-angle $1/\Gamma$ (where Γ is the bulk Lorentz factor); as the invariant mass of a two-photon system is given by $2E_{\gamma_1}E_{\gamma_2}(1-\cos\theta)$, where θ is the angle between the two photons, forcing $\theta \le 1/\Gamma$ increases the threshold for e^+e^- pair production, allowing the γ-ray photons to escape.

In principle, a relativistic outflow could still be spherically symmetric, although other relativistic outflows known in astrophysics are collimated jets. However, there is direct observational evidence that GRB outflows are jet-like, in the shape of "jet breaks" observed in the afterglow spectra of some GRBs.

Jet breaks, i.e. sudden changes in the power law index of the GRB spectrum, occur at the point where the half-angle of the aberration cone equals the half-angle of the jet, as shown in figure 4.17. When the aberration cone is narrower than the jet, the angle over which the fast particles responsible for the emission can contribute to the observed flux is limited by $1/\Gamma$, regardless of whether the particles are emitted isotropically or collimated into a jet. However, when the opening angle of the aberration cone is greater than the jet opening angle, $\theta_R > \theta_J$, this is no longer the case, and the observed flux is reduced compared to the case of isotropic outflow by a factor of $(\theta_J/\theta_R)^2 = \Gamma^2\theta_J^2$. In addition to this edge effect, the time at which $\theta_J \simeq 1/\Gamma$ also happens to be the time at which the jet starts to expand sideways, because sound waves have had enough time to travel transversely across the jet[406].

In the simplest case of adiabatic expansion into a medium of constant density, the energy in the two back-to-back jets is given by

$$E = \frac{2\pi}{3}\theta_J^2 R^3 \Gamma^2 n m_p c^2,$$

essentially as in equation (4.15) except that the solid angle is $2\pi\theta_J^2$ instead of 4π. This leads to $\Gamma \propto R^{-3/2}$ as discussed above. After the jet break, θ_J is replaced by $\theta_R = 1/\Gamma$, leading to

$$E = \frac{2\pi}{3} R^3 n m_p c^2.$$

For the adiabatic condition this would imply that the shock radius remains constant after this point, since if E is constant, R must also be constant[407]. More precise calculations (see references in [406] and [407]) indicate that R continues to increase slowly and that Γ declines exponentially with R.

The scaling relation given in equation (4.18) predicts that the time of the jet break is related to the jet half-angle by

$$\theta_J \propto t_{\rm obs}^{3/8};$$

a more precise calculation yields[223]

$$\theta_J = 0.13 \left(\frac{t_J}{1+z}\right)^{3/8} \left(\frac{n}{E_{\rm iso}}\right)^{1/8}, \qquad (4.23)$$

where t_J is the time of the jet break in days, n is the number density of the circumburst medium in cm^{-3}, $E_{\rm iso}$ is the isotropic energy (i.e. the energy calculated assuming spherical expansion) in units of 10^{52} ergs (10^{45} J), z is the redshift of the GRB, and θ_J is measured in radians. The factor of $(1 + z)$ simply reflects cosmological time dilation: our expressions up to this point have assumed an observer somewhat dangerously located just outside the GRB, whereas in fact most GRBs are at high or very high redshift.

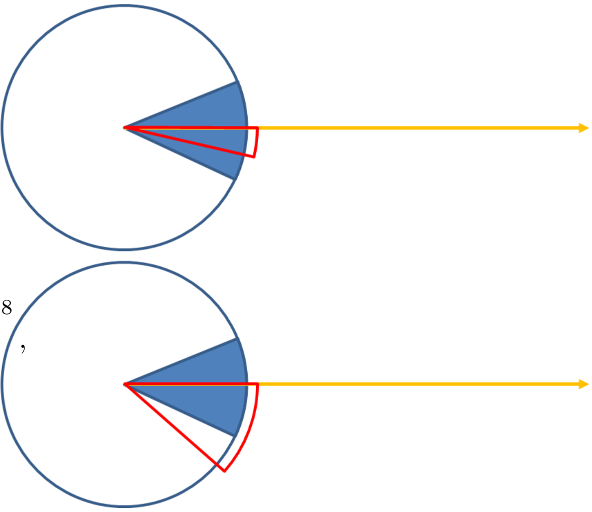


Figure 4.17: Schematic diagram of the mechanism of jet breaks in GRB spectra. In the upper panel, higher-energy photons are emitted by higher-energy particles in a narrow cone. The angular range over which particles can contribute to the observed flux is determined by the opening half-angle of the relativistic beaming, $\theta_R = 1/\Gamma$ (red cone) and does not depend on whether the emission is isotropic or collimated (blue cone). In contrast, the lower-energy photons can be emitted by lower-energy particles whose emission is less strongly beamed. As shown in the lower panel, this implies that the angular range of contributing particles is set by the jet opening half-angle, θ_J, for a collimated outflow, hence reducing the observed flux compared to isotropic emission.

Equation (4.17) predicts that if R is approximately constant, $\Gamma \propto t_{\rm obs}^{-1/2}$. Putting this into the synchrotron frequency expressions predicts a post-break lightcurve form[406]

$$f_\nu \propto \begin{cases} \nu^{1/3} t_{\rm obs}^{-1/3} & \nu_a < \nu < \nu_m; \\ \nu^{-(\delta-1)/2} t_{\rm obs}^{-\delta} & \nu_m < \nu < \nu_c; \\ \nu^{-\delta/2} t_{\rm obs}^{-\delta} & \nu > \nu_c \end{cases} \qquad (4.24)$$

for the case of slow cooling. Note that the time dependence of the lightcurve is achromatic (not dependent on frequency) above ν_m (above ν_c in the case of fast cooling); this behaviour can be used to confirm that we are seeing a jet break and not some other feature.

Jet breaks have been widely observed in long GRBs [223], with inferred jet half-angles of a few degrees (see figure 4.18). This corresponds to a break time t_J of order a day or two, which is very challenging for short bursts: short GRB afterglows are much fainter than those of long GRBs, and typically fade beyond visibility after about a day. If no jet break has been seen on this timescale, the result is a not very helpful lower limit of $\theta_J \geq 3°$ [223]. As can be seen in figure 4.18, there are nevertheless a few observations of jet breaks in short GRBs, albeit often on the evidence of a single waveband (thus with no confirmation that the post-break behaviour is achromatic), and a few *useful* lower limits, such as $\theta_J > 20°$ for GRB 050724, which exhibited no jet break in its X-ray lightcurve for 22 days after the burst[223].

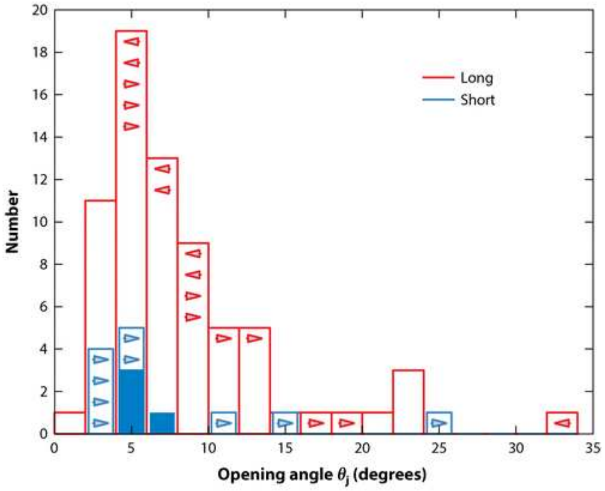


Figure 4.18: Jet opening angles inferred from afterglow spectral breaks for long (red) and short (blue) GRBs. Right (left) pointing arrowheads indicate lower (upper) limits. Figure from [223].

On the evidence of figure 4.18, it seems likely that most GRBs come from collimated outflows, although the evidence for short GRBs is currently weak. This is a significant finding for two reasons: it reduces the inferred energy by around two orders of magnitude, and it increases the inferred rate of GRB events by a similar factor (since most of them are not pointed at us). Because the emission *after* the jet break is nearly isotropic, it also implies the existence of so-called "orphan afterglows": if we are at some angle $\theta > \theta_J$ to the axis of the GRB, we will not see the burst itself or the initial stages of the afterglow, but we *should* be able to detect the afterglow as soon as $\Gamma < 1/\theta$. No such orphan afterglows have been convincingly detected to date, not entirely surprisingly: by definition you are looking for the faint late-time tail of the afterglow lightcurve, without the benefit of a GRB to show you where to point your telescope. Future survey instruments, particularly the Large Synoptic Survey Telescope (LSST)[408], are probably the best hope for detecting orphan afterglows: faint optical transients are one of the principal science goals of the LSST.

The burst itself: prompt γ-ray emission

The diagnostic feature of a GRB is the short, intense burst of soft (∼MeV) γ-rays. As shown in figure 2.49, the distribution of burst durations from the BATSE catalogue is clearly bimodal, with the division occurring at $t_{90} \simeq 2$ s; t_{90} is the time interval containing 90% of the observed flux. Apart from this, however, the lightcurves of GRBs are astonishingly diverse, as shown in figure 4.19: some have the fast-rise/slow-decline pattern common in astrophysical outbursts (e.g. supernovae, classical novae, flare stars), but many show multiple peaks, some apparently random, others with quasiperiodic features. This diversity must in some way relate to the way in which the GRB is generated, but so far no satisfactory explanation has been devised. Indeed, although there is general agreement on the likely progenitors of GRBs, the actual process of generating the γ-ray burst itself remains poorly understood.[406]

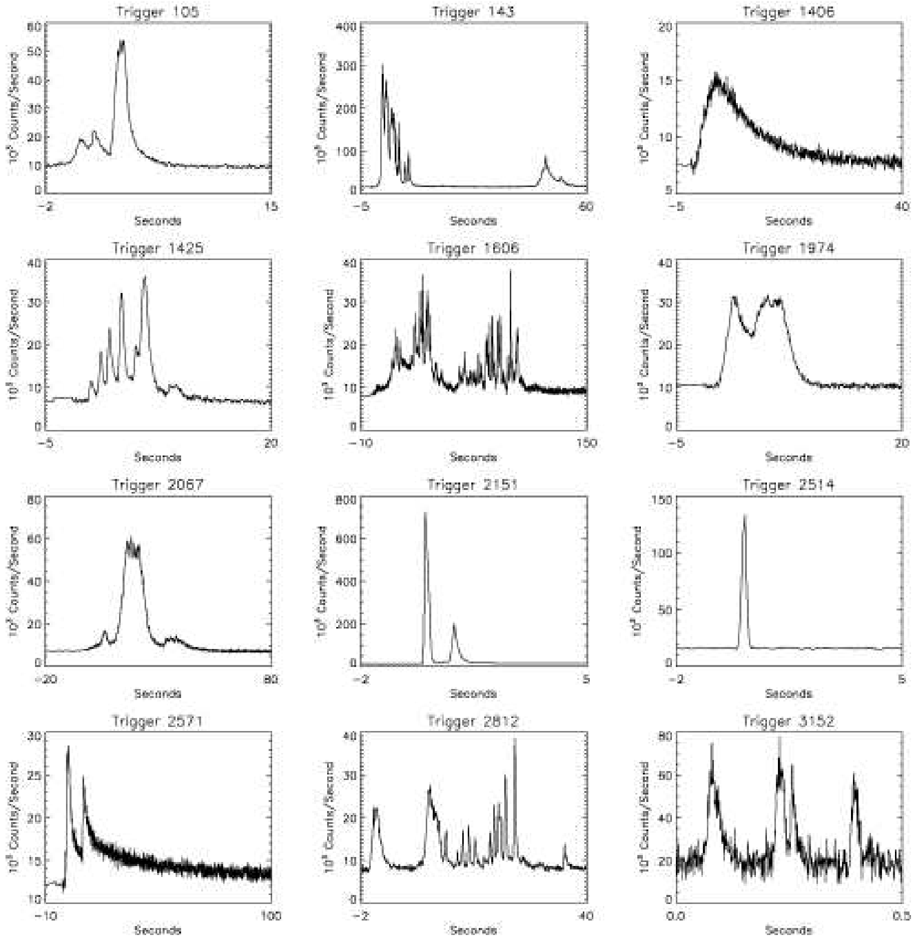


Figure 4.19: A sample of GRB lightcurves[409]. Note the different axis scales: for example, the superficially similar Trigger 1606 and Trigger 3152 differ in duration by two orders of magnitude.

Long and short GRBs

Despite the diversity in the detailed lightcurves, the basic division into two classes is robust: long GRBs have softer γ-ray spectra on average, are brighter and have brighter afterglows, occur at larger redshifts, are found in galaxies with younger stellar populations, and have a clear association with luminous Type Ic supernovae[223]. However, the conventional division at 2 s duration in the lab frame is somewhat arbitrary. First, because GRBs occur at high redshifts (extending to $z > 8$, e.g. GRB 090423 at $z = 8.26^{+0.07}_{-0.08}$[411]), lab-frame durations > 2 s ("long") may correspond to rest-frame durations < 2 s ("short"); in fact, GRB 090423 is an example of exactly this, as its lab-frame duration of 10.3 s[411] corresponds to only 1.1 s in its rest frame. Secondly, there exists a class of short GRBs with extended emission, accounting for 15–

25% of all short GRBs[223]: these have a short initial spike of γ-ray emission followed by a longer, softer "tail" extending for 10–100 s. Depending on the energy range and sensitivity of the detector and the distance of the GRB, some of these may be identified as "long" GRBs if the tail accounts for more than 10% of the observed emission. Third, the 2D plot of "hardness ratio" against t_{90} (figure 4.21) shows that the two populations are not completely disjoint: even taking the spectral difference into account, some GRBs that belong to the "short" population will be classified as "long", and vice versa.

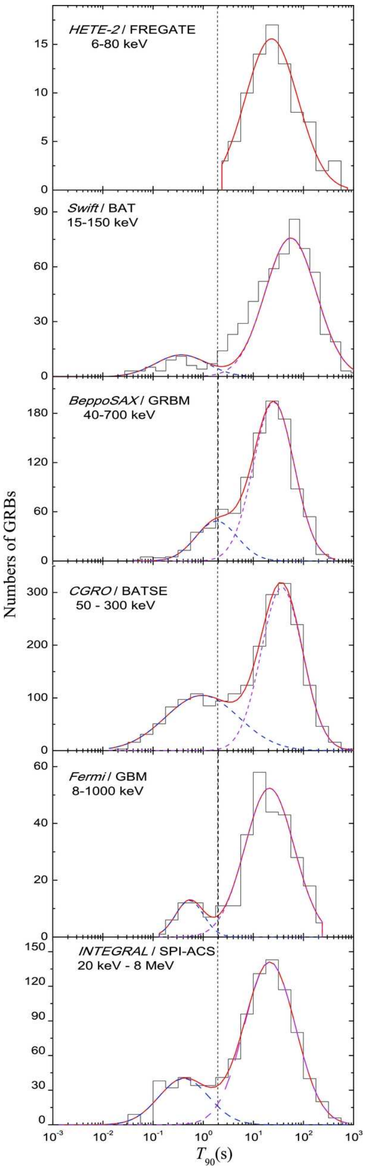


Figure 4.20: A compilation of t_{90} distributions from different detectors[410]. Although the bimodal structure is visible in most of the distributions, both its prominence and the position of the long-short divide vary.

There have been several attempts to introduce different classification schemes for GRBs, with the intention of replacing the 2 s boundary with someting more physically meaningful. Some of these attempt to classify bursts by progenitor type (e.g. compact object merger vs massive star core collapse), while others introduce different empirical parameters: Lü et al.[412], for example, use $\varepsilon = E_{\mathrm{iso},52}/E_{\mathrm{p},z,2}^{5/3}$ where $E_{\mathrm{iso},52}$ is the isotropic-equivalent γ-ray energy in units of 10^{52} ergs and $E_{\mathrm{p},z,2}$ is the rest-frame peak γ-ray energy in units of 10^2 keV. It is quite possible that some such refinement of the classification will be adopted in the future, but so far none has gained wide acceptance: in particular, as noted by Berger[223], those which rely on identifying the putative progenitors risk biasing samples in favour of still-unproven hypotheses (for example, any classification that automatically assumes that any GRB taking place in an elliptical host galaxy is a compact-object merger cannot then argue that the locations of such objects are evidence for their being compact-object mergers!).

The fireball model

One of the earliest models for GRBs, and still the most widely accepted, is the "hot fireball" model in which a large amount of energy ($10^{51} - 10^{54}$ ergs; $10^{44} - 10^{47}$ J) is released in a short time ($\sim$10 s) in a very compact region ($R \sim 10$ km), as might be expected in the formation of a neutron star or black hole. This model was originally suggested by Cavallo and Rees (1978), and in closer to its current form by Paczýnski (1986) and Goodman (1986)[413]. This energy is assumed to be initially in the form of an e^+e^- pair plasma consisting of $e^\pm$, photons and neutrinos in thermal equilibrium, but is optically thick to Thomson scattering: the result is that this initial thermal energy is converted into bulk kinetic energy of the protons in the outflow, accelerating them to

Lorentz factors of several hundred. It is assumed that this bulk kinetic energy is reconverted into photons at some distance from the initial fireball, with the aid of internal or external shocks: the most popular scenario is that the prompt γ-rays are produced by internal shocks, with the external shock responsible for the afterglow emission as discussed above.

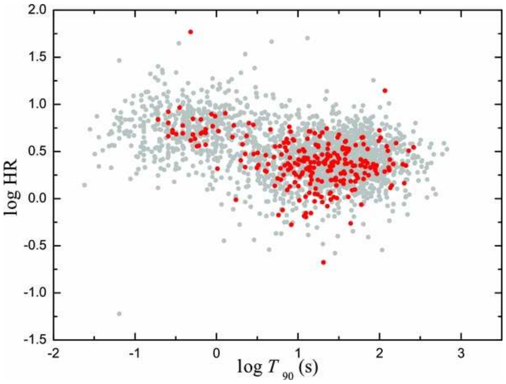


Figure 4.21: Plot of hardness ratio (defined as the fluence in the 100–350 keV band to that in the 50–100 keV band) against t_{90}, for BATSE (grey) and *Fermi*–GBM (red) GRBs, showing the harder spectrum of short GRBs. Figure from [410].

The basic energetics of the fireball model assume an adiabatic expansion of the fireball[414]. As the pressure is dominated by radiation, the ratio of specific heats is $\hat{\gamma} = \frac{4}{3}$. Given the adiabatic condition $pV^{\hat{\gamma}}$ = constant and the ideal gas law $pV = nkT$, we conclude that $TV^{\hat{\gamma}-1}$ = constant and therefore that $T \propto V^{-1/3}$. Since for an ultrarelativistic gas $T \propto \gamma$, where γ is the Lorentz factor corresponding to the random thermal motion of the gas, and $V \propto R^3$ where R is the fireball radius, we conclude that $\gamma \propto R^{-1}$: the temperature of the gas decreases as the fireball expands. However, the total energy must remain constant, so this decrease in *thermal* kinetic energy is compensated by an increase in kinetic energy of *expansion*, $\gamma\Gamma$ = constant where Γ is the Lorentz factor for the bulk motion of the expanding gas. The effect of the expansion is thus to convert thermal energy into bulk kinetic energy. This conversion continues until $\gamma \simeq 1$ or the medium becomes transparent to Thomson scattering, whichever happens first. If the medium remains optically thick, the final value of Γ is equal to the initial thermal Lorentz factor γ_0, and nearly all of the thermal energy of the burst has been converted to kinetic energy of protons. If, on the other hand, the medium becomes optically thin to Thomson scattering before this point, only part of the initial thermal energy is converted to bulk kinetic energy, the rest remaining mostly in the thermal photons (since the $e^\pm$ will annihilate to photons once the temperature drops below ~1 MeV).

The thermal distribution of the $e^\pm$ in the initial pair plasma, measured in the comoving (primed) frame, is given by[406]

$$n'_\pm = \frac{2(2\pi m_e kT')^{3/2}}{h^3} \exp\left(-\frac{m_e c^2}{kT'}\right), \tag{4.25}$$

where T' is the temperature in the comoving frame. The cross-section for pair annihilation is

$$\sigma_{e^+e^-\to\gamma\gamma} = \frac{\sigma_\mathrm{T}}{\langle\beta\rangle}$$

where $\langle\beta\rangle$ is the mean $e^\pm$ speed in units of c. For highly relativistic $e^\pm$ for which $\beta \sim 1$, the characteristic timescale for annihilation (in the comoving frame) is therefore

$$t'_{e^+e^-\to\gamma\gamma} = \frac{2}{\sigma_{e^+e^-\to\gamma\gamma} n'_\pm \langle\beta\rangle} \simeq \frac{2}{\sigma_\mathrm{T} n'_\pm c},$$

where we are assuming that $n'_{e^-} = n'_{e^+} = \frac{1}{2} n'_\pm$ (i.e. we are neglecting the electrons associated with protons and assuming that the overall electron density is dominated by the pair plasma).

The optical depth to Thomson scattering will drop precipitously when the $e^{\pm}$ annihilate, i.e. when the characteristic timescale for annihilation is equal to the dynamical timescale $r/c\Gamma(r)$. We saw above that $\Gamma(r) \propto r$, so $\Gamma(r)/r \simeq 1/R_0$ where R_0 is the initial radius of the fireball (recall that initially all the energy of the fireball is thermal and the bulk kinetic Lorentz factor $\Gamma \sim 1$). Hence this freeze-out occurs when

$$n'_{\pm} \sim \frac{2}{\sigma_{\mathrm{T}} R_0}.$$

Substituting this into equation (4.25) and solving for T' gives[406] $T' = 20.5$ keV, corresponding to $\Gamma_{\text{freeze}} \simeq 64$ and $R_{\text{freeze}} \simeq 1.7 R_0 \Gamma_{\text{freeze}}$.

Beyond R_{freeze}, the $e^{\pm}$ of the original pair plasma have annihilated, and the remaining electrons are those associated with the protons (assuming that the original gas is mostly in the form of hydrogen, there will be one electron for every proton). The proton number density can be calculated from the mass outflow:

$$n'_p = \frac{\dot{M}}{4\pi r^2 \Gamma m_p c} = \frac{L}{4\pi r^2 \gamma_0 \Gamma m_p c^3}, \tag{4.26}$$

where $\dot{M} = \mathrm{d}M/\mathrm{d}t$ is the mass outflow rate, L is the luminosity, i.e. energy output per unit time, and the initial thermal γ_0 is therefore $\gamma_0 = L/\dot{M}c^2$. The electrons associated with these protons will therefore dominate the Thomson scattering opacity when

$$n'_p\left(R_{\text{freeze}}\right) > \frac{2}{\sigma_{\mathrm{T}} R_0}.$$

Substituting in equation(4.26) and the numerical value for Γ_{freeze}, this condition corresponds to[406]

$$\gamma_0 \lesssim 2 \times 10^6\, L^{1/4} R_0^{1/2}$$

where L is measured in units of 10^{52} ergs (10^{45} J) and R_0 in units of 100 km. If this condition is satisfied, which—according to [406]—"is likely for most GRBs", there will be a smooth transition from $e^{\pm}$ pair-plasma dominated Thomson scattering to Thomson scattering by electrons associated with the ionised hydrogen plasma, and the fireball will continue to expand beyond R_{freeze}.

To determine whether the expansion stops at $\Gamma = \gamma_0$ or when the expanding plasma becomes transparent to Thomson scattering, we need to calculate the optical depth to Thomson scattering for a photon at radius r, which is given by[406]

$$\tau_{\mathrm{T}} = \int (1-\beta)\sigma_{\mathrm{T}} n_p \mathrm{d}r \simeq \int \sigma_{\mathrm{T}} n_p \frac{\mathrm{d}r}{2\Gamma^2} \simeq \sigma_{\mathrm{T}} n'_p \frac{r}{2\Gamma} \simeq \frac{L\sigma_{\mathrm{T}}}{8\pi r m_p c^3 \gamma_0 \Gamma^2},$$

and setting τ_{T} for the Thomson photosphere gives a photospheric radius[406]

$$R_{\mathrm{T}} \simeq 1.8 \frac{L}{\gamma_0 \Gamma^2}, \tag{4.27}$$

where R_{T} is measured in parsecs and L in units of 10^{52} ergs as before. The maximum bulk Lorentz factor occurs when $R_{\mathrm{T}} = \gamma_0 R_0$, i.e. full transfer of thermal energy to kinetic energy occurs just at the Thomson photosphere. Substituting in for γ_0, we find that

$$\Gamma_{\max} \simeq 8.5 \times 10^2 L^{1/4} R_0^{-1/4},$$

with L in units of 10^{52} ergs and R_0 in units of 100 km. If $\Gamma_{\max} < \gamma_0$, the medium becomes optically thin before all the thermal energy has been transformed to bulk kinetic energy. As $m_e c^2 \sim 0.5$ MeV, a Lorentz factor of 850 corresponds to an electron energy of ∼0.4 GeV. Those GRBs in which *Fermi*–LAT has detected photons with energies of tens of GeV must therefore have bulk Lorentz factors very close to this (admittedly order-of-magnitude) theoretical limit.

Origin of the γ-rays

The outcome of the fireball model as outlined above is a baryon-loaded relativistic outflow, possibly associated with high-temperature thermal photons if the medium has become optically thin before the transfer of energy from thermal to bulk kinetic is complete. However, the γ-rays of the prompt burst are not thermal: as seen in section 2.4.4, the spectrum is a broken power law, and the likeliest production mechanisms for such a spectrum are synchrotron radiation and inverse Compton scattering. A baryonic jet with a Lorentz factor of a few hundred will not generate ∼1 MeV γ-rays directly by either of these mechanisms: we need some method of reconverting some of the bulk kinetic energy into kinetic energy of a population of non-thermal high-energy electrons, or alternatively modifying a high-temperature thermal photon spectrum into the observed smoothly-broken power law.

The location at which this takes place is constrained by a number of factors. The optical depth to Thomson scattering and $e^{\pm}$ pair production must both be fairly small (< 1), since Thomson scattering would degrade the γ-ray energies below what is observed, and $e^{\pm}$ pair production would cause a sharp decrease in the number of γ-rays above ∼1 MeV, contrary to observation. Equation (4.27) indicates that, for Lorentz factors of order 100 and a luminosity of order 10^{52} ergs, this implies $R_\gamma \gtrsim 10^{10}$ m from the lack of Thomson scattering; the lower limit from $e^{\pm}$ pair production is a couple of orders of magnitude more stringent than this[406], depending on the highest photon energies observed (GeV photons need to come from further out than this, which is interesting in view of the fact that the GeV emission tends to be a few seconds later than the prompt burst of MeV γ-rays, as shown in figure 2.48).

The upper limit for R_γ is the deceleration radius, at which the outward motion effectively stalls, as discussed in the previous section. Beyond this point, there is effectively no excess energy to be converted to γ-rays.

There exist[415] prompt emission models which invoke every available scale from the Thomson photosphere to the deceleration radius. The most popular model invokes internal shocks, supplemented by direct emission from the photosphere in the case of those GRBs which have a “thermal bump” in the prompt spectrum on top of the Band power law. The radius of an internal shock can be estimated from $R_{\rm IS} \sim \Gamma^2 c\delta t$, where δt is the variability timescale of the prompt burst. As can be seen in figure 4.19, the variability timescale varies from burst to burst, and some bursts seem to display activity with multiple timescales, so $R_{\rm IS}$ can span a wide range of radii. It is envisaged that a GRB may have multiple internal shocks, which can produce efficient acceleration where they overtake each other; complex behaviour such as this might go some way to explaining the extraordinary diversity of burst profiles. In the internal shock model, the MeV γ-rays are produced as synchrotron radiation from high-energy electrons accelerated mainly at collisions between internal shocks. The GeV γ-rays may come from synchrotron-self-Compton emission, although it is not clear that the delay between the MeV and GeV γ-rays is well explained by this mechanism[416]. Alternative mechanisms for producing the GeV emission include hadronic cascades initiated by $p + \gamma \to p(n) + \pi^0(\pi^+)$: these would yield neutrinos as well as GeV photons. However[416, 417], it appears to be difficult to reproduce the observed GeV γ-ray spectra in hadronic models without requiring extremely high total burst energies and jet Lorentz factors. Another possibility is *two-zone models*[416], in which the keV–MeV and GeV emission come from different regions and have inherently different timescales. There are many such models with several different proposed sources for the GeV emission, both leptonic and hadronic. In some models it is proposed that the GeV emission comes from late internal shocks, whereas others suggest the external shock (which would relate the GeV emission to the earliest stages of the afterglow rather than the late stages of the prompt emission). None of the proposed models is without problems, and it is of course possible that different mechanisms apply to different GRBs.

The GRB central engine

As discussed in section 2.6.3 and by Kumar and Zhang[406], the central engines responsible for short and long GRBs are expected to be essentially the same despite the significant differences in the prompt emission (the differences in the afterglow reflect differences in the circumburst environment rather than the central power source). In both short and long GRBs, the central engine must be capable of producing energetic jets with Lorentz factors of $\gtrsim 10^2$, probably with intermittent activity to explain the episodic nature of the prompt emission in many cases (see figure 4.19)[4]. The principal suggested models are[406] rapidly-accreting stellar-mass black holes, in which case the GRB is powered by the accretion, and rapidly-rotating, highly magnetised neutron stars (magnetars), where the energy is provided by rapid spin-down of the magnetar.

Hyper-accreting black holes

The luminosity produced by accretion on to a black hole is

$$L_{\mathrm{BH}} = \zeta \dot{M} c^2, \tag{4.28}$$

where ζ is the efficiency with which accreted matter is converted to radiated energy and $\dot{M}$ is the accretion rate. For a GRB peak luminosity of order 10^{44} W and an efficiency of 0.01, this implies an accretion rate of around 0.05 $M_\odot\,\mathrm{s}^{-1}$; Kumar and Zhang[406] quote a range of "0.01 – several". The Eddington luminosity for a 10 solar mass black hole is around 10^{32} W: to achieve the peak luminosities of typical GRBs, the black hole must be accreting at a rate many orders of magnitude greater than this, hence the term "hyper-accreting black hole".

Such a rapid accretion rate implies a thick disc or torus of very hot, dense plasma around the black hole. This will be opaque to photons, which explains the low efficiency for radiative power (we normally think of accreting black holes, e.g. in active galactic nuclei, as having efficiencies $\zeta \sim 0.1$ rather than 0.01 or less). This type of radiatively inefficient accretion process is called an *Advection Dominated Accretion Flow* or ADAF. (Note that ADAFs usually arise in very *low* density accretion flows, where the infalling gas does not radiate because it is essentially collisionless. In this case, on the other hand, the gas is not radiating because it is opaque to photons. The ADAF condition simply refers to the lack

[4] However, Kumar and Zhang[406] point out that the variability in the prompt emission might be caused by relativistic turbulence at the emission site rather than variability of the central power source.

of radiation, not to the reason for this.) Under some conditions, the accretion flow becomes unstable to convection, producing a *Convection Dominated Accretion Flow* (CDAF)[418], in which angular momentum is transported inwards but accretion is strongly suppressed. A CDAF does not provide much energy to power a GRB, but can strongly affect the supernova explosion in a massive star core collapse since it will typically drive a (non-relativistic) outflow.

ADAF and CDAF conditions involve only plasma and photons. However, at extreme temperatures and densities such as might be found close to the inner edge of the plasma torus, the effects of neutrinos cannot be neglected. The accreting material then cools by neutrino emission, producing a *Neutrino Dominated Accretion Flow* or NDAF[406, 418].

In the NDAF regime, neutrinos are emitted by processes such as $e^+e^- \to \nu\bar{\nu}$, $p+e^- \to n+\nu_e$ and $n+e^+ \to p+\bar{\nu}_e$, with the latter two normally dominant[419]. Such reactions will occur for temperatures $kT \gtrsim (m_n - m_p)c^2$. The neutrinos generally escape, carrying away energy and cooling the gas. NDAFs produce efficient accretion in which nearly all of the mass in the accretion disc is actually accreted by the black hole, in contrast to CDAFs where accretion is highly suppressed and matter tends to be transported outwards.

A hyper-accreting black hole can launch a relativistic jet in several ways, the most studied of which[406] are neutrino annihilation[419] and the Blandford-Znajek process[420]. In the neutrino annihilation model, $e^\pm$ pairs are produced through $\nu\bar{\nu} \to e^+e^-$, producing a relativistic, $e^\pm$ dominated jet with appropriate properties to drive a GRB. In addition, neutrinos will interact with baryons via both charged-current and neutral-current processes, generating a baryonic wind; the resulting jet will therefore have[406] a certain degree of baryon loading, the extent of which depends on the black hole mass, spin rate and accretion rate.

The total rate of neutrino-antineutrino annihilation is given by[419]

$$\dot{N}_{\nu\bar{\nu}} = \dot{n}_{\nu\bar{\nu}} r^3,$$

where r^3 is the volume of the region in which annihilation takes place and $\dot{n}_{\nu\bar{\nu}}$ is the rate of annihilation per unit volume. The latter depends on the annihilation cross-section $\sigma_{\nu\bar{\nu}}$ and the number densities of neutrinos and antineutrinos, n_ν and $n_{\bar{\nu}}$:

$$\dot{n}_{\nu\bar{\nu}} \sim \sigma_{\nu\bar{\nu}} n_\nu n_{\bar{\nu}} \propto n_\nu E_\nu n_{\bar{\nu}} E_{\bar{\nu}},$$

since $\sigma_{\nu\bar{\nu}} \propto E_\nu E_{\bar{\nu}}$ as neutrino interaction cross-sections generally scale with neutrino energy (see section 2.5.2). The number density multiplied by the energy is just the energy flux, $F_{\nu(\bar{\nu})}$, so we have

$$\dot{n}_{\nu\bar{n}u} \propto F_\nu F_{\bar{\nu}}.$$

Because NDAF cooling is efficient, the energy fluxes $F_{\nu(\bar{\nu})}$ are given by the released gravitational energy per unit volume,

$$F_{\nu(\bar{\nu})} \propto \frac{M\dot{M}}{r^3}.$$

As the neutrino-dominated region is close to the inner edge of the disc, $r \propto r_{\rm ms}$, where $r_{\rm ms}$ is the radius of the innermost stable orbit around the black hole. This depends on the black hole spin, and can be expressed as

$$r_{\rm ms} = x_{\rm ms} R_{\rm S} = \frac{2GM}{c^2} x_{\rm ms}, \tag{4.29}$$

where $R_{\mathrm{S}} = 2GM/c^2$ is the Schwarzschild radius and the value of x_{ms} varies from 3 for a Schwarzschild (non-rotating) black hole to 0.5 for a maximally rotating black hole. Substituting into the expression for $\dot{N}_{\nu\bar{\nu}}$ gives[419]

$$\dot{N}_{\nu\bar{\nu}} \propto \frac{\dot{M}^2}{x_{\mathrm{ms}}^3 M}.$$

The total energy supplied by neutrino-antineutrino annihilation is roughly $\dot{E}_{\nu\bar{\nu}} = (E_\nu + E_{\bar{\nu}})\dot{N}_{\nu\bar{\nu}}$, and in a simple thermal model of the neutrino spectrum $E_{\nu(\bar{\nu})} \propto T_{\mathrm{eff}} \propto F^{(}_{\nu(\bar{\nu})}1/4)$. This gives a final result of[419]

$$\dot{E}_{\nu\bar{\nu}} \propto x_{\mathrm{ms}}^{-15/4} \dot{M}^{9/4} M^{-3/2}. \tag{4.30}$$

Using numerical integration to solve a more realistic model, Zalamea and Beloborodov[419] in fact found that the dependence on x_{ms} was somewhat stronger, $\dot{E}_{\nu\bar{\nu}} \propto x_{\mathrm{ms}}^{-4.8}$. Evaluating the constant of proportionality gives[419]

$$\dot{E}_{\nu\bar{\nu}} \sim 10^{45}\,\mathrm{W} \times x_{\mathrm{ms}}^{-4.8} \left(\frac{\dot{M}}{M_\odot/\mathrm{s}}\right)^{9/4} \left(\frac{M}{3M_\odot}\right)^{-3/2}. \tag{4.31}$$

The observed luminosities of GRBs can be achieved in this model for rapidly-spinning black holes (∼0.95 maximal) with accretion rates of a few tenths of a solar mass per second, which is not unreasonable. The dependence of this result on the black hole spin rate is very strong: for a non-rotating black hole, the required accretion rate is about ten times higher.

A alternative mechanism for launching a jet from a hyper-accreting black hole is the Blandford-Znajek process[420]. In this case, the energy to power the jet comes from the rotational energy of the black hole, which is extracted through magnetic braking.

In a rotating plasma torus, the movement of the charged particles generates a poloidal magnetic field (i.e. one that wraps around the body of the doughnut, threading through the central hole; a field that goes around the ring of the doughnut is a toroidal field). This can produce a very large magnetic field close to the black hole event horizon, magnetically coupling the black hole and the accretion disc. Assuming that the black hole is rotating faster than the disc, this transfers energy and angular momentum from the black hole to the disc, and can drive an electromagnetic jet.

The rotational energy of a black hole with mass M and angular momentum J is[406]

$$E_{\mathrm{rot}} = 1.8 \times 10^{47}\,\mathrm{J} \times f_{\mathrm{rot}}(a_s)\frac{M}{M_\odot},$$

where $f_{\mathrm{rot}}(a_s) = 1 - \sqrt{(1-q)/2}$, $q^2 = 1 - a_s^2$, and $a_s = Jc/GM^2$ is the dimensionless spin parameter of the black hole. As $0 \le a_s \le 1$, the maximum value of f_{rot} is $1 - 2^{-1/2} = 0.29$.

The total power extracted is[406]

$$\dot{E}_{\mathrm{BZ}} \simeq 1.7 \times 10^{43}\,\mathrm{J} \times a_s^2 \left(\frac{M}{M_\odot}\right)^2 B_{15}^2 F(a_s), \tag{4.32}$$

where B_{15} is the magnetic field in units of 10^{15} gauss (10^{11} T) and the numerical factor F is of order unity (it increases from $\frac{2}{3}$ to $\pi - 2$ as a_s goes from 0 to 1).

Expressed as an "efficiency" ζ in the sense of equation (4.28), this power can correspond to $\zeta > 1$, because the energy is coming from the black hole rotation and not from the accretion. The principal difficulty in estimating $\dot{E}_{\mathrm{BZ}}$ for a given system is determining the magnetic field B_{15}[406]: different methods of estimating B_{15} give different results and hence different values of $\dot{E}_{\mathrm{BZ}}$.

Regardless of the details, the fact that $\dot{E}_{\mathrm{BZ}} \propto \dot{M}^2$ implies that a high accretion rate is required for high BZ power, and this in turn implies that the $\nu\bar{\nu}$ annihilation driven jet mechanism will still operate in systems that have BZ powered electromagnetic jets. Neutrino-driven winds will also still exist, but the magnetic field will prevent protons from drifting into the BZ jet. However, the jet can still be baryon-loaded to a lesser degree as a result of neutron drift[406].

Effect of the progenitor system

As discussed in section 2.6.3, long-soft GRBs are, with a few unexplained exceptions, securely associated with Type Ic–BL supernovae, while short-hard GRBs are believed to arise from compact object mergers (NS–NS or NS–BH). This difference means that jets produced near the event horizon of a newly formed black hole need to propagate through a dense stellar envelope in the former case, but escape essentially unhindered in the latter. This has a number of observable consequences[406]:

- highly magnetised electromagnetic jets (e.g. BZ jets) are "protected" from the ambient material by their magnetic field, and thus require less energy to escape the stellar envelope than non-magnetic baryonic jets;
- Kelvin-Helmholtz instabilities[5] develop on the boundary layer between the jet and the surrounding envelope—these may generate variability in the jet even if the central engine is essentially stable;
- the presence of the stellar envelope forces jets, whether magnetised or not, to be more collimated than those produced in systems without an envelope.

It is also reasonable to expect the period of active hyper-accretion to be much shorter in the absence of a stellar envelope, which may account for the characteristic difference in timescales between long and short GRBs.

Millisecond magnetars

A magnetar is a neutron star with an extremely strong magnetic field (surface field $\sim 10^{11}$ T). A millisecond magnetar, as its name suggests, is a magnetar with spin period *sim*1 ms. The total spin energy of a magnetar is

$$E_{\mathrm{rot}} = \frac{1}{2} I \Omega^2 = \frac{1}{5} M \left(\frac{2\pi R}{P} \right)^2 \tag{4.33}$$

which yields, assuming uniform density, 2.2×10^{45} J for a period of 1 ms, a radius of 10 km, and a typical neutron star mass of 1.4 $M_{\odot}$. This represents the maximum possible total energy emitted by a magnetar-powered GRB. Lü and

[5] These are caused by a difference in velocity between two adjacent media, or within a continuous medium, and produce waves at the interface. Water waves produced by wind blowing across a lake or ocean are examples of KH instabilities; so are the patterns at the boundaries of cloud bands on Jupiter and Saturn. For clear diagrams and animations, see [421].

Zhang[422] classified *Swift* GRBs as likely or unlikely to be magnetar-powered, based on the shape of the X-ray light curve (see below), and found that this bound is respected by those considered plausible magnetar candidates, but not by those whose light curves did not match the predictions of the magnetar model.

Following Longair[171] section 13.5, if the magnetic dipole axis of a neutron star is misaligned with its spin axis, electromagnetic radiation is emitted according to

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mu_0 |\ddot{\mathrm{p}}_m|^2}{6\pi c^3}, \tag{4.34}$$

where p_m is the observed magnetic dipole moment. This equation is simply equation (2.15), with the electric dipole term $Q^2|\ddot{\mathbf{r}}|^2/4\pi\epsilon_0$ replaced by the magnetic equivalent $\mu_0|\ddot{\mathrm{p}}_m|^2/4\pi$. For a rotating magnetic dipole observed at a large distance,

$$|\mathrm{p}_m| = p_{m0} \sin \Omega t,$$

where p_{m0} is the component of the magnetic dipole perpendicular to the spin axis, and Ω is the angular velocity. Therefore

$$|\ddot{\mathrm{p}}_m| = -\Omega^2 p_{m0} \sin \Omega t$$

and hence the average power radiated is

$$-\left\langle \frac{\mathrm{d}E}{\mathrm{d}t} \right\rangle = \frac{\mu_0 p_{m0}^2 \Omega^4}{12\pi c^3} \tag{4.35}$$

(note: Longair has a factor of 2 greater than this, but I can't see how he avoids introducing a factor of $\frac{1}{2}$ in averaging over $\sin^2 \Omega t$, and other sources give this result).

Assuming that the magnetic field of the magnetar is a simple dipole, the surface magnetic field B and the magnetic dipole moment are related by

$$p_{m0} \simeq \frac{4\pi R^3 B}{\mu_0},$$

and substituting this into equation (4.35) gives

$$-\left\langle \frac{\mathrm{d}E}{\mathrm{d}t} \right\rangle = \frac{4\pi \Omega^4 R^6 B^2}{3\mu_0 c^3}.$$

Now, from equation (4.33), the energy radiated must be

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = -I\Omega \frac{\mathrm{d}\Omega}{\mathrm{d}t}$$

(the minus sign is because the energy radiated is the energy *lost* by the magnetar). Equating the above two expressions gives

$$\frac{\mathrm{d}\Omega}{\mathrm{d}t} = -\frac{4\pi \Omega^3 R^6 B^2}{3c^3 \mu_0 I}. \tag{4.36}$$

Separating the variables, we have

$$\int\limits_{\Omega_0}^{\Omega_t} \frac{\mathrm{d}\Omega}{\Omega^3} = -\frac{4\pi R^6 B^2}{3c^3 \mu_0 I} \int\limits_0^t \mathrm{d}t,$$

where Ω_0 is the initial angular velocity and Ω_t is the angular velocity at time t. This gives

$$\frac{1}{2\Omega_t^2} - \frac{1}{2\Omega_0^2} = \frac{4\pi R^6 B^2 t}{3c^3\mu_0 I},$$

or

$$\frac{1}{\Omega_t^2} = \frac{1}{\Omega_0^2}\left(1 + \frac{8\pi R^6 B^2 \Omega_0^2 t}{3c^3\mu_0 I}\right)$$

which can be written

$$\Omega_t = \Omega_0\left(1 + \frac{t}{t_0}\right)^{-1/2}, \tag{4.37}$$

where

$$t_0 = \frac{3c^3\mu_0 I}{8\pi R^6 B^2 \Omega_0^2}.$$

Therefore, we expect the luminosity of a GRB powered by magnetar spin-down to be given by

$$L_{\rm MSD}(t) = \frac{L_0}{\left(1 + \frac{t}{t_0}\right)^2} \simeq \begin{cases} L_0, & t \ll t_0 \\ L_0(t/t_0)^{-2}, & t \gg t_0 \end{cases} \tag{4.38}$$

where $L_0 = I\Omega_0^2/2t_0$. For a magnetar with a mass of 1.4 $M_\odot$, an initial period of 1 ms, a radius of 10 km and a surface magnetic field of 10^{12} T, this gives $t_0 \simeq 23$ s and $L_0 \simeq 10^{44}$ W, consistent with the duration and peak luminosity of a typical long GRB.

In fact, a simple dipole spin-down is not adequate to describe the early stages of a magnetar-powered GRB[406, 423]. The newly born magnetar is extremely hot, and neutrino-driven winds are produced, much as in the black hole model. However, unlike the black hole case, the magnetar has a real surface, and neutrino interactions can drive mass loss from the neutron star itself. This leads to a heavy baryon loading of the wind, and the resulting outflow is quite slow. Only as the neutron star cools down so that neutrino interactions become less important does the outflow accelerate to relativistic speeds. After ∼30 s, the proto-magnetar becomes entirely transparent to neutrinos, and the magnetisation parameter $\sigma_0 = \phi^2\Omega^2/\dot{M}c^3$ (where ϕ is the magnetic flux per unit solid angle) increases sharply. In the model of Metzger et al.[423], this increase in σ_0 prevents further jet acceleration and terminates the prompt phase of the GRB (see figure 4.22), though Kumar and Zhang[406] express some doubts about the rapidity of the turn-off.

As the GRB in this model is generated by a rapidly-spinning magnetic neutron star, one expects the development of a striped wind as shown in figure 4.15, with the consequent possibility of particle acceleration by magnetic reconnection, as opposed to or in addition to the internal-shock hypothesis discussed earlier. When jet acceleration is terminated by the sudden rise in σ_0, a significant fraction of the magnetar's spin energy is still available to power later phenomena. In particular, the presence in some GRBs of a "plateau" in the X-ray emission immediately following the GRB proper is often ascribed to magnetar spin-down[423], and it was this feature that was used as a diagnostic by Lü and Zhang[422] in classifying *Swift* GRBs as likely or unlikely to be magnetar-powered. Somewhat surprisingly, Lü and Zhang found this feature in short as well as long GRBs: one would naïvely expect that the product of a merger of two neutron stars would be much too massive to produce anything other than a black hole, but the very high spin rate can stabilise a supramassive neutron

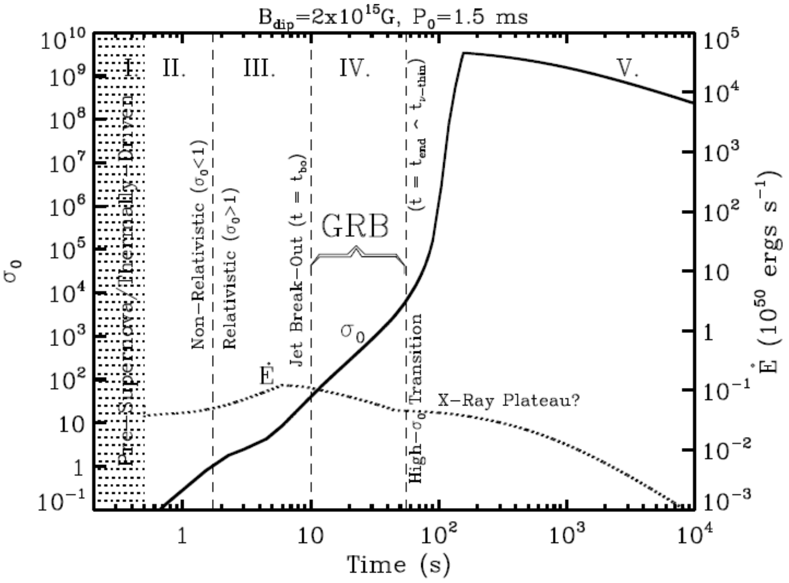


Figure 4.22: Time evolution of a magnetar-drive long GRB showing the wind magnetisation σ_0 (solid line, left-hand scale) and the wind power $\dot{E}$ (dotted line, right-hand scale) as a function of time. This is a less extreme model than that described in the text, with initial period 1.5 ms and surface field 2×10^{11} T. The GRB prompt emission occurs from about 10 to 55 s after the core collapse: at $t < 10$ s the jet is still inside the stellar envelope, and the GRB is terminated by the sudden rise in σ_0 at $t > 55$ s. Figure from Metzger et al.[423].

star[406], at least for a short time. An issue here is that the ~20 s timescale of magnetar spin-down does not seem consistent with the < 2 s duration of a short burst. Some short GRBs do have a period of "extended emission"[223], softer than the prompt spike and lasting for the requisite tens of seconds: in some models[223, 423], such GRBs are produced by the accretion-induced collapse (AIC) into a millisecond magnetar of a white dwarf in a close binary. (AIC is perhaps more likely than core collapse to produce a millisecond magnetar, because the angular momentum of the accreted material (or orbital angular momentum in the case of coalescence of a white-dwarf binary) should spin up the produced neutron star.) In this model, the initial short burst is powered by accretion on to the magnetar, and the extended emission by magnetar spin-down.

Central engine diagnostics

Both the hyper-accreting black hole model and the millisecond magnetar model appear to satisfy the basic energy requirements of typical GRBs. Are there observational properties which might distinguish between the two models, and, if so, which gives better agreement?

One possible diagnostic has already been mentioned: the maximum total energy available from a millisecond magnetar, given plausible parameters, is around 2×10^{45} J. Any GRB whose energy exceeds this is most unlikely to be powered by a magnetar. The difficulty here lies in estimating the opening angle of the jet: for a given measured flux, a narrow jet requires less total energy than a wide jet, because the emission covers a smaller fraction of the total solid angle. Where the jet opening angle has been measured from a spectral break (see section 4.4.1), it is usually found that the total energy required is less than the magnetar limit[406]; however, some of the brightest GRBs, such as

GRBs 050820A, 050904 and 070125, seem to have collimation-corrected total energy release well in excess of 10^{45} J and therefore pose a serious problem for magnetar models[424]. However, even if we conclude from this that these rare "hyper-energetic" GRBs must be powered by a black hole, it does not follow that all GRBs are so powered.

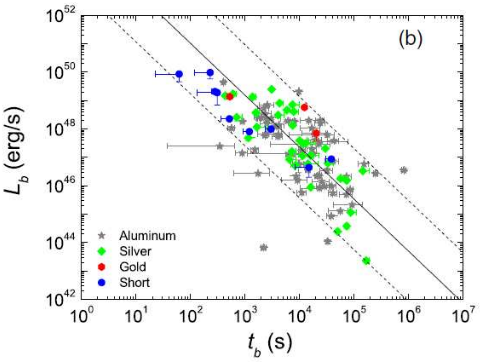


Figure 4.23: Anticorrelation of plateau duration and luminosity for *Swift* GRBs with X-ray plateaus[422]. The "gold", "silver" and "aluminum" categories represent decreasing levels of confidence in the magnetar model: "gold" events have a clear plateau inconsistent with an external shock origin, "silver" events are highly consistent with magnetar model predictions (but do not exclude alternative interpretations), and "aluminum" events are not entirely consistent with simple magnetar models but might be explained by modefied models. The fitted line is $\log L_b = (-1.83 \pm 0.20) \log t_b + (0.20 \pm 0.18)$, where L_b is measured in units of 10^{49} ergs/s and t_b in units of 1000 s. Only the "gold" and "silver" GRBs are included in the fit. Figure from Lü and Zhang[422].

Another challenge for magnetar models lies in the fact that the X-ray emission decreases sharply at the end of the prompt burst, which is interpreted[406] as an indication that the central engine turns off very rapidly. This is easier to explain in a hyper-accreting black hole model, where the transition from NDAF to ADAF naturally produces such a turn-off, than in the magnetar model, where the spin-down power decreases gradually. There are magnetar models which produce a rapid turn-off, such as the argument by Metzger et al.[423] that the rapid rise in σ_0, which occurs when the magnetar becomes transparent to neutrinos, can cut off jet acceleration, but this argument has not convinced others in the field[406].

On the other hand, the X-ray plateau observed in some GRB light curves is probably easier to explain in the magnetar model. Black hole models can produce such a plateau during accretion of the outer part of the pre-supernova star's envelope[406], but getting this to match observations requires some degree of fine tuning. In the magnetar model, the plateau is associated with the magnetar spin-down timescale as indicated by equation (4.38), and the luminosity during the plateau (L_0) should therefore be inversely correlated with its duration (t_0). This is in principle testable, and the data presented by Lü and Zhang[422] do appear to show such an anticorrelation, albeit with a slope closer to –2 than –1 (see figure 4.23).

As first observed by *Swift*, a large minority of GRBs show late-time X-ray flares[425]. These flares occur in both long and short GRBs, and in the closely related X-ray flashes[6]. They appear to be correlated with the prompt emission rather than the afterglow; many are essentially impossible to explain with an external shock model, and the light curve of a typical flare is very like that of a typical peak in the prompt γ-ray emission. They are therefore

[6] As noted on page 134, X-ray flashes (XRFs)[297, 426] have properties very similar to long GRBs, but the prompt emission consists of soft (few keV) X-rays rather than γ-rays. Their relation to classical GRBs is clear, but its nature is not well understood: they may be classical GRBs seen slightly off-axis, but some features of their light curves seem inconsistent with this interpretation.

generally regarded as evidence of late-time activity by the central engine. Such activity seems inherently more likely with a magnetar engine than with a hyperaccreting black hole, which might be expected to turn off rather sharply at the close of the NDAF phase. The observed properties of X-ray flares, and in particular their pronounced time evolution such that later flares are softer, broader and weaker[425], have been interpreted in terms of both black hole and magnetar models, not entirely successfully: Kumar and Zhang[406] argue that the time evolution of the flare luminosity, $E_{\rm XF} \propto t^{-2.7}$, is natural in black hole models but not in magnetar models, whereas the sharp rise and fall of individual flares is natural in magnetar models but not in black hole models.

Overall, the choice between black hole and magnetar central engines remains open: some ultraluminous GRBs are almost definitely black hole powered, but may not be typical, while Lü and Zhang's "gold" sample matches magnetar predictions very well[422], but again represents a small subset of the total sample. It is by no means impossible that both central engine types occur, with as yet undetermined relative frequency.

Progenitors

As discussed in section 2.6.3, the long-soft/short-hard empirical division of GRBs is physically real (see, e.g., [223]) and corresponds to two distinct progenitor types. Broadly speaking, long GRBs are believed to be caused by massive star core collapse and short GRBs by compact object mergers.

Because most GRBs are at high redshift, and hence even a supernova would be a faint object, only a handful of long GRBs have solid associations with supernovae. Hjorth and Bloom[299] list 30 candidates, of which five have solid spectroscopic evidence, six have solid photometric evidence (a supernova-like light curve superimposed on the GRB afterglow) with some spectroscopy, and a further eight have no spectroscopy but a clear "bump" in the light curve consistent with an underlying SN similar to the 11 with spectroscopy. The remaining 11 have a bump in the light curve, but it is poorly sampled, of low significance, or not consistent with the properties of the confirmed associated SNE, or the GRB does not have a measured redshift so the properties of the putative SN cannot be accurately determined. Since the publication of [299], further GRB/SN associations have been reported: see table 4.2 for a list. Overall, the evidence indicates that the majority of long GRBs are associated with Type Ic supernovae, although the converse is not true: even given the fact that most GRB events will not be seen by us because the jets do not point in our direction, it is most unlikely that all SNe–Ic launch GRBs[299, 406].

There do exist GRBs classified as "long" which are definitely *not* associated with luminous Type Ic supernovae. Figure 4.24 shows GRB 060505 and GRB 060614, nearby long bursts with strong upper limits on light from an associated supernova. In the case of GRB 060614, this may be a quirk of the classification system: this GRB has an initial short intense spike followed by ∼100 s of less intense emission, and some authors, e.g. Barthelmy[436] class it as a short GRB with extended emission rather than a long GRB. The issue is that if the extended emission contributes <10% of the γ-ray energy, it will not be included in the burst duration T_{90} and the burst will be classed as "short" on the basis of its initial spike, whereas a very similar burst with slightly more intense extended emission may have a much longer nominal duration because the calculated T_{90} will include the extended emission period. Thus, $T_{90}(\text{GRB}\,060614) = 103$ s, but the duration of the initial spike is only ∼5 s—still nominally long, but in the

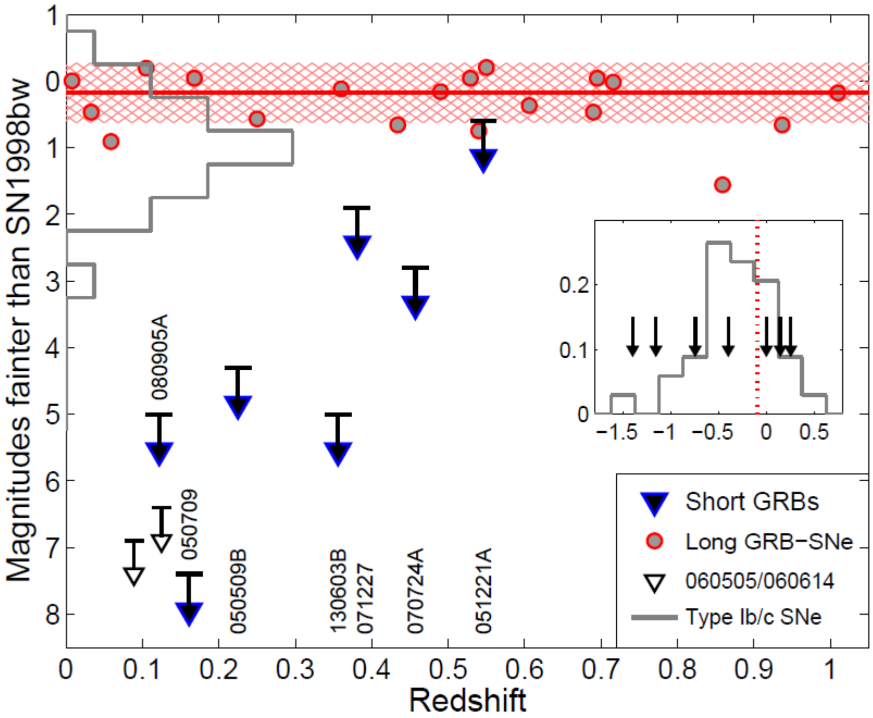


Figure 4.24: Association (or otherwise) of GRBs and supernovae[223]. The filled circles represent the supernovae associated with long GRBs[299]; the red line and hatched band show the mean and standard deviation of this distribution. These SNe are brighter than typical SNe Ibc, whose absolute magnitude distribution (relative to SN 1998bw) is shown in the grey histogram on the y axis. Blue arrows show the upper limits on associated SN magnitude for short GRBs: six of the seven measurements available at the time this plot was made clearly rule out SNe similar to those associated with the long GRBs. The two black arrows are the anomalous long GRBs 060505 and 060614, which despite being nearby objects did not have any detectable associated supernova. Finally, the inset plot shows the duration distribution (in units of $\log_{10} t$) of short GRBs, with the seven from the main plot shown as arrows. It has been argued that the division between "short" and "long" GRBs in the *Swift* sample should be placed at 0.8 s rather than the canonical 2 s: this value is shown by the red dashed line in the inset plot. Three of the plotted short GRBs are on the "long" side of this line, but none has an associated SN. Figure from Berger[223].

dip between the two peaks of figure 2.49 rather than incontestably in the "long" peak. The location of GRB 060614 is also atypical of long GRBs[223]: its host galaxy has a lower-than-average specific star formation rate, and the GRB itself was well off-centre and not in a region with an obvious massive star population. Overall, it may well be fair to lump GRB 060614 in with the subclass of short GRBs with extended emission as opposed to the classical long GRB class.

GRB 060505 is also quite short by long GRB standards, with a burst duration of 4–5 s, but it had a positive spectral lag (low-energy photons arrive later than high-energy photons, a common characteristic of long GRBs not seen[223] in short GRBs), occurred in a star-forming region and generally looked like a long GRB at the short end of the T_{90} distribution, not a short GRB at the long end. In the context of the association of long GRBs with supernovae, it is significantly more difficult to dismiss GRB 060505 as a misclassified short GRB than is the case with GRB 060614.

van Putten et al.[437] also classify GRB 061021, which does not appear in

Spectroscopically confirmed GRB–SNe				
GRB	**SN**	z	**Class**	**Reference**
980425	1998bw	0.0085	A	[299]
011121	2001ke	0.362	B	[299]
020903		0.251	B	[299]
021211	2002lt	1.006	B	[299]
030329	2003dh	0.1685	A	[299]
031203	2003lw	0.1055	A	[299]
050525A	2005nc	0.606	B	[299]
060218	2006aj	0.0334	A	[299]
081007	2008hw	0.530	B	[299]
100316D	2010bh	0.0591	A	[299]
101219B		0.552	B	[299]
111209A	2011kl	0.677	A	[427]*
120422A	2012bz	0.283	A	[428, 429]
130215A	2013ez	0.597	A	[430]
130427A	2013cq	0.3399	A	[431]
130702A	2013dx	0.145	A	[432, 433]
130831A	2013fu	0.4791	A	[434, 430]

Table 4.2: GRBs clearly associated with spectroscopically confirmed supernovae. The first 11 are as listed by Hjorth and Bloom[299], with their taxonomy (class A: strong spectroscopic evidence; class B: strong photometric evidence with some spectroscopy); the remaining 6 are spectroscopically confirmed associations observed since the publication of [299].
* GRB 111209A belongs to the proposed class of ultra-long GRBs[435] and its associated supernova is spectroscopically and photometrically distinct from the typical GRB-associated SNe[427].

figure 4.24, as a long GRB with no associated supernova[7], presumably on the basis of the lack of any bump-like features in its light curve: there seems to be no published formal upper limit on SN light for this GRB. Like GRB 060614, it has an initial γ-ray spike consistent in its properties with short GRBs, followed by a softer tail consistent with long GRBs.

Owing to the very small sample, it is difficult to draw strong conclusions about the nature of long GRBs without supernovae. On the basis of GRB 060614, and to a lesser extent GRB 061021, it would be tempting to assimilate them into the class of short GRBs with extended emission, as do van Putten et al.[437] and (for the former) Barthelmy[436], but it is more problematic to invoke this as an explanation for the rather short, but otherwise typical, GRB 060505.

Further complicating the issue, some long GRBs are *extremely* long, with T_{90} values of hours rather than minutes. While some authors, e.g. Virgili et al.[438], regard these objects as simply the long-duration tail of the long GRB distribution, others, e.g. Levan et al.[435], argue that they represent a distinct population. The different viewpoints may arise partly from a difference in definition: Virgili et al.[438] use GRB 091024A (duration ~1300 s) as their prototype and include all bursts with durations $\gtrsim$ 1000 s, whereas Levan et al[435] consider only the three most extreme cases, with durations of $\sim 6000 - 10000$ s (noting that such long durations have considerable associated uncertainty, not least because they are considerably longer than the *Swift* satellite's 90-minute

[7] 2006 seems to have been a vintage year for long GRBs without supernovae...

orbital period).

Two of the three ultra-long GRBs selected by Levan et al.[435] have identified host galaxies: in both cases the host is faint, small and blue (somewhat fainter and more compact than the typical long GRB host galaxy, but it is not clear that this is meaningful given the tiny sample). GRB 111209 is associated with a supernova (SN 2011kl) which, though a Type Ic like other GRB-associated supernovae, differs significantly from the SNe plotted in figure 4.24, being a factor of 3 ($\sim$ 1 magnitude) brighter and apparently of lower metallicity[427]. Greiner et al.[427] argue that the properties of SN 2011kl suggest that it was produced by the core collapse of a massive star to form a magnetar, with the high luminosity stemming from the additional energy provided by magnetar spin-down.

Regardless of the status of ultra-long GRBs and long GRBs without supernovae, the host galaxies of classic long GRBs, and the location of the GRBs within those hosts, support the identification of long GRBs with massive stars[223, 406, 439, 440, 441]. Typically, the host galaxies of long GRBs are faint, blue, low-metallicity systems with high specific star formation rates, though there is substantial evolution with redshift. The range of UV (160 nm) absolute magnitudes in the TOUGH survey[441] is –14 to –21.4 magnitudes, with a median value of –18.9: these values indicate that most host galaxies are sub-luminous, although the distribution does stretch up to "normal" luminosities. The distribution evolves considerably with redshift: (relatively) local GRBs ($z < 1$) are much more likely to be found in fainter hosts, whereas more distant bursts ($1 < z < 3$) are more likely to occur in brighter hosts[441]. This may be a consequence of the fact that GRBs appear to occur preferentially in low-metallicity systems (for example, Krühler et al.[440] find that only (18±7) % of GRBs at $z < 1$ occur in galaxies with metallicity higher than solar, whereas $\sim$50% would be expected if GRBs traced the star formation rate): in the local universe, only dwarf galaxies are likely to be metal-poor. The trend towards brighter galaxies reverses for $z > 3$, with nearly all GRBs in this redshift range having hosts fainter than the median[441], but the statistics in this redshift bin are low, and the result may be affected by the presence in the TOUGH sample of some host galaxies without well-determined redshifts.

Long GRBs are not found in galaxies without active star formation, and the locations of the GRBs within their hosts tend to track the star formation activity[223, 406]. Overall, it seems fair to conclude that long GRBs are strongly associated with star formation and prefer low-metallicity systems. The fact that the associated supernovae are all Type Ic, i.e. they lack both H and He lines and thus originate from stripped stellar cores, points to either Wolf-Rayet stars or stars in close binaries as progenitors, and the metallicity dependence is not surprising given that metallicity is known to have a strong effect on star formation and evolution.

In contrast, the host galaxies of short GRBs are not exclusively star-forming: about 20% of short GRBs appear to be associated with early-type host galaxies with old stellar populations[223]. This is conclusive evidence that short GRBs do not (all) have massive star progenitors. However, the rate of short GRBs per unit stellar mass is several times higher in late-type (star-forming) galaxies than in early-type galaxies[223], so at least some progenitors are drawn from younger populations. These are not massive stars in the sense of long GRBs: the stellar population ages of short GRB hosts are systematically higher than those of long GRB hosts even in the late-type galaxies, and there is no preference for low-metallicity systems. Berger[223] concludes that short GRBs “track star formation with a delay of hundreds of millions of years to several gigayears”—that is, assuming that they represent the deaths of their progenitor stars, those stars are likely to be of intermediate to low mass.

This pattern of occurrence in all types of galaxies, but with rate per unit stellar mass increasing from early to late galaxy types, is stikingly reminiscent of the rate of Type Ia supernovae[442], where observations indicate a delay time distribution $\propto t^{-1}$. This observed time dependence matches theoretical predictions assuming the double-degenerate progenitor model for SNe Ia, i.e. the explosion originates from the coalescence of a binary system of two white dwarfs. Since the theoretically preferred model for short GRBs is the coalescence of a binary system of two neutron stars, the similarity of the decay time distributions is likely to reflect the similarity of the progenitor systems. Note that although neutron stars themselves have massive star progenitors, whereas white dwarfs have intermediate to low mass progenitors, the relevant timescale for the GRB/SN Ia is the coalescence of the binary (though we would expect the first GRBs to take place before the first SNe Ia, since the time to evolve to a compact object binary in the first place would be shorter).

The location of short GRBs within their host galaxies also differs from that of long GRBs, as shown in figure 4.25. In general, long GRBs are located within a few kpc of the host centre, with a median offset of just over 1 kpc, whereas short GRBs have a median offset of 5 kpc and range out to 75 kpc[223]. Of particular note is the significant fraction of very large offsets: about 10% of short GRBs have projected offsets >20 kpc from the centre of the putative host. This is much larger than the visible size of a typical galaxy, so the positions of such GRBs do not lie within the optical image of the assigned host. The possibility that they are in fact associated with a much fainter galaxy not visible in the relevant optical image must therefore be considered. However, the upper limits calculated for the brightness of such unseen hosts imply a very large redshift, requiring a bimodal redshift distribution for short GRBs with no difference in GRB properties to motivate this[223]. Also, as shown in figure 4.25, the observed offset distribution, including the tail to very large offsets, is actually quite well described by models of NS–NS mergers.

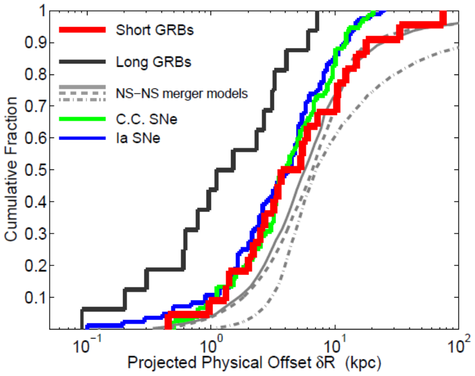


Figure 4.25: Projected offset of GRBs from the centre of the host galaxy[223], compared with core-collapse (green) and Type Ia (blue) supernovae and the predictions of NS–NS merger models (grey). Long GRBs (black) are located significantly closer to the host centre than the other classes, though one should recall that the host galaxies of long GRBs are smaller than average[440, 441]; short GRBs appear comparable to supernovae, but have a tail extending to larger offsets. This tail is well predicted by the merger models. Figure from Berger[223].

One should note that the appearance of figure 4.25, in which long GRBs appear very different from core-collapse supernovae while short GRBs seem similar, is biased by the fact that long GRBs preferentially occur in small galaxies.

If the offsets are normalised to the half-light radius r_e of the host galaxy[8], the median value of the scaled offset $\delta R/r_e$ is $\sim$1 for long GRBs and both classes of supernovae, but $\sim$1.5 for short GRBs, with 20% of short GRBs having $\delta R/r_e > 5$[223]—thus, taking host galaxy size into account, it is in fact the short GRBs that are the outliers and the other three that are comparable. Note that because of the definition of the half-light radius, any population whose distribution tracks the starlight would be expected to have a median $\delta R/r_e \sim 1$, so the short GRB distribution is more extended than the starlight. The likely explanation for this is that NS–NS systems have undergone core-collapse supernovae, and asymmetric supernova explosions impart "kicks" to their progenitor systems that result in rapid motion away from the explosion site (see, e.g., [443]). This would not happen to the WD–WD systems responsible for SNe Ia, because the formation of a white dwarf involves loss of the envelope via a stellar wind rather than an explosion. The models that correctly predict the long tail of large offsets include such kicks[223]. The median kick velocity required to account for the observed distribution is $\sim$60 km s^{-1}[223], which is not unreasonable (in particular, such a kick would not unbind the binary system).

Overall, the observed properties of short GRBs are consistent with the theoretical properties of compact object mergers. The fainter afterglows of short GRBs are naturally explained if short GRBs explode in lower density environments, so that the external shock sweeps up less material: this is entirely plausible if short GRBs are compact object mergers with only ambient interstellar medium surrounding them, while long GRBs are massive star supernovae in star-forming regions and probably surrounded by circumstellar ejecta. Although the statistics are low, there is some evidence that short GRB jets are less collimated than those of long GRBs (see figure 4.18): this is again expected if long GRB jets are collimated by having to punch through a surrounding stellar envelope while short GRB jets are not. The overall energy release in short GRBs is about two orders of magnitude less than for long GRBs[223], but is consistent with predictions from merger models employing the Blandford-Znajek mechanism to launch the jets.

Are there observational "smoking guns" available to test the merger model of short GRBs, in the way that the association of nearby long GRBs with visible SNe Ic confirms the core-collapse model of long GRBs? The most obvious candidate is the detection of a gravitational wave signal, since NS–NS mergers are also the favoured candidates for detection by Advanced LIGO. A coincidence between a short GRB and an aLIGO signal would have multiple benefits, confirming both the reality of the aLIGO signal and the compact object merger origin of the GRB.

The main problem with this scenario is that the design range of aLIGO for NS–NS mergers is only $\sim$200 Mpc[444], corresponding to a redshift of $z \sim 0.05$ (although coincidence with a short GRB would permit softer cuts, approximately doubling this to $z \sim 0.1$[445]), whereas the median redshift of those short GRBs with well determined redshift values is 0.48[223]. Therefore, the fraction of observed GRBs lying inside aLIGO's horizon is small: in fact, as of late 2013 *no* short GRB had a confirmed redshift $z < 0.1$[223]. Conversely, since GRB emission is beamed, many GW signals observed by aLIGO will not correspond to observable GRBs because the jets are not pointing in the right direction. Overall, estimates of the coincidence rate are of order 0.3 per year (quoted by [223]; other sources give comparable values) for aLIGO's design

[8] The half-light radius is, as its name indicates, the radius enclosing half the light of the galaxy.

range, which may take some time to reach. Therefore, we cannot rely on such coincidences to validate the model in the near future.

Rather more promising is the likely association of short GRBs with an event variously termed a “kilonova” [446] or “macronova” [447], which, as its name suggests, is an optical/near-infrared transient with properties intermediate between a classical nova and a supernova. This feature is produced by r-process nucleosynthesis in the neutron-rich environment around the merger[223]: the r-process produces highly unstable, super-neutron-rich nuclides, and it is the decay of these short-lived radioactive species that produces the kilonova emission. The abundance of heavy elements in the ejecta results in heavy line-blanketing, so that the kilonova emission peaks in the near infrared, around 1–3 μm[223, 446]. This is somewhat unfortunate, as the deep, wide-field imaging capacity required for prompt follow-up of short GRBs without detected optical afterglows (and hence with rather large position error boxes) is not currently available at near IR wavelengths. (This is even more of a problem if kilonovae are sought in coincidence with gravitational wave signals, since the angular resolution of aLIGO is very poor even by GRB standards!)

To date, candidate kilonovae have been observed in association with short GRB 130603B[446], and possibly with the “long/short” GRB 060614 discussed earlier[447]. GRB 130603B, a short GRB with $T_{90} \simeq 0.1 - 0.2$ s (depending on waveband) and $z = 0.356$, was localised to high precision owing to the detection of an optical afterglow and hence could be imaged using the HST. Differencing images taken 9 and 30 days after the burst showed a clear transient source at $\lambda = 1.6\,\mu$m which had disappeared in the 30-day image and was not present in optical (600 nm) images[446]. The absence of an optical signal makes it most unlikely that this transient is part of the GRB afterglow, which would not be expected to undergo such a drastic colour change, and the brightness and timescale are consistent with kilonova models[446]. The luminosity of the kilonova suggests that GRB 130603B ejected $\sim 0.05\,M_\odot$ of r-process material[223] (give or take a factor of two), indicating that compact object mergers could be the principal source of r-process nuclides.

The candidate kilonova associated with GRB 060614 is based on careful re-examination of the late-time light curve of its optical afterglow. The evidence is a small but significant excess in the HST F814W light curve around 14 days after the burst. No excess is observed in the VLT R band, again pointing to a near-infrared rather than optical transient[447] (the VLT R-band covers the wavelength range ~570–740 nm, the HST F814W band ~710–970 nm)[9].

Kilonovae are faint compared to supernovae, and therefore likely to be observable only to modest redshifts; on the other hand, their emission is isotropic rather than beamed. This probably makes them more promising as optical counterparts to gravitational wave signals than actual short GRBs: it should be possible to observe the kilonova even when the GRB jets are not aligned along our line of sight. The problem, as noted above, is that there are no wide-field near-IR survey instruments suitable for conducting a rapid search over the likely error box of a GW signal detection. However, the standard photometric I band, centred on 806 nm, and the SDSS i' and z' bands, centred at 770 and 910 nm respectively, are probably red enough to detect kilonovae (the I band is very similar to the HST F814W band) and are available on optical survey instruments. Further detections of kilonovae in conjunction with short GRBs

[9] A recent preprint[448] claims that a small excess can also be seen in the VLT R and I bands. This is based on using the early-time observations (1.7–3 days post burst) to constrain the afterglow contribution.

would confirm the merger model and might—especially if spectroscopic information were obtained—provide insight into the r-process in compact object mergers.

4.4.2 Gamma-ray bursts as sources of UHE cosmic rays

GRBs are one of the favoured candidate sources for ultra-high-energy (UHE) cosmic rays, which are generally assumed to be extragalactic because they are too energetic to be magnetically confined in the Milky Way. The basic energetics of GRBs seem to be broadly consistent with this possibility, and the presence of relativistic shocks and—particularly in the magnetar model—very large magnetic fields provides plausible sites for particle acceleration. The principal questions that have to be asked are:

- how much does the lack of observed high-energy neutrinos associated with GRBs[449] constrain GRBs as sources of UHE cosmic rays, and
- is the level of baryon loading required to generate the observed UHE cosmic ray flux consistent with our models of GRB emission?

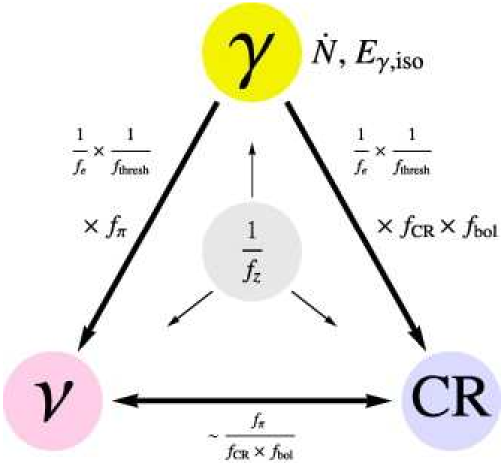

γ
$\dot{N}$, $E_{\gamma,\mathrm{iso}}$
$\frac{1}{f_e} \times \frac{1}{f_\mathrm{thresh}}$
$\frac{1}{f_e} \times \frac{1}{f_\mathrm{thresh}}$
$\times f_\pi$
$\times f_\mathrm{CR} \times f_\mathrm{bol}$
$\frac{1}{f_z}$
ν
CR
$\sim \frac{f_\pi}{f_\mathrm{CR} \times f_\mathrm{bol}}$

Figure 4.26: Schematic of the relationships between the observables relevant to GRBs as the source of UHE cosmic rays[450]. The *gamma*-ray observables are the number of observable GRBs per year ($\dot{N}$) and the equivalent isotropic energy per GRB in γ-rays ($E_{\gamma,\mathrm{iso}}$. The various "fudge factors" connecting the different messengers are the cosmic evolution factor f_z (> 1), the baryon loading of the GRB jets f_e^{-1} ($\gtrsim 10$), the correction for instrument flux threshold f_thresh ($\sim$0.2–0.5), the fraction of baryonic energy going into cosmic rays f_CR and into pion production f_π, and a bolometric correction factor f_bol ($\ll 1$). Figure from Baerwald, Bustamente and Winter[450].

The observables are the (upper limit on) the GRB-associated neutrino flux, the diffuse neutrino flux at high energies, the GRB γ-ray luminosity, the GRB rate, and the observed UHE cosmic ray flux. The relations between these involve various "fudge factors" (see figure 4.26[450]), which are model dependent and uncertain to varying degrees.

The fudge factors shown in figure 4.26 arise from various sources. As discussed in section 2.2.3, cosmic rays with energies above $\sim 10^{19}$ eV have limited range ($\lesssim 100$ Mpc) owing to interactions with the cosmic microwave background, whereas observed GRBs are mainly at high redshift, so the factor f_z is introduced to describe the evolution in GRB rate over time. We expect this factor to be > 1, i.e. GRB rate increases at higher redshift, because of the association of long GRBs with massive stars, and hence with the star formation rate (and also their preference for low metallicity host galaxies, see above); values from the literature are typically of order 10. The threshold factor, f_thresh, is introduced to correct the *observed* GRB rate, which depends on the detection and trigger thresholds of the various γ-ray telescopes, to the total rate: there will surely be GRBs which do not produce a high enough γ-ray flux to trigger the detectors, but which contribute to the cosmic-ray production. The value of f_thresh depends on the individual instrument as well as the assumed luminosity spectrum of GRBs; Baerwald et al.[450] quote a range 0.2–0.5 based on simulation studies, primarily based on *Swift*–BAT capabilities.

It is assumed that cosmic rays are produced when baryons accelerated by the source escape. The measured γ-ray flux, however, is sensitive to the electromagnetic energy emitted by the GRB. The factor f_e^{-1}, the **baryon loading** of the jet, is therefore required to describe the ratio of energy in protons to energy in electrons (assumed to be in equipartition with γ-ray energy, because of pair production and annihilation in the central regions of the GRB). This factor is model dependent: Baerwald et al.[450] quote a value of $\sim$10. Not all the baryons escape to produce cosmic rays, so we need a factor $f_{\rm CR}$ to describe the fraction of available baryon energy that goes into cosmic-ray production: this is clearly < 1, but its exact value depends on source conditions. Although the modelling of extragalactic cosmic-ray sources is based on the ultra-high energy range $10^{19} - 10^{21}$ eV, the sources surely produce lower-energy CRs as well, so this factor has to be further modified by a "bolometric correction" $f_{\rm bol}$, which is the fraction of the total CR energy contained in the UHE range. This is very dependent on the assumed spectral index of the CR energy spectrum: for the canonical produced spectral index of 2, it is $\ln 10^2/\ln 10^{12} = 0.17$, but for steeper spectral indices it is much smaller—note, however, that lower-energy protons are much less likely to escape, because they will be magnetically confined.

High-energy neutrinos come from the decay of charged pions produced when the high-energy baryons interact within the source, so a factor f_π is introduced to describe the fraction of the total baryon energy that goes into pion production (which is *not* the fraction converted to pions in a single interaction, because the baryons may undergo multiple interactions before escaping). The relationship between the factors $f_{\rm CR}$ and f_π depends on whether the cosmic rays escape as neutrons (in which case there *must* have been a prior reaction of the form $p+X \to n+\pi^+ +X'$, since one can't accelerate neutrons) or as protons: neutron escape implies $f_\pi \sim 0.2$ and $f_{\rm CR} \sim 0.4$ from the decay kinematics, but if the escaping particles are protons we can have $f_{\rm CR} \gg f_\pi$.

The basic observable for extragalactic cosmic rays is the observed rate of UHE cosmic rays, and the local rate at which energy has to be injected into the UHECR population to account for this. Based on the compilation of Gaisser, Stenev and Tilav[87], Baerwald et al. quote a value of 1.5×10^{37} J Mpc^{-3} yr^{-1} for this, about a factor of 3 lower than the original Waxman and Bahcall estimate quoted on page 120 (this is a result of including more recent data). The number of observed GRBs per year is $\dot{N} \sim 1000$.

Baerwald et al.[450] define their evolution factor f_z as

$$f_z = \frac{1}{4\pi D_H^3}\int_0^\infty \mathcal{H}(z)\frac{{\rm d}V/{\rm d}z}{1+z}\,{\rm d}z, \tag{4.39}$$

where $\mathcal{H}(z) = \dot{n}_{\rm GRB}(z)/\dot{n}_{\rm GRB}(0)$ is the comoving GRB rate at redshift z divided by the local GRB rate at $z = 0$, V is volume, and $D_H = c/H_0$ is the Hubble distance. Conveniently, with $H_0 \simeq 70$ km s^{-1} Mpc^{-1} their volume normalisation factor $4\pi D_H^3 \simeq 1000$ Gpc^3 (they quote 968, with somewhat unjustified precision), so it approximately cancels numerically with $\dot{N}$. Using the above values for the energy injection rate and $\dot{N}$, we obtain

$$E_U \simeq 1.5 \times 10^{46}\ {\rm J}\ \times f_{\rm thresh} f_z$$

for the average energy per GRB emitted as UHE cosmic rays (this scales directly as the energy injection rate $\dot{\mathcal{E}}_U$ and inversely as the number of GRBs).

Given that f_z in most models is of order 10 and $f_{\rm thresh} > 0.1$, this is an uncomfortably large number, implying a very high baryon loading in the GRB jets. The relation between E_U and the observed γ-ray energy (calculated as if isotropic) is, by the definitions above,

$$E_U = \frac{f_{\rm CR} f_{\rm bol}}{f_e} E_{\gamma,\rm iso},$$

and $\sim 10^{46}$ J is often quoted as a typical value of $E_{\gamma,\rm iso}$. This means that we need

$$\frac{f_{\rm CR} f_{\rm bol}}{f_e} \gtrsim f_{\rm thresh} f_z \sim 3$$

if GRBs are to supply enough energy to account for the observed flux of UHE cosmic rays.

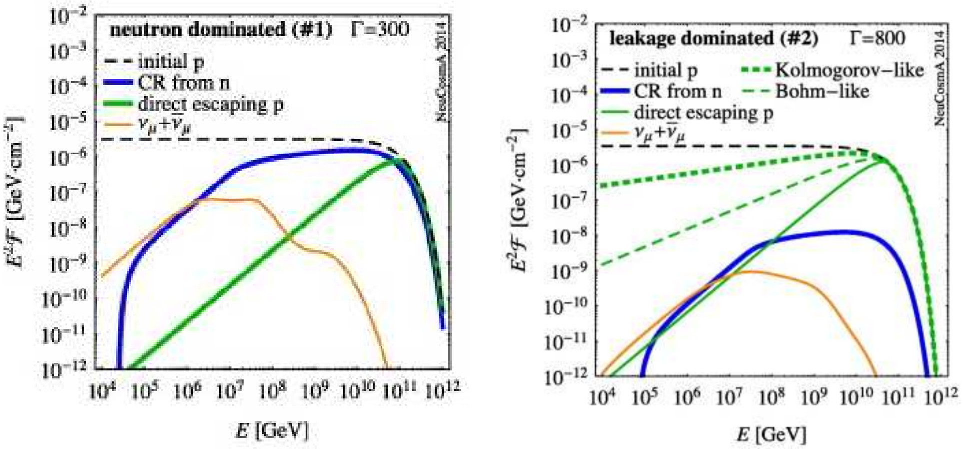


Figure 4.27: The expected CR spectrum from a typical GRB with neutron-dominated (left) or proton-dominated (right) escape mechanisms. Also shown is the expected neutrino spectrum (orange): note that this is about two orders of magnitude higher for neutron-dominated escape. Figure from Baerwald, Bustamente and Winter[450].

A diagnostic for the baryon loading f_e^{-1} (all baryons) or $f_{\rm bol}/f_e$ (UHE baryons) is the neutrino flux, which unlike the γ-ray flux is generated directly from the baryons (recall that the γ-rays from GRBs are *very soft*, at energies of order 1 MeV—π^0 decay is not a factor). The neutrino flux and the cosmic-ray flux scale by a factor $\propto f_\pi/(f_{\rm CR} f_{\rm bol})$: this factor is much larger for neutron escape models (in which the factor $f_\pi/f_{\rm CR}$ is of order $\frac{1}{2}$) than it is for proton escape models, with obvious consequences for the significance of the non-observation of GRB-associated neutrinos by IceCube[449].

There are three possible scenarios for cosmic ray escape:

- all protons are magnetically confined—cosmic rays escape as neutrons which subsequently decay to protons;
- protons escape if they are within their Larmor radius of the edge of the expanding gas shell in which they are being accelerated—this will produce an extremely hard spectrum, as the highest-energy protons will have the greatest chance to escape;
- lower-energy protons can escape by diffusing towards the edge of the shell—this softens the escaping proton spectrum by an amount which depends on the model adopted for the dependence of the diffusion coefficient on energy.

Of course, in practice all of these will occur at some level: calling a particular model a "neutron" model or a "direct escape" model is a statement about which process *dominates* the observed CR flux, not a statement that *only* that process operates.

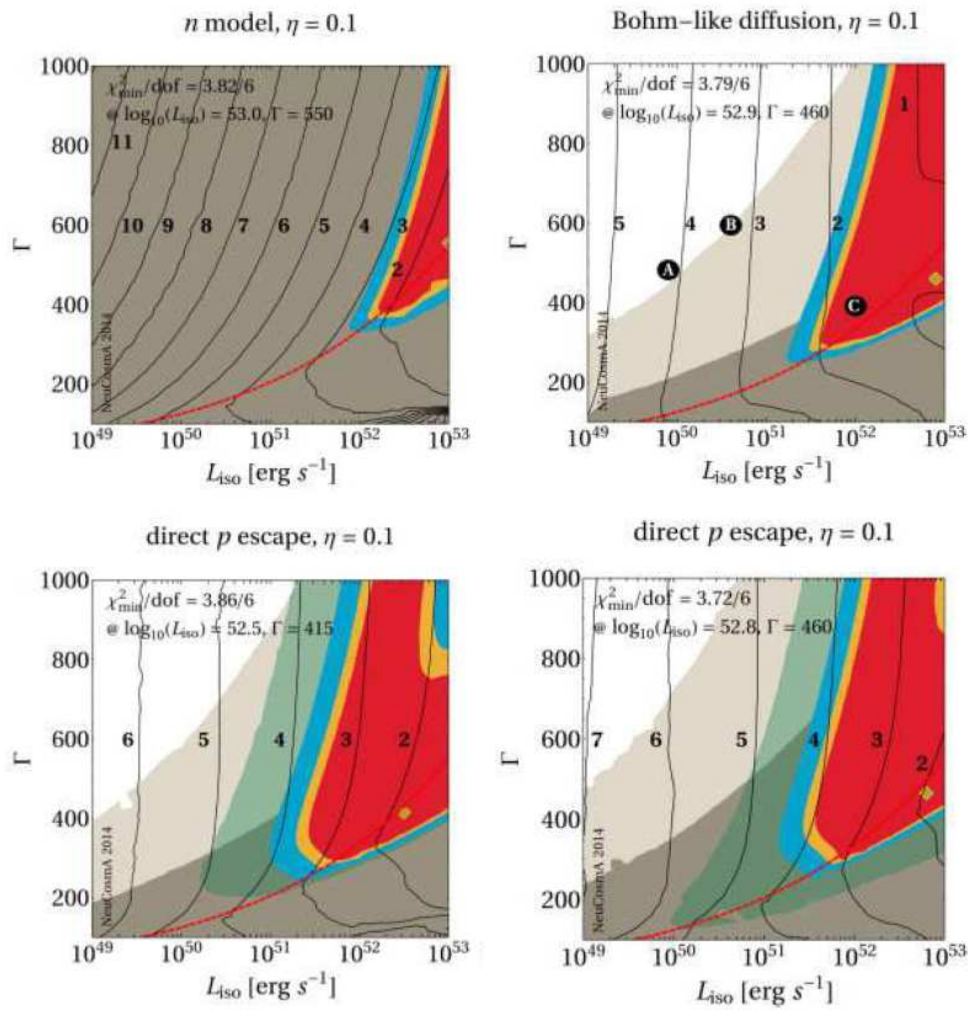


Figure 4.28: Parameter scan of GRB models with acceleration efficiency $\eta = 0.1$[450]. The red, yellow and blue filled areas represent 90%, 95% and 99%, respectively, allowed regions from Telescope Array CR data. The dark grey area represents the region excluded by the 2012 IceCube analysis[278]; the light grey region is the expected exclusion region assuming a null result after 15 years' data, and the green region could be excluded by 15 years' data on cosmogenic neutrinos. The current IceCube limit is somewhere in the light grey region.
Top row, neutron-dominated and Bohm diffusion models, assuming that the GRB rate tracks the star formation rate. Bottom left, direct-escape model with the same assumption; bottom right, direct-escape model with an additional factor of $(1 + z)^{1.2}$ in the GRB evolution. Γ is the bulk Lorentz factor, and the numbered contours are for $-\log_{10} f_e$. Figure from Baerwald, Bustamente and Winter[450].

Figure 4.27[450] shows the expected proton and neutrino spectra for neutron-dominated escape (left) and direct proton escape (right), as well as the proton spectra for two different models of diffusive escape: Bohm diffusion, in which the diffusion coefficient is proportional to the energy, and Kolmogorov diffusion, in which it is proportional to $E^{1/3}$ (in the source rest frame). As expected, the neutrino flux is much higher for the neutron model than for the proton model.

Baerwald, Bustamente and Winter[450] conduct a model parameter scan for neutron-dominated, direct-escape dominated, and Bohm-diffusion models, and apply constraints on the cosmic ray flux as measured by the Telescope Array and the neutrino flux as not seen by IceCube—note that their IceCube curves predate the recent update to this limit[449] and are therefore conservative.

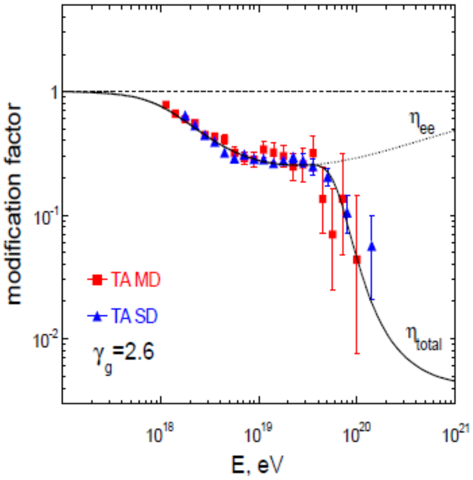


Figure 4.29: The effect of pair-production and the GZK cut-off on the cosmic ray spectrum at high energies[451]. The curves show the effect on an underlying spectral index of 2.6 caused by e^+e^- pair production (dotted line) and the GZK cut-off (solid line). The data are from the Telescope Array in monocular (red) and stereo (blue) mode.

Selected results from this analysis are shown in figure 4.28. The first noteworthy finding is that the neutron-dominated model (top left), with its necessarily high neutrino flux, appears to be entirely ruled out by the 2012 IceCube limits[278]. Allowed regions with plausible parameter values remain for direct-escape and Bohm-diffusion models; in the models shown, these will be detected or ruled out by higher IceCube statistics. (The allowed regions in the 2015 IceCube analysis[449] are unfortunately not plotted against the same parameters, and therefore cannot be directly compared, but seem broadly similar in allowing bulk Lorentz factors of a few hundred (depending on model) combined with baryon loadings of order 10–30, somewhat lower than those preferred by the direct-escape model but consistent with the diffusion model.) For higher acceleration efficiencies, there are other allowed regions at low $E_{\gamma,\rm iso}$ which are not testable by neutrino observations, but these require high bulk Lorentz factors and/or very high baryon loading, and will probably be better tested by improvements in the UHE cosmic-ray data.

The plots shown in figure 4.28 assume that the extragalactic component of the cosmic-ray flux need only describe the flux beyond the "ankle" region at $\sim 10^{19}$ eV (see figure 2.4) and that the injection spectrum in this region has a spectral index of around 2. However, the expected cosmic ray spectrum from supernova remnants (see section 4.3.1) does not reach to the ankle, so there are alternative models which assume that the extragalactic regime starts at $\sim 10^{18}$ eV, at the "dip" feature in figure 2.4, with a much steeper injection spectrum (spectral index $\sim 2.5-2.7$); the form of the spectrum above 10^{18} eV is dictated by e^+e^- pair production, $p+\gamma_{\rm CMB} \rightarrow p+e^+ + e^-$, whose threshold energy is $\sim 10^{18}$ eV, and by the GZK cut-off, $p+\gamma_{\rm CMB} \rightarrow p+\pi^0$ or $n+\pi^+$ (see section 2.2.3), above $\sim 2\times10^{19}$ eV. This "dip model" is successful at describing the shape of the UHE cosmic-ray spectrum, as shown in figure 4.29[451], but cannot be credibly explained by GRBs in the parameter scans of Baerwald et al.[450], because the need to account for a larger proportion of the cosmic ray flux requires unrealistically high baryon loading.

In summary, if extragalactic sources of cosmic rays are required to account for the entire cosmic ray spectrum above $\sim 10^{18}$ eV, then GRBs appear to be ruled out as the sole or principal source. If we require extragalactic sources to account for only the very highest energy cosmic rays, above $\sim 10^{19}$ eV, then they remain a possibility, although the non-observation of associated neutrinos is already constraining the parameter space significantly[449, 450] and has the potential to confirm or refute this model within the next decade.

4.4.3 Radio-loud active galactic nuclei

As can be seen in figures 2.52 and 2.60, non-transient extragalactic sources of GeV and TeV photons are overwhelmingly blazars (BL Lac objects and flat spectrum radio quasars): in the third *Fermi*–LAT AGN catalogue[452], 71% of the detected high-latitude ($|b| > 10°$) γ-ray sources are AGN, and 98% of these are blazars. Any source that emits TeV photons is necessarily accelerating electrons to extremely high energies, although this is not proof that it is a cosmic ray source (an e^+e^- pair plasma with acceleration by magnetic reconnection, which is a viable model for pulsar wind nebulae, see section 4.3.4, need not accelerate any hadrons at all). Nevertheless, the proven existence of UHE electrons in this class of objects surely marks them out as candidate sources for high-energy cosmic rays.

Taxonomy of AGN

"Active galaxies" is a catch-all term for galaxies whose energy emission is not dominated by integrated starlight. Over the course of the 1960s and 70s it became clear that, although the non-stellar emission can originate from a very large region indeed (most obviously in the classical double-lobed radio galaxies, whose emission can stretch over ~ 1 Mpc), it is powered by energetic processes in the very centre of the galaxy. The term "active galactic nucleus" seems to have been coined by the Armenian astrophysicist Viktor Ambartsumian[453] and came into widespread use in the 1970s.

Observationally, the term "active galaxy" or "active galactic nucleus" covers a wide range of apparently disparate phenomena. Most were discovered long before the current picture of nuclear activity caused by an actively accreting supermassive black hole (see below) was in place, and the nomenclature is therefore very unsystematic: categories overlap, and the original definitions may not survive improved observations (for example, BL Lac objects were originally defined as lacking optical emission lines, but more sensitive measurements have since shown that many, including BL Lac itself, do in fact have emission lines, albeit weak).

It should be noted that the nature of the selection criteria applied has a very significant effect on the final AGN sample. Obviously, radio-selected samples will not select the majority of AGN, which are radio quiet, but it is also true that low-excitation radio galaxies (see below) often lack the diagnostic features used by optical surveys, so optical searches for AGN may miss many low-luminosity radio galaxies[455]. Even X-ray selection is compromised by the existence of strong X-ray absorption in many obscured, "Compton-thick", AGN[454]: Brandt and Hasinger[454] point out that well-known local AGN such as NGC 1068 (M77, a type 2 Seyfert at a distance of $\sim$13 Mpc) would only be detectable out to $z \sim 0.1$ in the *Chandra* Deep Fields, because of strong X-ray absorption.

Over the last 25 years, various "unified schemes" for AGN classification have been developed, e.g. [456, 457, 458, 292]. The common theme in these schemes is that the observed morphology of the AGN is strongly dependent on its orientation to the line of sight, so that the distinction between, say, a radio galaxy and a radio-loud quasar is purely an orientation effect and does not reflect a real difference in the parent population. More recently, the initial belief that this might reduce the number of underlying AGN types to precisely two—radio-quiet and radio-loud—has been increasingly questioned[459, 460, 461], and it appears that at least one further criterion—the nature of the accretion on to the black hole—needs to be taken into account.

The basic criteria that are used in classifying AGN are:

Radio loudness

There is a clear distinction between radio-loud AGN, which launch a highly collimated relativistic jet from close to the central engine, and radio-quiet AGN, which do not. The majority of the "classical" AGN (80–95% depending on the exact definition adopted) are radio-quiet. Radio-quiet AGN show no evidence of particle acceleration and, while obviously interesting in their own right, are not relevant to particle astrophysics.

Optical emission lines

The other clear observational distinction is between those AGN which exhibit broad permitted and semi-forbidden lines (corresponding to Doppler velocities of $\mathcal{O}(10000)$ km s^{-1}), known as Type 1 AGN, and those in which the permitted and forbidden lines are both "narrow" (corresponding to Doppler velocities of typically several hundred km s^{-1}, so narrow in this context only!), known as Type 2[462]. In the orientation-based unified schemes, this distinction is not real, but is caused by the presence of an optically thick structure, usually called the "central torus", surrounding the central engine and obscuring the region of broad-line emission when the galaxy is seen from the side. In support of this hypothesis, some (not all) Type 2 AGN are seen to exhibit "hidden" broad lines when viewed in polarised (i.e. scattered) light: light from the obscured broad line region is being scattered into our line of sight by gas clouds further from the central engine.

Radio luminosity and morphology

Fanaroff and Riley[463] divided classical double-lobe radio galaxies into two classes, FR I and FR II, based on their radio luminosity. The difference in radio luminosity is strongly correlated with morphological differences: in the low-luminosity FR I galaxies, the radio emission comes mainly from two jets either side of the central engine and is *edge-darkened*, whereas the high-luminosity FR II galaxies have faint or invisible, usually asymmetric, jets with the radio emission coming mainly from two *edge-brightened* "lobes" of emission at the ends of the jets, typically with radio "hot spots" at the very far edge of the lobe.

Ionisation species of emission lines

A more recent dichotomy is the distinction[464] between *high-excitation* and *low-excitation* Type 2 radio galaxies, based originally on comparing the [OIII] forbidden line at 500.7 nm with Hα. This is strongly but not completely correlated with the FR classification: most (but not all) FR I galaxies are low-excitation galaxies; most FR II galaxies are high-excitation, but a significant minority are not. Type 1 (broad line) radio galaxies are always high-excitation. (This distinction was in fact first noticed by Hine and Longair back in 1979[465], but not used in classification until much later.)

In the classic unified scheme for radio-loud AGN[456, 458], blazars (BL Lac objects and highly-variable flat-spectrum radio quasars) are seen essentially down the jet, and their observed spectrum is dominated by relativstically beamed emission from the jet. As we move away from the jet axis, we see progressively flat-spectrum radio-loud quasars (FSRQs), steep-spectrum radio-loud quasars (SSRQs—these designations relate to the spectral index of the radio emission), broad-line radio galaxies (BLRGs), and narrow-line radio galaxies (see, e.g., figure 5 of Tadhunter[462]); in the last category, the broad line emission region (BLR) is entirely obscured by the dusty torus. This is illustrated in the top half of the left panel of figure 4.30. There is an analogous progression for the more common radio-quiet AGN, with radio-quiet quasars and Type 1 Seyfert galaxies being the analogues of radio-loud quasars and BLRGs, and Type 2 Seyferts the equivalent of NLRGs. In the presumed absence of a jet, there is no equivalent to blazars in the radio-quiet sequence.

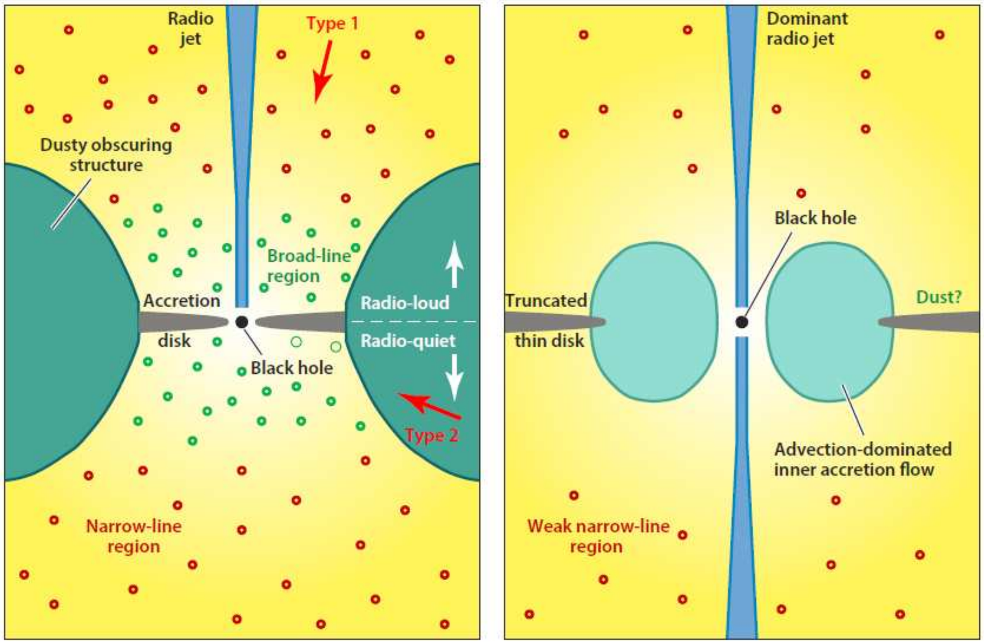


Figure 4.30: Two models of AGN[460]. Left panel, radiative-mode AGN producing high-excitation spectrum, with radiatively efficient accretion from a thin accretion disc and an outer dusty obscuring structure or "torus". Only ∼10% of this class of AGN launch a jet and are radio-loud (top half of panel), the rest being radio-quiet (bottom half). Right panel, jet-mode AGN with radiatively inefficient accretion from a geometrically thick inner accretion flow, no broad line region, and a low-excitation spectrum. This class of AGN always launches jets, and the jet kinetic energy is the dominant mode of energy loss. Figure from Heckman and Best[460].

This simple scheme, which Tadhunter[462] calls the "Perfect Unification Principle", and Antonucci[458] the "Straw Person Model", does not cope at all well with the low-excitation radio galaxies (or, in Tadhunter's nomenclature, weak line radio galaxies, WLRGs). It is very difficult to see how differences in orientation could cause such large differences in emission-line intensity and ionisation state, and essentially impossible to imagine how this could be correlated so strongly with the FR class, since that depends on radio emission taking place at much larger distance scales than those associated with the central torus. As a result, more recent studies of AGN[459, 460] often focus more strongly on the distinction between high- and low-excitation galaxies, arguing that this reflects a fundamental difference in the nature of the central engine, and consequently in the geometry of the innermost regions of the AGN. This picture is summarised in figure 4.30[460], with the left panel representing the high-excitation, “radiative-mode”, AGN typified by quasars and NLRGs, and the right the low-excitation, “jet-mode”, AGN powering BL Lac objects and most FR I radio galaxies.

In this context, it should be noted that the high-luminosity FR II radio galaxies and quasars are *extremely* rare in the local universe. As pointed out by Hardcastle[466], there are *no* FR II radio galaxies within the notional 100 Mpc range of UHE cosmic rays (the closest is 3C 98, at a redshift of 0.0305 and therefore a nominal distance of ∼130 Mpc; Cygnus A is nearly twice as distant, with a redshift of 0.0561 and nominal distance of ∼240 Mpc). Although this is not a hard cut-off, it is clear that if UHE cosmic rays are produced by AGN, the low-luminosity FR I radio galaxies are by far the most likely suspects: van Velzen et al.[467] count 74 within a redshift of 0.03, including the well-known local examples Centaurus A (3.7 Mpc) and M87 (16.6 Mpc).

The AGN central engine

The central engine of active galactic nuclei is universally acknowledged to be accretion on to a supermassive black hole. The basic justification for this is

- the rapid variability of many AGN, on timescales of days to years for most AGN, and less than a day for blazars;
- the high energy output.

Variability timescales Δt essentially set a limit on the size of the source of $R < c\Delta t$ (because otherwise the rise-time of the variation would be “blurred” by the difference in arrival times between photons from the near and far sides of the source), though this can be modified somewhat by anisotropy or by beaming. The high energy output implies a large mass from the Eddington limit (see below); the combination of high mass and small size has long implicated massive black holes as the most likely suspects. This conclusion has been reinforced in recent times by the clear evidence that essentially all large galaxies have central supermassive black holes, with a tight correlation between the mass of the black hole and the mass of the galaxy’s spheroidal component (the whole galaxy in the case of elliptical galaxies, just the bulge for spirals and lenticulars)[468].

The Eddington luminosity

An estimate for the maximum power output from a supermassive black hole can be obtained by assuming that the limiting factor is the outward radiation pressure, which will prevent further accretion if it exceeds the gravitational attraction of the black hole. There is a slight complication in that the radiation pressure acts mostly on electrons, whereas the gravitational force is dominated by protons, but if we assume that the accreted material is mostly hydrogen then there is one electron for every proton, and electromagnetic forces will ensure that they remain together (even if the plasma is completely ionised, you can’t blow away the electrons and expect the protons to continue accreting).

The gravitational force on one proton at distance r from the black hole is simply

$$F_g = \frac{GM_\bullet m_p}{r^2},$$

where $M_\bullet$ is the mass of the black hole.

The number density of photons at this distance is

$$n_\gamma = \frac{L}{4\pi r^2 h\nu},$$

where L is the luminosity and ν is the frequency (this is going to cancel out later, so we don't have to worry about the fact that the luminosity is not monochromatic).

Each photon carries momentum $h\nu/c$, so the radiation pressure per electron at distance r is

$$F_p = \sigma_{\rm T} n_\gamma h\nu/c = \frac{L\sigma_{\rm T}}{4\pi r^2 c},$$

where $\sigma_{\rm T}$ is the Thomson cross-section defined in equation (2.38).

Equating F_g and F_p gives the limiting luminosity

$$L_{\rm Edd} = \frac{4\pi G M_\bullet m_p c}{\sigma {\rm T}} = 1.3 \times 10^{31} \frac{M_\bullet}{M_\odot}\ {\rm W}, \tag{4.40}$$

where $L_{\rm Edd}$ is known as the **Eddington luminosity**.

The accretion rate $\dot{M}_{\rm Edd}$ required for a black hole to achieve its Eddington luminosity is given by

$$\dot{M}_{\rm Edd} = \frac{L_{\rm Edd}}{\xi c^2}$$

where ξ is the efficiency with which accreted mass is converted to radiated energy. The maximum possible value of ξ depends on the spin of the black hole, and varies from 0.06 for a non-rotating (Schwarzschild) black hole to 0.42 for a maximally rotating (Kerr) black hole with favourable accretion geometry. Note that this efficiency far exceeds the 0.007 achieved by nuclear fusion: this is one reason why accretion on to black holes was so quickly seized on as a likely power source for quasars[469].

The Eddington luminosity and associated accretion rate is a useful yardstick in discussing AGN central engines. In particular, many authors (e.g. [459, 460]) attribute the difference between high-excitation and low-excitation AGN to a difference in accretion rate: AGN accreting at $> 0.01\dot{M}_{\rm Edd}$ have thin accretion discs and efficiently convert accreted mass to emitted radiation, whereas AGN accreting at lower rates have low-density, geometrically thick accretion flows which do not radiate effectively: they are advection-dominated accretion flows (ADAFs).

If gas is to be accreted on to a black hole, it must lose angular momentum. Angular momentum is a conserved quantity, so this implies that gas elsewhere in the system must gain angular momentum. This transfer of angular momentum is achieved through the viscosity of the gas. Non-zero viscosity will also heat the gas, by converting ordered motion into random thermal motion; this heat may be radiated away, or it may be advected inwards with the gas. Another effect of viscous dissipation of energy is that the gas will settle into an accretion disc, because a thin disc has less kinetic energy per unit of angular momentum. In other words, a radiatively efficient accretion flow is likely to be accreting cold gas from a thin accretion disc: the timescale for radiative cooling is short enough to allow the heat generated by viscous interactions to be radiated away effectively.

The physics of accretion discs in complicated and extensively studied: there are entire books on the subject[470, 471]. As we are not interested here in the physics of AGN in general, but only on their relevance to particle astrophysics, we shall only scratch the surface of this vast subject: a little more detail is given in Longair[171] sections 14.3, 14.4 and 20.7.

Thin accretion discs

In view of the above discussion of radiatively efficient accretion, it is important to understand the conditions under which we might expect to have a thin accretion disc. This derivation follows Longair section 14.3.

Consider an accretion disc of thickness H at distance r from the black hole. Assuming that the mass of the disc is negligible compared to the mass of the hole, $M_{\text{disc}} \ll M_\bullet$, the equation of hydrostatic equilibrium for a mass element at a height z above the plane of the disc is

$$\frac{\partial p}{\partial z} = -\frac{GM_\bullet \rho \sin\theta}{r^2},$$

where θ is the angle subtended by the height z at the position of the black hole; for a thin disc $\sin\theta \simeq z/r$. If we replace the derivative by a simple division, $\partial p/\partial z \sim p/H$, this expression becomes

$$\frac{p}{H} \simeq \frac{GM_\bullet \rho H}{r^3}.$$

Assume that the inward drift is slow enough that at any given radius r the gas can be considered to be moving in a circular orbit with speed v_ϕ, where from Newton's laws

$$v_\phi^2 = \frac{GM_\bullet}{r};$$

if we substitute this into the previous equation we have

$$\frac{p}{\rho} = v_\phi^2 \frac{H^2}{r^2}.$$

Now the speed of sound in a gas is given by $c_s^2 = \partial p/\partial \rho$, where the derivative is evaluated at constant entropy. Therefore, to order of magnitude we can write $p/\rho \simeq c_s^2$, giving

$$\frac{c_s}{v_\phi} = \frac{H}{r}. \tag{4.41}$$

The accretion disc will be thin ($H \ll r$) if the rotational speed v_ϕ is much greater than the sound speed in the gas. Note that in an ideal gas, $p/\rho = kT/\mu$ where μ is the mean particle mass: the speed of sound in an ideal gas is directly proportional to its temperature.

Radiative-mode and jet-mode AGN

As mentioned above, the observational distinction between high-excitation and low-excitation AGN cannot be explained by invoking differences in orientation, but must reflect a real difference in the central engine. In recent years, it has become generally accepted[459, 460] that there are two quite distinct modes of AGN activity powered by different modes of accretion, as illustrated in figure 4.30. These have been given various names in the literature: I will follow Heckman and Best[460] in referring to the mode underlying high-excitation AGN as "radiative-mode" (it is also known as "standard-mode", "quasar-mode" and "cold-mode") and that underlying low-excitation AGN as "jet-mode" (it is also called "radio-mode"—a particularly unfortunate name as ∼10% of radiative-mode AGN are also radio loud, including most of the highest luminosity radio galaxies—and "hot-mode").

Radiatively-efficient AGN are the classic active galaxies epitomised by quasars (as examples of Type 1, unobscured, AGN) and Type 2 Seyfert galaxies (or, for the radio-loud minority, narrow-line radio galaxies). AGN of this type can be identified optically by their strong emission lines, in the mid-infrared by the thermal emission from the dusty "torus" (it probably isn't a simple torus!) or "obscuring structure" (as per Heckman and Best), or in hard X-ray emission[454]. The radio-loud minority can also be identified by non-thermal radio emission. As noted above, all of these search strategies will typically produce different samples. Optical surveys need to use emission-line diagnostics to separate galaxies containing AGN from pure starburst galaxies: depending on the severity of the cut, this may either accept some non-AGN into the AGN sample or discriminate against genuine AGN in actively star-forming galaxies, while X-ray surveys are effective at rejecting non-AGN but will miss some Compton-thick obscured AGN; IR surveys will presumably miss AGN which do not have a significant obscuring structure; radio surveys will obviously only select the radio-loud subset of radiative-mode AGN (though they are probably the only way of effectively selecting jet-mode AGN, see below). Nevertheless, Heckman and Best[460] comment that differently selected samples of radiative-mode AGN are broadly comparable in their general properties.

The established theoretical model for radiative-mode AGN is shown schematically in the left-hand panel of figure 4.30. It includes the following ingredients[460, 461] (see also Longair[171] Chapter 20):

- a central supermassive black hole ($M_\bullet \gtrsim 10^6 M_\odot$, size scale $\lesssim$ 100 AU);
- accretion of cold gas from an accretion disc that is optically thick but usually, though not necessarily, geometrically thin (size scale < 1 pc);
- a "corona" of very hot gas around the accretion disc, which Compton-scatters optical and UV photons from the disc to produce the observed hard X-ray emission;
- a population of hot, relatively dense ionised gas clouds with a velocity dispersion of several thousand km s^{-1}, created by photoionisation from the accretion disc and corona and responsible for the broad emission lines (size scale $\sim 0.01 - 1$ pc, larger for higher luminosities);
- a dusty structure larger than, but in the plane of, the accretion disc, containing dust grains and molecular gas and having an optical depth sufficient to absorb all photons up to and in some (Compton-thick) cases including hard X-rays (column density[460] $\sim 10^{23} - 10^{25}$ cm^{-2}, size scale 0.1–10 pc depending on luminosity);
- extending in a cone perpendicular to the obscuring structure, clouds of lower-density photoionised gas with a velocity dispersion of a few hundred km s^{-1} producing both permitted and forbidden "narrow" emission lines (size scale hundreds to thousands of parsecs).

In the unified models of AGN[458, 292], Type 2 AGN are being observed at an angle such that the dusty structure obscures our view of the accretion disc and broad-line region, leaving only the narrow emission lines as diagnostics in the optical band. However, Netzer[461] argues that "part of the confusion in the present unification scheme results from the presence of several subgroups that may not belong to it in the first place": in particular, a subset of Type 2 AGN show no broad lines despite a lack of other evidence for obscuration, and some authors interpret these as AGN lacking a broad line region altogether. These are typically less luminous than Type 2 objects with evidence of obscuration, and some authors, e.g. Merloni et al.[472], argue that they are really Type 1 objects whose broad lines are unobserved because the low-luminosity AGN is invisible against the background light from the host galaxy. The fraction of AGN that are classified as Type 2 by optical diagnostics is also strongly dependent on X-ray luminosity[472], with a far higher proportion of low-luminosity AGN classed as optically obscured: this is not the case when the definition of obscured/unobscured is made on the basis of X-ray evidence. This would be consistent with systematic misclassification of low-luminosity AGN as a consequence of dilution of the optical AGN spectrum by galaxy background, as argued by Merloni et al.[472]; it would also be consistent with a failure of low-luminosity radiative-mode AGN to excite a broad line region at all, which seems to be the model favoured by Netzer[461]. Tadhunter[462] shows an example of a “Type 2” NLRG (as originally classified) which proved to have weak but clear broad lines when subsequently investigated at higher signal to noise, thus demonstrating that optical misclassification does occur in some cases.

Only a minority ($\sim$10%) of radiative-mode AGN are radio-loud. In contrast, jet-mode AGN are radio-loud by definition, and are often missed by non-radio AGN surveys because the characteristic emission lines, blue optical continuum (thermal radiation from the accretion disc) and X-ray emission are weak or absent. This suggests that this class of AGN is powered by radiatively inefficient accretion through an optically thin, geometrically thick advection-dominated accretion flow as in the right-hand panel of figure 4.30. In this model, the bulk of the energy emitted from the central black hole is in the mechanical energy of the jet, rather than in electromagnetic radiation. Jet-mode AGN are difficult to survey accurately, because correlating radio and optical catalogues is non-trivial: the fact that the radio emission from a radio galaxy is typically extended means that radio surveys with good angular resolution, such as the FIRST survey[473] tend to catalogue a single radio galaxy as multiple sources and may also “resolve out” faint extended emission (i.e. the intensity per pixel is too low to register, even though the overall flux is significant), whereas radio surveys with poorer angular resolution such as the NVSS[474] measure fluxes more precisely and generally identify a radio galaxy as a single source, but the reported position may not match the optical position (if the radio source is asymmetric, as many are) or may be too imprecise to yield a unique optical counterpart. The net result of this is that automatic cross-matching of radio and optical surveys is quite challenging (visual identification is less so, but is not practical when dealing with modern very large catalogues: even the most acquiescent graduate student would baulk at being asked to comb through the entire SDSS catalogue to cross-match with FIRST...). However, correlation techniques which combine NVSS and FIRST with “collapsing algorithms” designed to identify and combine multiple radio components from the same source, are now able to provide optical-radio cross-matched samples with $\sim$95% completeness and $\sim$99% reliability[460].

There *are* radio-quiet sources with low-ionisation emission lines: the so-called LINERs (for Low Ionisation Nuclear Emission-line Region). The nature of these objects is not entirely clear: they are probably a fairly heterogeneous group, and not all contain an actively accreting supermassive black hole. A significant body of opinion[475] holds that the characteristic LINER emission emanates from old stars and is not dependent on the existence of an active AGN. Perhaps fortunately, we do not need to consider this class of galaxy further, since radio-quiet galaxies do not show evidence for significant particle acceleration (above and beyond that produced by supernova remnants and pulsar wind nebulae as discussed in the previous section) and are hence not relevant to particle astrophysics.

Given that we know[468] that supermassive black holes (SMBH) are a ubiquitous feature of large galaxies, the questions that arise from the above summary are:

- What causes some SMBH to power AGN?
- What determines whether a given AGN operates in the radiative mode or the jet mode?
- Why are some radiative-mode AGN radio-loud, when the majority are radio-quiet?

In (perhaps partial) answer to the first question, only SMBH that are actively accreting material can possibly power an AGN: an SMBH that is sitting quietly surrounded by material in stable orbits is not going to generate any energy. Active accretion requires a gas supply—unlike gas, which can transfer angular momentum by viscous interactions as discussed above, stars are essentially a collisionless gas and have no obvious way to lose enough angular momentum to fall into the SMBH. Therefore, it is not unreasonable to find that most galaxies, despite containing a central SMBH, do not display significant AGN activity.

AGN demographics

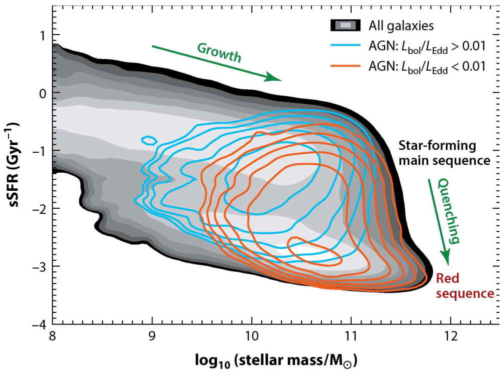


Figure 4.31: Demographics of AGN[460]. The greyscale shading shows the galaxies in the SDSS main galaxy sample: the two regions of higher concentration (lighter shading) represent the "blue cloud" of star-forming galaxies (top) and the "red sequence" of elliptical galaxies (bottom). The blue and red contours show the loci of AGN with high and low accretion rates respectively; as shown later, these correspond to radiative-mode and jet-mode AGN respectively. The contours are logarithmic: each contour represents a factor of 2. The y-axis is the log to base 10 of the specific star-formation rate (star-formation rate in solar masses per gigayear, divided by galaxy mass in solar masses); the log has been missed off the axis label. Figure from Heckman and Best[460].

Figure 4.31[460] shows the local galaxy population from the Sloan Digital Sky Survey (greyscale) and AGN (contours). The galaxy population shows two distinct concentrations, one at high specific star formation rate and one at low: these represent the "blue cloud" of star-forming galaxies and the "red sequence" of elliptical galaxies respectively. The former is called a cloud, and the latter a sequence, because they evolve differently with redshift: star-formation rates decrease from about redshift 2 to the present day, so if this were a galaxy sample with greater depth in redshift, the red sequence would stay more or less unchanged, but the blue cloud would extend upward to higher star formation rates, making it broader and less well-defined. The less well-populated region between the blue cloud and the red sequence is colloquially known as the "green valley": these galaxies are not, of course, really green—they just lie in between blue and red. The superimposed contour plots show the population of AGN in the SDSS, separated into high (blue) and low (red) accretion rates. It is clear that these two classes correspond broadly to intermediate-mass blue and "green" galaxies and high-mass red galaxies, respectively. This makes logical sense: red galaxies have low specific star-formation rates, which implies a lack of cold gas to form stars, which in turn implies a lack of cold gas to fuel an AGN.

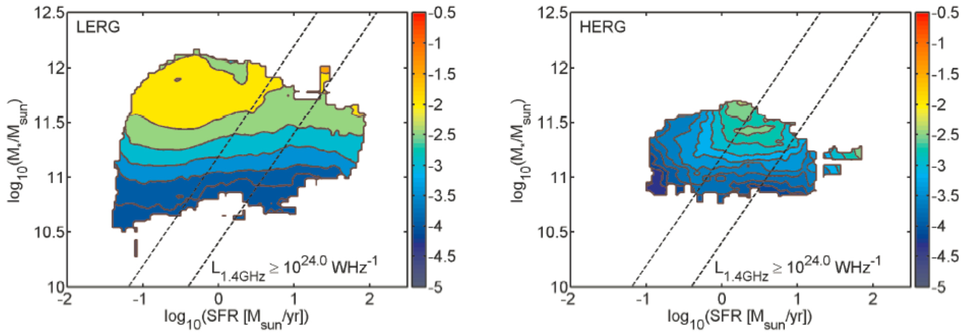


Figure 4.32: Demographics of radio galaxies[476]. The fraction of SDSS galaxies that are low-excitation (left) or high-excitation (right) radio galaxies, as a function of the galaxy's mass and total star formation rate. The two diagonal lines mark the approximate locations of the transitions between the red sequence (left), the green valley (centre) and the blue cloud (right). Figure from Janssen et al.[476].

A similar picture, also based on the SDSS, is presented by Janssen et al.[476] and shown in figure 4.32. Low-excitation radio galaxies (LERGs) are predominantly located in very massive galaxies, with little evidence for a dependence on star-formation rate (the lack of LERGs to the right of the "blue cloud" dividing line seems to stem simply from a lack of very massive star-forming galaxies), whereas high-excitation radio galaxies (HERGs) prefer more active star formation and therefore a somewhat lower galaxy mass—the contours clearly fall off from right to left as well as from top to bottom. This is natural if radiative-mode AGN require a supply of cold gas to form an accretion disc, whereas jet-mode AGN can be fuelled by accreting hot gas from the hot interstellar medium of an elliptical galaxy.

The correspondence between accretion rate and excitation is shown explicitly by Best and Heckman[459]: see figure 4.33. The galaxies that they classify as HERGs consistently have an energy output, expressed as a fraction of their calculated Eddington luminosity, that is an order of magnitude higher than those they class as LERGs. An essentially identical distribution, with a somewhat higher overall normalisation (HERGs peaking at ∼20% instead of ∼3%) is found by Mingo et al.[477]. In both cases there is some overlap in the distributions, so the choice of radiative-mode or jet-mode activity is not made on the basis of accretion rate alone: some other factor(s) must be involved. However, the separation is noticeably cleaner than one based on radio luminosity alone: the left panel of figure 4.33 shows that, although the fraction of HERGs clearly increases rapidly with increasing luminosity, both modes are present over nearly the whole available range, with HERGs definitively dominant only at radio powers above 10^{26} W Hz^{-1} (as measured by the NVSS).

The global picture from studies such as these is therefore[460] that jet-mode AGN occur preferentially in very massive galaxies, usually with low star-formation rates and hence presumably low reserves of cold gas. The AGN is

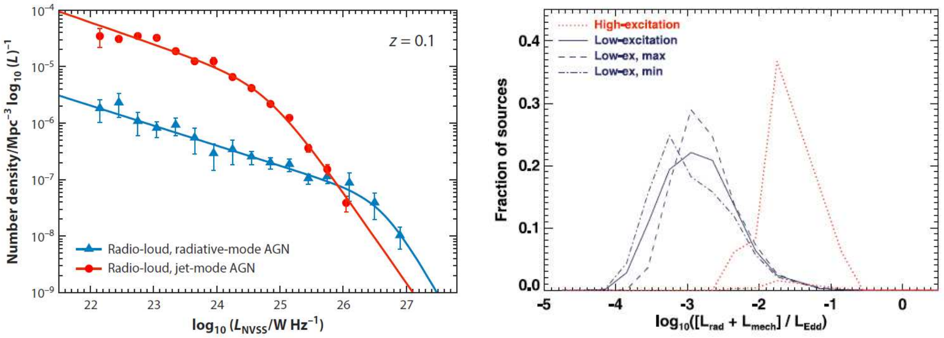


Figure 4.33: Luminosity and accretion rate of radio galaxies. Left panel, radio luminosity function of radiative- and jet-mode AGN in the local universe, from Heckman and Best[460]. Radio-loud radiative-mode AGN (HERGs) tend to be more luminous than jet-mode AGN (LERGs), but both types are seen across (almost) the whole range of possible luminosities. Right panel, energy output as a fraction of the Eddington luminosity, from Best and Heckman[459]. Assuming that the energy output tracks the accretion rate, HERGs are typically accreting at 1% to 10% of the Eddington rate, LERGs at <1%.

a massive black hole, typically $> 10^7\,M_\odot$, accreting at a low rate, typically $\lesssim 0.01 L_{\rm Edd}$ via a geometrically thick, optically thin advection-dominated accretion flow. Radio-loud radiative-mode AGN occur in galaxies with higher star-formation rates, i.e. with available reserves of cold gas; the host galaxies are typically less massive, but this may be a selection effect caused by the lack of very massive star-forming galaxies. The AGN is a lower-mass black hole, typically $10^6 - 10^7\,M_\odot$, accreting at $\mathcal{O}(10\%)$ of its Eddington rate via an optically thick, usually geometrically thin accretion disc.

There still remains the question of why a minority of radiatively efficient AGN launch jets and are radio loud, when the majority do not. There are many properties of the AGN, its host galaxy, and its environment which could influence this, and most of them have been implicated at some point. A popular conjecture is that a high black hole spin rate is necessary to launch a jet: this may be true, but very high spin rates have been measured, using X-ray reflection, in radio-quiet AGN[478], and it is therefore clear that high spin rates are not *sufficient* to provoke radio-loud behaviour. Tadhunter et al.[479] find that HERG host galaxies have higher dust masses than typical elliptical galaxies, implying a larger reservoir of cool gas; in earlier work this group also found evidence of tidal tails and similar merger/interaction diagnostics, and they therefore suggest that the gas may be delivered, and the radio activity triggered, by such interactions; however, again they find that the majority of interacting elliptical galaxies do *not* host radio-loud AGN, so this is not a sufficient condition. A minority of double-lobed radio galaxies have "double-double" structures with distinct inner and outer lobes, suggesting individual episodes of jet formation separated by quiescent periods in which the jets are absent: this switching between different accretion modes is well attested in X-ray binaries[480].

Interestingly, the very similar (though obviously on a much smaller scale!) radio jets emitted by black hole X-ray binaries are associated with the so-called "low-hard" state in which overall luminosity is low and the X-ray spectrum hard. In this state there is no evidence for an optically thick accretion disc, whereas during occasional outbursts (to the "high-soft" state) the luminosity becomes dominated by thermal radiation from an optically thick accretion disc[481]. The similarity between this and jet-mode/radiative-mode AGN seems too close to be accidental, though the observed differences between radio-quiet and radio-loud radiative-mode AGN in respect of host galaxies and environment suggest that it is not *simply* a case of an AGN "duty cycle" in which all radiative-mode AGN spend ~10% of their time in a radio-loud state. Overall, this question remains undecided and an active subject of research.

Launching the radio jet

It is clear from the observed morphologies of both FR I and FR II radio galaxies that the radio emission comes from relativistic jets emitted from close to the black hole. The fact that the radio emission is synchrotron radiation makes it clear that magnetic fields are involved, and the general assumption is that the jets are magnetically launched and collimated.

The magnetic field of the host galaxy is trapped in the plasma around the black hole and advected inwards. The pressure due to the magnetic field is

$$p_{\rm mag} \simeq \frac{B^2}{2\mu_0}$$

(the exact relation will depend on the exact form of the field), and this opposes the gravitational force

$$F_g = \frac{GM_\bullet \Sigma}{R^2}$$

where Σ is the surface density of the accretion disc. By conservation of mass, in a steady state the surface density of the disc at any radius r is given by

$$\dot{M} = 2\pi r v_r \Sigma,$$

where v_r is the inward velocity at radius r and $\dot{M}$ is the accretion rate—in other words, mass flows steadily through the accretion disc towards the black hole, and the disc itself neither gains nor loses mass.

Accretion from the disc will be halted if the magnetic pressure exceeds the gravitational attraction. This will occur when

$$\frac{B^2}{2\mu_0} = \frac{GM_\bullet \dot{M}}{2\pi R^3 v_R}.$$

For convenient application to AGN we write

$$\begin{array}{ll} \dot{M} = \dot{m}\dot{M}_{\rm Edd} & \text{where } \dot{M}_{\rm Edd} = \alpha M_\bullet \\ R = rR_{\rm S} & \text{where } R_{\rm S} = 2GM_\bullet/c^2 \\ v_R = \epsilon v_{\rm ff} & \text{where } v_{\rm ff} = \sqrt{2GM_\bullet/R} = c/\sqrt{r}, \end{array}$$

where $\dot{M}_{\rm Edd} = L_{\rm Edd}/c^2$ and in SI units (i.e. with $M_\bullet$ in kg) $\alpha = 7.3 \times 10^{-17}\,{\rm s}^{-1}$. Substituting these into the equation for B^2 gives

$$B^2 = \frac{\mu_0 \alpha c^5}{8G^2\pi} \frac{\dot{m}}{\epsilon r^{5/2} M_\bullet}$$

If we now evaluate the numerical factors, we get

$$B_{\rm max} \sim 3 \times 10^4 \text{ T } \times \frac{\dot{m}^{1/2}}{\epsilon^{1/2} M_\bullet^{1/2} r^{5/4}},$$

where $M_\bullet$ is measured in solar masses. This condition produces a **magnetically arrested disc** or MAD[482]. The radius of the magnetosphere can be found by integrating B to find the total magnetic flux Φ:

$$\Phi(r_{\mathrm{m}}) = B_0\sqrt{\frac{\dot{m}}{\epsilon M_\bullet}}\int_0^{R_{\mathrm{m}}} 2\pi R r^{-5/4}\mathrm{d}R = B_0\sqrt{\frac{\dot{m}}{\epsilon M_\bullet}}\left(\frac{2GM_\bullet}{c^2}\right)^2 2\pi\int_0^{r_{\mathrm{m}}} r^{-1/4}\mathrm{d}r,$$

where $B_0 = 3\times 10^4$ T and in the last step we change variables from the physical radius R to the dimensionless parameter r. The integral yields $\frac{4}{3}r_{\mathrm{m}}^{3/4}$, and evaluating the numerical coefficients gives

$$r_{\mathrm{m}} \sim 3\times 10^{27}\Phi^{4/3}\epsilon^{2/3}\dot{m}^{-2/3}M_\bullet^{-2},$$

where $M_\bullet$ is expressed in solar masses and Φ in T pc^2. With typical values of $M_\bullet = 10^8 M_\odot$, $\dot{m} = 0.01$, $\epsilon = 0.01$ and $\Phi = 10^{-5}$ T pc^2 [482], this gives $r_{\mathrm{m}} \sim 66000$, so the accretion disc should be disrupted at quite a large radius.

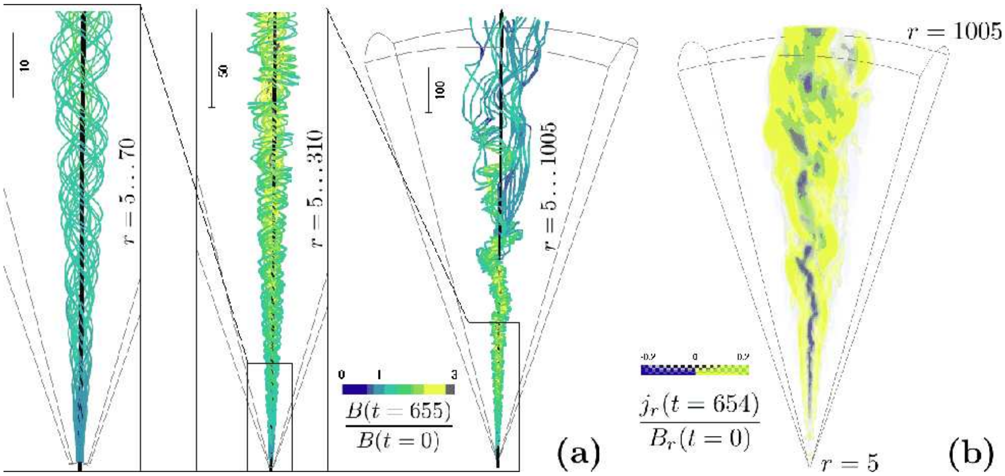


Figure 4.34: 3D simulation of a magnetically-driven jet[484]. From left to right, the first three panels (a) show magnetic field lines on increasing length scales, while the right-hand panel (b) shows the current density ($\nabla\times\mathbf{B}$) corresponding to the last set of field lines. Note that this simulation is non-relativistic, whereas real AGN jets have Lorentz factors of a few, so the details may not correspond to reality. Picture from Moll[484].

Some of the magnetic flux advected inwards will actually thread the black hole event horizon (electric charge, along with mass and spin, is one of the three properties a black hole is allowed to have). By applying the expression for the magnetic flux, $\Phi \propto M_\bullet^{1/4}\dot{M}^{1/2}R^{3/4}$ at R_{S} and using $R_{\mathrm{S}} = 2GM_\bullet/c^2$ to eliminate M, we find that

$$\Phi_{\mathrm{BH}} = \phi(\dot{M}cR_{\mathrm{S}^2})^{1/2}, \qquad (4.42)$$

where ϕ is a dimensionless constant which simulations indicate is of order 25[483][10]. If the maximum flux advected on to the black hole is greater than this, then jets are produced by the Blandford-Znajek mechanism[420] as discussed on page 212 in the context of GRBs. The spin of the black hole twists the trapped magnetic field into a helical configuration, which will initially collimate the jet, although 3D simulations[484] suggest that instabilities disrupt the

[10] Sikora and Begelman[483] say $\phi \sim 50$, but they are using $r_g = GM/c^2$ as their parameter instead of R_{S}.

helical structure at larger distances, as shown in figure 4.34. The Blandford-Znajek mechanism launches an electromagnetic jet, but FR I jets in particular probably entrain baryonic material (i.e. gas) at the jet boundary (**external entrainment**) and from blowing material off stars unfortunate enough to be caught in the jet (**internal entrainment**)[485]. Many models therefore envisage a layered jet with an inner "spine" of pair plasma and an outer "sheath" of baryonic matter[486]. There is some observational evidence for such structures, notably polarisation structure in radio jets where the polarisation is transverse to the projected magnetic field in the centre of the jet but along the magnetic field at the edges[487].

The magnetic flux threading parsec-scale jets, which is a measure of Φ_{BH}, can be inferred from the **core-shift effect**[488]: the position of the "optically"-thick (actually, radio-emission-thick) core of the jet is defined by the location at which the optical depth $\tau = 1$. This can be shown to yield a 3D offset between the central engine and the observed jet core of

$$r = (B^{k_b} F/\nu)^{1/k_r}$$

where B is the magnetic field at 1 pc, F is a calculable constant dependent on various source properties, $k_b = (3 - 2\alpha)/(5 - 2\alpha)$ where α is the synchrotron spectral index, and k_r is calculable with some model dependence, but can be checked from observations if more than two frequencies ν are measured. The measured core position offsets, which are the projection of r on to the plane of the sky, can then be used to infer B. This has recently been done for a sample of 76 radio-loud (mostly) radiative-mode AGN[489]. The results are in good agreement with equation (4.42), with the fitted parameter $2\phi = (52 \pm 5)\Gamma\theta_j$, where Γ is the bulk Lorentz factor and θ_j is the jet opening angle; the value of $\Gamma\theta_j$ at the jet point of origin is expected to be ~ 1 according to model calculations, which would give a value of ϕ in excellent agreement with the expectation from MAD simulations.

4.4.4 Particle acceleration in AGN

Magnetic confinement

Radio-loud AGN have always been a favoured potential source for the highest energy cosmic rays[490], because the high radio luminosity shows that particle acceleration is indeed taking place, if only of electrons and not necessarily to EeV energies, and the magnetic fields inferred from equipartition or minimum-energy arguments are sufficient to confine particles of the required energies within the radio lobes (see equation (4.1) and figure 4.1). Following Hardcastle[466], we can write the electron energy distribution as a power law, $N(E_e) = N_0 E_e^{-\delta}$, and assume that the high-energy electron population and the magnetic field are a factor of ζ away from equipartition, so that

$$U_e = N_0 \int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} E_e E_e^{-\delta} \mathrm{d}E_e = N_0 I = \zeta \frac{B^2}{2\mu_0}. \tag{4.43}$$

In this case the integral I can be done numerically:

$$I = \int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} E_e^{1-\delta} \mathrm{d}E_e = \begin{cases} \ln\left(E_{\mathrm{max}}/E_{\mathrm{min}}\right) & \delta = 2 \\ \frac{1}{2-\delta}\left[E_{\mathrm{max}}^{2-\delta} - E_{\mathrm{min}}^{2-\delta}\right] & \delta \neq 2 \end{cases} \tag{4.44}$$

In section 2.3.5, we derived the functional dependence of the synchrotron-radiation emissivity j_ν on ν and B,

$$j_\nu = C(\delta) N_0 B^{(\delta+1)/2} \nu^{-(\delta-1)/2} \tag{4.45}$$

(see equation (2.48), and the full analysis in Longair[171] section 8.5.2 gives the coefficient of proportionality as

$$C(\delta) = \frac{\sqrt{3}e^3}{8\pi\epsilon_0 cm_e(\delta-1)}\left(\frac{2\pi m_e^2 c^4}{3e}\right)^{-(\delta-1)/2}\frac{\sqrt{\pi}\,\Gamma\left(\frac{\delta}{4}+\frac{19}{12}\right)\Gamma\left(\frac{\delta}{4}-\frac{1}{12}\right)\Gamma\left(\frac{\delta}{4}+\frac{5}{4}\right)}{\Gamma\left(\frac{\delta}{4}+\frac{7}{4}\right)} \tag{4.46}$$

if the pitch angle of the electrons with respect to the magnetic field is isotropic. Combining equations (4.43) and (4.45) to eliminate N_0 gives

$$j_\nu = \frac{C(\delta)\zeta}{2I\mu_0}\nu^{-(\delta-1)/2}B^{(\delta+5)/2}. \tag{4.47}$$

In order to confine cosmic rays of charge Z and energy $E_{\rm CR}$ in the region in question, its size must satisfy the Hillas condition

$$R > \frac{E_{\rm CR}}{ZeBc} \tag{4.48}$$

(see also equations (3.38) and (4.1)). We can use this to eliminate B, yielding the inequality

$$j_\nu = \frac{C(\delta)\zeta}{2I\mu_0}\nu^{-(\delta-1)/2}\left(\frac{E_{\rm CR}}{ZeRc}\right)^{(\delta+5)/2}, \tag{4.49}$$

and finally multiply by $\frac{4}{3}\pi R^3$ to obtain the total luminosity

$$L_\nu = \frac{2\pi C(\delta)\zeta}{3I\mu_0}\nu^{-(\delta-1)/2}\left(\frac{E_{\rm CR}}{Zec}\right)^{(\delta+5)/2}R^{-(\delta-1)/2}. \tag{4.50}$$

This equation can be used, with suitable assumptions about $E_{\rm min}$, $E_{\rm max}$, R, ζ and δ, to decide which radio galaxies are capable of accelerating cosmic rays to the maximum energies observed. Note that equation (4.50) depends on the cosmic-ray rigidity, $E_{\rm CR}/Ze$ (see equation (2.2); strictly, rigidity is defined in terms of cp rather than E, but as we are dealing here with the high-energy tail of the cosmic ray spectrum it is reasonable to assume that all species are ultra-relativistic, $E \gg mc^2$), rather than the energy itself, because it is rigidity and not energy that determines how a charged particle responds to a magnetic field.

If $\delta = 2$ as expected in diffusive shock acceleration, the dependence of equation (4.50) on the details of the electron spectrum is only logarithmic, and the dependence on the size of the source is also quite weak ($L \propto R^{-1/2}$). The principal determinant of the required luminosity is the assumed maximum rigidity: $L \propto (E_{\rm CR}/Ze)^{7/2}$. Putting in numbers, we have (for a single lobe)

$$L(408\text{ MHz}) > 2.0\times 10^{24}\zeta\left(\frac{E_{20}}{Z}\right)^{7/2} r_{100}^{-1/2}\text{ W Hz}^{-1}, \tag{4.51}$$

where $E_{20} = E_{\rm CR}/10^{20}$ eV, and $r_{100} = R/100$ kpc. If we take $r_{100} = 2.5$ (250 kpc being a reasonable maximum size scale for one lobe of a large radio galaxy), $E_{20} = 1$, $Z = 1$ (i.e., protons) and $\zeta = 1$ (strict equipartition), and multiplying by 2 since classical radio galaxies have two lobes, we get a monochromatic luminosity at 408 MHz of 2.5×10^{24} W Hz^{-1}, only about a factor of 10 less than the FR I/FR II break at this frequency: that is, only the most luminous FR I radio galaxies are likely to be able to contribute to the flux of 10^{20} eV cosmic rays if such cosmic rays are protons[466].

The numerical factor in equation (4.51) depends on assumptions about the electron distribution, specifically the minimum energy and spectral index of the power law spectrum (the dependence on the maximum energy is weak). In the canonical case of $\delta = 2$ the variation with E_{min} is only logarithmic, but equation (4.44) shows that it becomes more severe as δ moves away from 2. However, as we saw in chapter 3, most acceleration models produce δ values quite close to 2, so this uncertainty is probably not very significant. In general, a steeper spectrum reduces the minimum luminosity: this is because the equipartition argument relates the magnetic field strength to the *total* electron energy density (dominated by low-energy electrons), whereas the synchrotron luminosity depends primarily on *high* energy electrons (of which there will be fewer, for a given total electron energy density, if the power law is steeper).

Note that, as pointed out by Hardcastle[466], if the radio lobes can confine UHE cosmic rays, then UHE cosmic rays produced elsewhere in the AGN may still *appear* to emanate from the lobes, because they are likely to propagate into the lobes and be confined there (and perhaps accelerated further) for some time before escaping.

Possible sites of particle acceleration

There is no doubt that blazars accelerate electrons to energies $\gg 1$ TeV, because the overwhelming majority of extragalactic TeV and GeV γ-ray sources are blazars[257, 452]. The fact that blazars, where we are looking more-or-less directly into the jet, so dominate the high-energy γ-ray source catalogue suggests that the production is correlated with the jet: an isotropic mechanism, say associated with the immediate vicinity of the central engine, would not require alignment of the jet axis and the line of sight. However, as the jets are relativistic, this may be partly a selection effect: any radiation emitted by material moving with the jet is strongly Doppler boosted, and, as we saw for GRBs, relativistic beaming reduces the cross-section for pair production and therefore increases the chances that high-energy photons will escape. Of course, because of the deflection of charged particles by magnetic fields, a nearby AGN could contribute significantly to the observed flux of UHE cosmic rays even if its radio jets are not directed close to our line of sight: the magnetic fields of radio jets are expected to be very complicated (see figure 4.34) and could certainly eject charged cosmic rays at large angles.

There are a number of possible sites in a typical radio-loud AGN where hadrons could be accelerated to high energies.

Close to the central engine (sub-parsec scales):

The radio jets are launched very close to the central supermassive black hole: VLBI interferometry[491] derives a size of 5.5 ± 0.4 Schwarzschild radii for the base of the jet in M87. Particle acceleration could also take place on this scale: the formation of the helical magnetic fields believed to account for jet collimation would presumably offer appropriate geometries for acceleration by magnetic reconnection. The TeV γ-rays from Centaurus A observed by H.E.S.S.[492] appear to come from the core, although the resolution of the observation is not good enough to exclude the inner jet. Kachelrieß, Ostapchenko and Tomàs[493, 494] demonstrate that TeV photons produced by protons accelerated in the core of Cen A could escape (despite the high level of background photons inducing pair production), but argue that an unexpectedly large magnetic field (about a factor of 10 higher than predicted by equipartition arguments) would be required to reach energies of $\sim 10^{20}$ eV.

At shocks in the inner radio jets (0.1–1 kpc scales):

The jets of FR II galaxies, and the inner jets of FR I galaxies, have a "knotty" or "lumpy" appearance that is usually assumed to be caused by the presence of shocks. This provides a natural environment for diffusive shock acceleration. These knots are bright X-ray sources (see figure 4.35), and the X-ray emission appears to be synchrotron radiation[495]. Since the frequency of synchrotron radiation and the rate of energy loss from synchrotron radiation both depend on E_e^2 (see equations (2.45) and (2.37)), the population of electrons respeonsible for X-ray synchrotron emission must have been accelerated very close to the origin of the X-rays: there simply is not time for them to have propagated in from elsewhere. Thus, the inner jets of FR I radio galaxies are definitely sites of particle acceleration; if it is diffusive shock acceleration, and if the jets contain protons, we would expect that protons would be accelerated (although the X-ray synchrotron emission is *prima facie* evidence only for electron acceleration). FR I jets appear to be relativistic only close to the central engine: they entrain material and decelerate as they move outwards. Hence, even if the jet is originally a pair plasma as expected if it is launched by the Blandford-Znajek mechanism, it will subsequently entrain proton-rich gas to provide the seed material for cosmic rays.

In the extended lobes ($\mathcal{O}(100)$ kpc scales):
We saw above that the extended lobes of FR II and bright FR I radio galaxies are capable of magnetically confining UHE cosmic rays. There is distributed X-ray[497] and GeV γ-ray[496] emission in the giant lobes of Centaurus A, suggesting that particle acceleration is taking place there; as in the case of the inner jets, the X-ray emission is best fitted by a synchrotron origin, indicating that the acceleration is taking place locally. The lobes are likely to be full of rapidly moving, disorganised magnetic field, so second-order Fermi acceleration (stochastic acceleration), which we dismissed in Chapter 3 as too slow to do anything useful, may make an important contribution here[317].

In the hotspots at the ends of FR II lobes ($\mathcal{O}(100)$ kpc scales):
The hotspots at the ends of FR II lobes are interpreted as jet termination shocks, i.e. they are produced when the highly supersonic jet is stopped by the intergalactic medium. Unlike jets in FR I galaxies, FR II jets remain fast-moving and collimated out to very large distances ($\mathcal{O}(100)$ kpc, see for example Cygnus A). As with the inner jets of FR I galaxies, the hotspots produce optical and X-ray synchrotron radiation, implying local electron acceleration; unlike the FR I jets, the magnetic fields have been directly measured in some cases by comparing inverse Compton luminosity ($\propto U_{\rm rad}\beta^2\gamma^2$, see equation (2.67)) with synchrotron ($\propto U_{\rm mag}\beta^2\gamma^2$, see equation (2.40))[466], and the results indicate that UHE cosmic rays could be successfully confined. However, it should be noted that FR II galaxies are rare in the local universe: although they may be responsible for the acceleration of cosmic rays with energies of order 10^{20} eV in the Universe at large, they are unlikely to contribute many such particles to the cosmic ray flux *observed on Earth*, because of the GZK limit.

It is quite possible, indeed likely, that particle acceleration occurs in multiple sites in the same AGN. Our direct evidence—X-ray synchrotron radiation—relates to the acceleration of leptons in the inner jets of FR I radio galaxies and the outer hotspots of FR II radio galaxies, but distributed emission in the outer

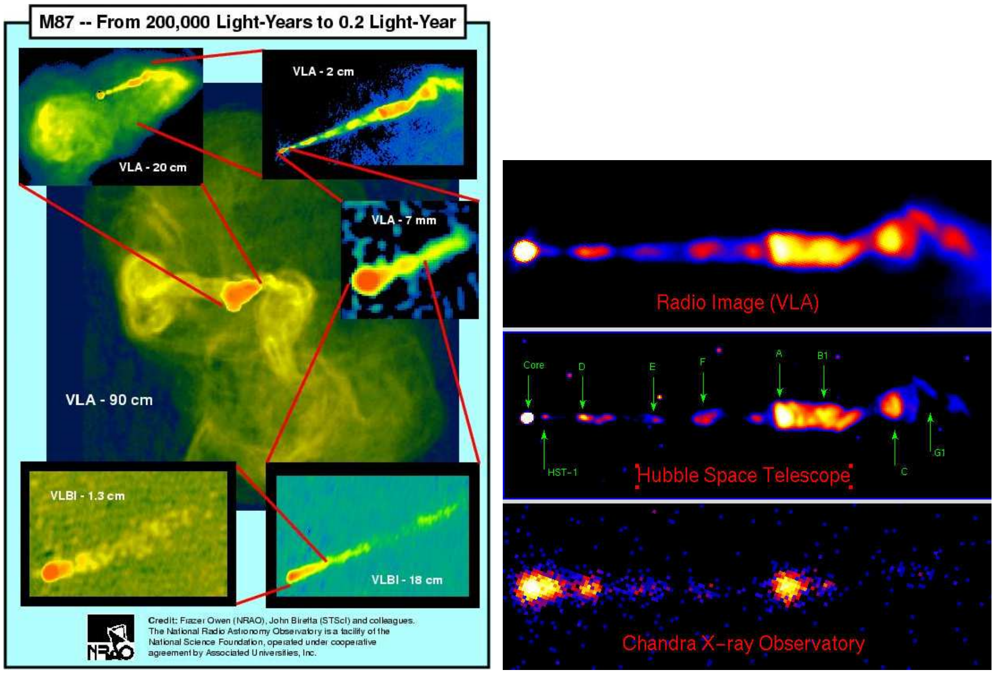

M87 -- From 200,000 Light-Years to 0.2 Light-Year
VLA - 2 cm
VLA - 20 cm
VLA - 7 mm
VLA - 90 cm
VLBI - 1.3 cm
VLBI - 18 cm
Credit: Frazer Owen (NRAO), John Biretta (STScI) and colleagues. The National Radio Astronomy Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc.
NRAO
Radio Image (VLA)
Core
D
E
F
A
B1
HST-1
Hubble Space Telescope
C
G1
Chandra X-ray Observatory

Figure 4.35: The nearby AGN M87. M87 is optically an apparently normal giant elliptical galaxy, apart from a faint jet extending out of the nucleus. In radio, it is an extremely complex object with structure on many scales (left panel). The jet (right panel) can be observed at wavelengths from X-ray to radio, with corresponding structures visible at all wavelengths, although some are much brighter in some wavebands than in others. Left panel Frazer Owen (NRAO), John Biretta (STScI) and colleagues[498]. Right panel from *Chandra* website[499]: radio image (top) F Owen, F Zhou, J Biretta; HST optical (middle) E Perlman et al., NASA/STScI/UMBC; X-ray (bottom) H Marshall et al., NASA/CXC/MIT.

lobes of nearby FR I AGN suggests that acceleration also takes place there. As our analysis of magnetic confinement suggests that UHE cosmic rays could be successfully confined by the outer lobes of the more luminous FR I radio galaxies, it is quite possible that the *same population* of high-energy protons might be accelerated at more than one site: protons accelerated by diffusive shock acceleration in the jets could subsequently enter the lobes and be further accelerated there before escaping. It is not necessary to posit a simple one-site model covering the whole energy range from thermal to 10^{20} eV.

Acceleration mechanisms

As with sites, so with mechanisms: essentially every mechanism described in chapter 3 can be accommodated in AGN. The association of X-ray emission from synchrotron radiation with features plausibly interpreted as shocks suggests diffusive shock acceleration; some of these shocks are relativistic, as demonstrated by the apparent superluminal motion of “knots” in AGN jets. This is an optical illusion caused when the jet axis is close to the line of sight (see the last problem in chapter 2), and though the required velocity is not actually superluminal it *is* relativistic: an observed jet speed of κc requires an actual velocity such that the Lorentz factor $\Gamma \simeq \kappa$. Values of κ up to around 50 are observed, but the typical jet Lorentz factor is likely to be much less than this. Because of relativistic beaming, jets with high Lorentz factors will appear brighter than slower jets, and will therefore be over-represented in flux-limited surveys.

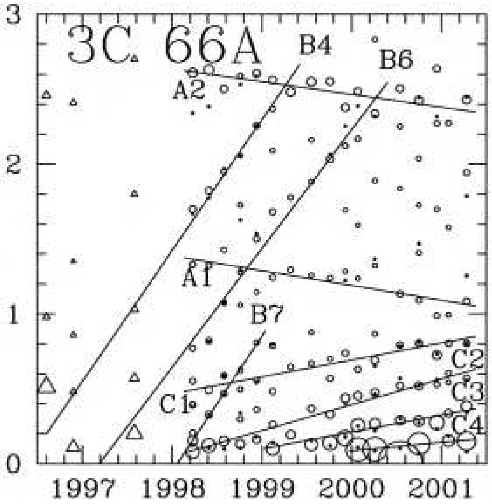


Figure 4.36: Motion of knots in the jet of blazar 3C 66A, as observed with the VLBA[500]. The y axis shows distance from the core in units of milliarcseconds (mas). This object has a redshift of 0.444 and an angular-diameter distance of ~1.1 Gpc[501]: the C components of the jet are mildly superluminal and the B components highly superluminal. The apparent reverse motion of the A components may just reflect a change in their brightness distributions over the observing period rather than motion of the entire feature.

Not all knots in AGN jets are superluminal: some appear stationary or move very slowly. These may represent reverse shocks: as we saw when considering supernova remnants, outward-moving ejecta can generate both forward and reverse shocks when colliding with ambient media or overtaking slower ejecta. A particularly complicated case, the BL Lac object 3C 66A, is shown in figure 4.36[500]: this source appears to have three distinct families of knots, the B series being highly superluminal with apparent speeds of 22–27c, the C series much slower (1.5–5c), and the A series apparently moving back towards the core, although Jorstad et al.[500] are inclined to interpret this as an artefact caused by changes in their brightness distributions shifting the position of the centroid. They suggest that the B components (which are weak) represent forward shocks and the C components reverse shocks. For radio galaxies in general, it is assumed that the plasma of the jet proper passes through the shocks, and that the fastest-moving knot therefore gives the lower limit for the speed of the jet material.

The brightness of knots can vary on rapid timescales: Jorstad et al.[502] report a multiwavelength study of the blazar 3C 454.3, a flat-spectrum radio quasar at a redshift of 0.859, which undergoes very fast γ-ray outbursts with rise-times as short as 3–5 hours (for a flux change of at least a factor of 2). The flares are seen at all wavebands from *Fermi*–LAT γ-rays (0.1–200 GeV) through X-ray and optical to 1.3 mm, with essentially no time-lag (<1 day) between wavelengths. These outbursts are associated with changes in the polarisation of superluminal knots in 3C 454.3's parsec-scale jet and appear to occur when knots pass through the VLBI "core" of the radio jet, which is located about 15–20 pc from the central engine[502]. The spectrum of the emission suggests a synchrotron radiation origin for the wavelengths from mm up to optical/UV, and inverse Compton for the X-ray and γ-ray emission.

Such fast changes in synchrotron emission suggest that the timescale for particle acceleration must be comparably short. Jorstad et al.[502] offer three different theoretical models for the behaviour of 3C 454.3: passage of the jet plasma through standing recollimation shocks, magnetic reconnection events, and current-driven instability created where the jet changes from being magnetically dominated to matter dominated. The last mechanism causes turbulent flow downstream of the magnetic- to matter-dominated transition, which could create the conditions for more efficient particle acceleration and/or for more magnetic reconnection. More than one of these mechanisms could operate concurrently: they are not mutually exclusive.

Magnetic reconnection has been suggested as an acceleration mechanism in AGN by several authors[503, 504, 505, 506]. For suitable magnetic field configurations, it can be both efficient and fast, which is an advantage given the observed rapid variability of synchrotron radiation in blazars. Although early work[503] suggested fairly modest maximum energies of tens of GeV, more recent authors such as Giannios[504] claim that energies of $\sim 10^{20}$ eV are achievable for conditions appropriate to luminous AGN, though perhaps not for the less luminous FR I AGN that populate our local neighbourhood. Sironi et al.[506] further argue that magnetic reconnection is preferable to acceleration in relativistic shocks, because the latter mechanism works only for favourable magnetic geometry (as we saw earlier, if the magnetic field lines are oriented parallel to the shock front, particles are unlikely to achieve the repeated shock crossings needed to attain high energies). The underlying acceleration mechanism in magnetic reconnection is essentially first-order Fermi, with the particles bouncing off the converging magnetic flows instead of magnetic turbulence as in diffusive shock acceleration, with some drift acceleration along magnetic field gradients[505]. Results of 2D and 3D simulations differ significantly[505], so this is an area where increasing computational power and therefore increasingly sophisticated simulations should provide an improved understanding.

Finally, when we dismissed second-order Fermi acceleration (stochastic acceleration) as a mechanism for Galactic cosmic-ray emission in section 3.3, we did so on the grounds that it is too slow ($\propto (v/c)^2$, where v is tens of kilometres per second) and that it has an injection problem: it cannot accelerate protons from a standing (or at least thermal) start, because ionisation energy loss is faster than the acceleration. However, in the giant radio lobes of AGN, there is a strong, tangled magnetic field which may have some turbulent modes moving at speeds that are not small compared to c. This suggests[507] a much more favourable environment for stochastic acceleration: $(v/c)^2$ is $\mathcal{O}(0.01)$ or more, and there are many mechanisms that can supply a suprathermal seed population. For Centaurus A, Hardcastle et al.[507] calculate a stochastic acceleration timescale of a few Myr, which, while not suitable for supplying a rapidly varying luminosity, is entirely consistent with the spectral age of the lobes. However, tangled magnetic fields also offer opportunities for magnetic reconnection, as suggested by Stawarz et al.[497] in their analysis of compact X-ray emitting regions observed by the *Suzaku* X-ray satellite in Cen A's giant lobes.

In summary, the environments within AGN central regions, jets, lobes and hotspots all provide suitable conditions for particle acceleration, by a variety of mechanisms. Synchrotron radiation in the X-ray band provides direct evidence of local particle acceleration—because electrons radiating at these energies have very short energy-loss timescales, and cannot have travelled far—at the outer hotspots of FR II radio galaxies, in the inner parsec-scale jets of both FR I and FR II objects, and in the radio lobes of nearby FR I galaxies. Stochastic acceleration, acceleration at shocks and acceleration by magnetic reconnection have all been suggested in some or all of these locations: it is very likely that a variety of different mechanisms operate, even within the same galaxy.

AGN and high-energy cosmic rays

AGN certainly accelerate particles to very high energies, as shown by TeV photon emission from blazars. The observed flux of high-energy cosmic rays is easily accommodated by AGN: Hardcastle et al.[507] calculate that only 0.1% of the power supplied by Cen A's jets need go into accelerating cosmic rays in order to account for the observed cosmic ray flux from its direction. However, there are a few problems in identifying AGN as the origin of UHE cosmic rays. One of these is the lack of a significant spatial correlation: after promising initial results, the correlation of UHE cosmic rays as observed by the Pierre Auger Observatory with the local AGN population has steadily declined in significance, and the latest results[508] offer no statistically significant correlations, with the lowest chance probabilities (for arrivals within 18° of AGN at distances $<$130 Mpc from the *Swift* catalogue, and within 15° of Centaurus A) being 1.3% and 1.4% respectively (the choices of angular radius, minimum CR energy and—for the AGN catalogue—maximum distance are not fixed *a priori* but optimised for the best signal, so there is a substantial "look-elsewhere effect", but Auger say that the probabilities have been corrected for this). If the UHE cosmic rays are protons and they come from AGN, we should do better than this: Farrar et al.[509], using the Jansson-Farrar model of the Galactic magnetic field, predict a deviation of only 3.8° for a 60 EeV proton coming from Cen A. However, if the highest energy cosmic rays are heavier nuclei, as suggested by the $X_{\max}$ shower depth distribution[510], this problem essentially goes away: for the same energy, higher Z ions have lower rigidity and will be deflected more by the same magnetic field. Data from the Telescope Array, albeit with lower statistics, remain consistent with a pure proton composition at the highest energies[511]; the two experiments use different analysis strategies and hence have different systematics, so the significance of the apparent disagreement is unclear.

Meanwhile, the Telescope Array's "hotspot" (see figure 2.26), an extended region with a Gaussian σ of 10°, has a significance of 3.4σ including "look-elsewhere" correction[144], but is not obviously associated with a good candidate source.

In short, if we are to identify nearby AGN as the sources of UHE cosmic rays, it seems that a pure proton composition would be a real problem. The TA data strongly disfavour a very heavy (iron-dominated) composition, and indeed the Auger data do not indicate this either. It might, however, be possible to reconcile the directional data with a composition including a large fraction of light nuclei. If the maximum cosmic-ray energies produced in local AGN are not far above the GZK cut-off, as suggested for example by Hardcastle[466], an increase in mean atomic mass would be expected as a consequence of the dependence of rigidity on Z (see page 46), and this could result in a reduced proton content at these energies.

We must also, as usual, consider the question of whether we have evidence for the acceleration of *hadrons* in AGN, as synchrotron and inverse Compton radiation require only accelerated leptons. This is particularly relevant to AGN jets, since there is significant evidence from equipartition arguments that the jets are likely to consist of e^+e^- pair plasma rather than ionised gas[512]. The issue is that where the magnetic field energy density has been measured (by comparing inverse Compton luminosity with synchrotron), it is found to be within a factor of a few of the electron energy density[513]. However, if protons and electrons are travelling down the jet with comparable Lorentz factors, the proton energy density would be much higher than the electron energy density, owing to the higher mass of the proton. If we instead assume that they have similar kinetic energies, $(\gamma_e - 1)m_e c^2 = (\gamma_p - 1)m_p c^2$, then we end up concluding that $\gamma_p = 1 + (\gamma_e - 1)m_e/m_p$, which for electron γ factors of order 10 results in protons that are non-relativistic. It is surely unlikely that electrons and protons in an e^-p jet travel down the jet with significantly different speeds, so this favours e^+e^- jets (which are what the Blandford-Znajek mechanism naturally produces).

This is not an insuperable objection. FR I jets decelerate and entrain material from their surroundings, which will presumably be normal galactic and intergalactic gas. The jet termination shock in FR II galaxies is by definition the region where the jet head collides with the ambient intergalactic medium. Therefore, in both these cases there is a local population of protons and heavier nuclei to accelerate, even if the jet material itself is leptonic.

In this context, it is noteworthy that the TeV emission from Centaurus A does not fit an inverse Compton spectrum extrapolated from X-ray and GeV γ-ray data[514]. This can be accounted for by positing a hadronic origin for this component, e.g. pion decay[515] or photodisintegration of heavy nuclei[516]. However, as shown in figure 4.37, the angular resolution of the TeV signal is not sufficient to exclude a conservative explanation in which the TeV emission comes from a different region of the source, and it is possible to fit the TeV emission (by assuming an electron population with lower numbers but higher energy) without disturbing the fit to the radio-to-GeV data.

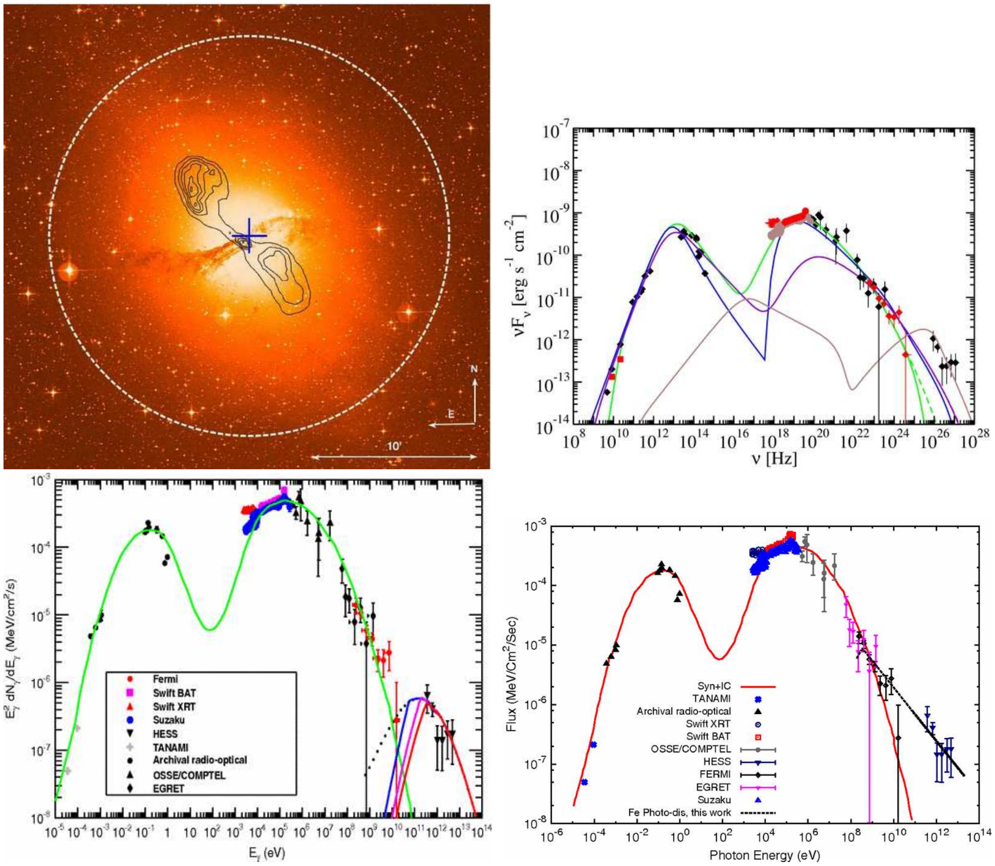


Figure 4.37: TeV emission from Centaurus A. Top left panel, the location of the TeV emission (blue cross, with the length of the arms indicating the 1σ error), overlaid on an optical image and a radio contour map of the inner lobes. The dashed circle marks the 95% confidence level upper limit of a possible extended component to the TeV emission (which is consistent with a point source). Figure from H.E.S.S.[492]. Top right, interpretation in terms of two sources, one accounting for the radio-to-GeV emission and the other for the TeV emission[514]. Bottom left, interpretation in terms of π^0 decay[515]; bottom right, interpretation as resulting from photodisintegration of heavy nuclei[516]. All three models offer good fits to the observations.

As in all cases of putative cosmic-ray sources, the key evidence would be detection of a nearby AGN as a point source of neutrinos. So far, such a detection has not been forthcoming: IceCube[517] find no significant signals for point or extended sources in four years of data taking. A stacked analysis looking specifically for correlations with *Fermi*–LAT blazars[518] also finds no signal, with the worst probability value for the null (no-signal) hypothesis being a deeply unimpressive 6%. The unavoidable conclusion from this is that γ-ray emitting blazars are *not* responsible for most of the observed astrophysical neutrinos. Whether this conclusion presents a problem for the hypothesis that UHE cosmic rays originate from AGN is debatable: neither [517] nor [518] makes any such claim, but other authors do.

For example, Jacobsen et al.[519] consider two models of Centaurus A in which protons are accelerated by shocks, in one case (Koers and Tinyakov[520]) close to the base of the jet and in the other (Becker and Biermann[521]) rather further out, in the parsec-scale inner jet. These models do not overproduce neutrinos for Cen A itself, but extrapolating from Cen A to the general AGN population (scaling by means of the X-ray luminosity) predicts a diffuse astrophysical neutrino flux very much in excess of the IceCube limit. Yoshida and Takami[522], considering the optical depth for photomeson production implied by the IceCube results, conclude that “none of the known extragalactic astronomical objects can be simultaneously a source of both PeV and trans-EeV energy cosmic rays”—they argue that the constraints imposed by IceCube on sources of $\mathcal{O}(10)$ PeV cosmic rays (assumed to be the parents of $\mathcal{O}(1)$ PeV neutrinos) become very difficult to satisfy if the same sources are also to generate the UHE cosmic rays with energies around a thousand times higher. This is similar to the conclusions reached by Baerwald et al.[450] regarding gamma-ray bursts, as discussed earlier.

Despite the failure to identify point sources, some information about the physics of neutrino emission can be gained from a study of the diffuse astrophysical neutrino flux. A recent IceCube paper[523] combines all the different IceCube studies in a joint maximum-likelihood analysis. The conclusion of this paper is that the spectrum of astrophysical neutrinos between 25 and 2800 TeV is best described by a power law with spectral index 2.50 ± 0.09 and an electron-neutrino fraction of $18 \pm 11\%$; the flux at 100 TeV is $\left(6.7^{+1.1}_{-1.2}\right) \times 10^{-18}$ GeV^{-1} s^{-1} sr^{-1} cm^{-2}. Since neutrinos carry off, on average, a fixed (energy-independent) fraction of the energy of their parent proton, this spectral index should mirror that of the originating proton population. The observed value is not incompatible with the cosmic-ray spectral index of $\sim$2.7, given that this spectrum is steepened by the preferential escape from the Galaxy of higher-energy cosmic rays (see section 3.8), but it is not consistent with the canonical value of 2 expected from diffusive shock acceleration (the IceCube analysis rejects a spectral index of 2.0 at 3.8σ significance, although this is weakened to 2.1σ if the power law is allowed to cut off exponentially at high energy). There is no evidence of any spectral break. The electron-neutrino fraction is interesting because it disfavours models in which cosmic rays escape from their sources as neutrons, which subsequently decay—even allowing for neutrino oscillation, this would lead to a higher fraction of ν_e.

Overall, it is clear that IceCube now has the potential to provide important constraints on AGN as cosmic-ray sources, as it is already doing for GRBs. The finding that *Fermi*–LAT blazars do not dominate the astrophysical neutrino flux is already a significant result, and the steep observed spectrum places interesting constraints on acceleration models. Because the models for particle acceleration in AGN are so varied, it is more difficult to draw firm conclusions than it is for sources such as supernova remnants where there are fewer degrees of freedom, but the next few years of IceCube data taking should prove very interesting.

In summary, the question of whether radio-loud AGN are the sources of the highest-energy cosmic rays remains undecided. The issue is made more difficult by the sheer number of different models for particle acceleration: does it take place primarily near the core, in the inner jets, or in the outer lobes; is the acceleration mechanism magnetic reconnection, diffusive shock acceleration, or stochastic acceleration; what, if anything, is the connection with TeV γ-ray emission; how strong is the evolution with cosmic epoch? The lack of detected neutrino point sources, and in particular the conclusion that the observed diffuse flux is not due principally to *Fermi*–LAT γ-ray blazars, is an interesting negative result, but not conclusive because it need not be true that "neutrino-loud" AGN and γ-ray emitting AGN are the same population. The lack of spatial correlation between local radio-loud AGN and UHE cosmic rays may be significant, or it may merely signal that the UHE cosmic rays are not predominantly protons. Overall, more data from both IceCube and the large cosmic-ray detectors are surely necessary to help resolve these issues.

4.5 Summary

In this chapter, we have considered some case studies of potential cosmic-ray sources. Shocks in the solar system, although they only accelerate particles to very modest energies, are of particular importance because they are close enough that we can measure their properties directly, allowing us to test our theoretical models and computer simulations; they also have practical importance through their effect on "space weather". However, for the sources of high-energy cosmic rays, we need to look further afield.

The cosmic ray energy spectrum spans about 12 orders of magnitude, from below 1 GeV to above 10^{11} GeV. It is clear that the lower end of this range is of Galactic origin—the gyroradius of a proton at these energies is much smaller than the dimensions of the Galaxy—and clear that the very top end is extragalactic; the exact point at which the switch occurs is still debated. The two main features of the cosmic-ray spectrum are the "knee" at $\sim 5 \times 10^6$ GeV and the "ankle" at around 5×10^9 GeV, with some experiments finding a "second knee" at $\sim 5 \times 10^8$ GeV[524]. When the spectrum is flattened by multiplying by a suitable power of E, the "ankle" structure appears as a pronounced dip in the spectrum at about 5×10^9 GeV. This is approximately consistent with the energy required for pair production off the cosmic microwave background, $p + \gamma_{\rm CMB} \rightarrow p + e^+ + e^-$: if this is indeed the correct explanation, it presupposes that the cosmic rays at this energy are predominantly protons, since the pair-production energy depends on the mass of the incoming cosmic ray. Finally, there is a sharp decline in the cosmic ray flux above $\sim 5 \times 10^{10}$ GeV, consistent with expectations from the "GZK cut-off", $p + \gamma_{\rm CMB} \rightarrow p + \pi^0$ (or $n + \pi^+$).

The composition of cosmic rays at high energy is determined indirectly by measuring the depth into the atmosphere at which the air shower created by the incoming cosmic ray reaches its maximum. There is general agreement among a range of experiments that the average atomic mass of cosmic rays increases from 10^6 GeV to $\sim 5 \times 10^7$ GeV and then declines towards the dip (see figure 2.14). There is much less agreement about what it does after the dip, with Auger data indicating a return to heavier composition while Telescope Array results remain close to the expectation for protons. As the two experiments use different analysis strategies and have different systematics, the extent of the disagreement may be less significant than it appears—the error bars are clearly correlated—but the lack of consensus complicates the interpretation.

These features of the cosmic ray spectrum lead to the following questions:

1. Where is the transition from Galactic to extragalactic cosmic rays? Is it related to the knee, the second knee, or the ankle?

2. Is the fall-off above the ankle due purely to the GZK cut-off, or does it reflect the maximum energy attainable in the sources?

Both of these questions are important in addressing the question of the origins of cosmic rays. If the transition from Galactic to extragalactic cosmic rays occurs at relatively low energies, it is easy to account for the Galactic cosmic rays, but the energetics of extragalactic cosmic rays present problems for gamma-ray bursts; if the transition is at high energies, GRBs remain a viable source for extragalactic cosmic rays, but it is not clear what Galactic source can achieve cosmic ray energies of order 10^9 GeV or more.

The question of the transition from Galactic to extragalactic sources is intimately related to the Galactic magnetic field and to the composition of the cosmic rays. Charged particles will "leak" out of the Galaxy if their gyroradii are of the same order as the thickness of the Galactic disc; lower-energy charged particles may also random-walk out if their gyroradii are of the same order as the characteristic length scale of the Galactic magnetic field. This is likely to be smaller than the dimensions of the Galaxy itself, as the field is complex and involves not only the large-scale Galactic field proper, but also smaller-scale random components due to turbulent flow of ionised gas and to the magnetic fields of individual objects such as supernovae[525]: these have a typical length scale of only ~100 pc, as opposed to ~300 pc for the scale height of the Galactic thin disc. Because the gyroradius is given by E/ZeB (for particles with $E \gg mc^2$), the leakage rate is related to both the strength of the magnetic field and its coherence length.

Owing to the Z-dependence of the gyroradius, heavier ions are confined to higher energies than protons. This is relevant both to the escape of cosmic rays from the Galaxy and to the confinement of proto-cosmic rays within an astrophysical accelerator: we should expect that a particular astrophysical object will be capable of accelerating cosmic rays up to a fixed rigidity (recall $R = cp/Ze$) rather than a fixed energy. In this context, the steady increase in $\langle \ln A \rangle$ between 10^6 and 5×10^7 GeV seen in figure 2.14 is significant, because it suggests that a particular class of source is "turning off" over this energy range. However, because the same constraints apply to both confinement in the source and confinement in the Galaxy, the same signature could also indicate that this is the energy range over which Galactic cosmic rays leak out and the transition to extragalactic sources takes place.

There is general agreement in the field that cosmic rays below the knee are Galactic in origin, and that they originate in supernova remnants, as discussed in section 4.3.3. The circumstantial evidence for this is strong: the observed flux of cosmic rays can be accounted for if ~10% of the SN energy goes into accelerating protons, which is consistent with the efficiencies found in simulations of diffusive shock acceleration; SNRs emit copious amounts of synchrotron radiation and are often γ-ray sources, indicating high levels of electron acceleration; there is clear evidence of shocks that could provide the acceleration mechanism; some SNRs have high-energy photon spectra which appear to be best explained by π^0 decay, which implies the presence of high-energy hadrons. Although there are other Galactic sources that could contribute, the case for supernova remnants being the dominant sources appears strong. Unfortunately the expected neutrino fluxes are quite small: IceCube[517] give limits for the Galactic supernova remnants Cas A and IC 443 that are more than an order of magnitude higher than the predicted flux, so it will take a *very* long time for the statistics to improve to the point where they might see something (the situation is better for SNR G40.5–0.5, where the observed limit is only a factor of 3 above the prediction—this could realistically be observable with 10–20 years of IceCube data). This is partly because the SNR neutrino energies of around 10 TeV are a bit low for IceCube, whose sensitivity peaks at around 100–1000 TeV (depending on source declination).

Between the knee and the ankle, there is much less consensus. Scenarios considered in the literature include:

- The knee is caused primarily by the natural end of the SNR cosmic-ray spectrum, intensified by the faster escape of higher-energy particles (see, e.g., [526]). The region between the knee and the ankle requires a second source, which may be Galactic or extragalactic.
- The knee is caused primarily by the faster escape of higher-energy particles, intensified by the natural end of the SNR cosmic-ray spectrum (the “escape model” of Giacinti et al.[527]). The flux above this is ascribed primarily to radio-loud AGN[528].
- The knee is a purely Galactic phenomenon, representing a transition from SNR cosmic rays to a second Galactic source type dominated by heavy nuclei (see, e.g., model A of [529]). The transition from Galactic to extragalactic cosmic rays takes place at the ankle.

If the ankle does not mark the transition between Galactic and extragalactic cosmic rays, it may represent a transition between two different extragalactic sources, or it may be an artefact caused by the combination of the pair-production dip and the GZK cut-off, as shown in figure 4.29. The latter model implicitly requires a proton-rich composition of UHE cosmic rays, because the pair-production and GZK energies are defined specifically by *protons* interacting with the CMB, not heavy nuclei; it is probably no coincidence that proponents of this model tend to compare their predictions with HiRes/Telescope Array data, which are consistent with a light composition at very high energies, rather than Auger data, which prefer a transition to heavier nuclei above the dip. Resolution of the UHE composition issue would be very helpful in constraining models.

The key problem for models with the Galactic/extragalactic transition at the ankle is understanding how Galactic sources manage to accelerate particles up to a few times 10^9 GeV; the key problems for models with transitions starting at the knee are how to explain the spectrum through the transition (it is easy to explain a transition occurring in association with a hardening of the spectrum—the high-energy source still contributes at lower energies, but is masked by the lower-energy source because of the latter’s steeper spectral index—but much harder to explain a smooth transition involving a steepening of the spectrum), and whether the required cosmic-ray flux is consistent with the IceCube neutrino data.

In general, the maximum energy attainable in a source of size R, magnetic field B and shock speed βc (assuming diffusive shock acceleration) is

$$E_{\max} = Ze\beta cRB,$$

where Ze is the charge of the ion being accelerated. If we consider $R \sim 3$ pc, $B \sim 0.3$ nT and $\beta \sim 0.01$, which are reasonable values for the size and blastwave speed of a supernova remnant and the Galactic magnetic field, we get a maximum energy of $\sim 10^5$ GeV for protons, which is about a factor of 50 or so below the knee. Allowing for amplification of the magnetic field to $\mathcal{O}(10)$ nT in the region of the supernova blast wave, as discussed on page 188, the maximum energy goes up enough to reach the knee with protons, and up to a factor of 26 ($= Z$) higher for heavier nuclei up to iron. This would fit the observed increase in $\langle \ln A \rangle$ between 10^6 and 5×10^7 GeV (see figure 2.14). Thus, a simplistic picture of particle acceleration in supernova remnants appears to prefer a model in which the Galactic component of cosmic rays shuts off at around 10^8 GeV and a proton-rich extragalactic component takes over.

If we wish to extend the Galactic contribution all the way to the ankle, $E \sim 5 \times 10^9$ GeV, we need an extra Galactic (as opposed to extragalactic!) contribution. Drury[530] suggests that very young supernova remnants with faster shocks may supply this; Blasi[531] considers supernovae of type IIb, but regards the required parameters as unrealistic ("it is clear however that all parameters need to be pushed to their extreme values in order to realize this situation") and prefers young pulsars and pulsar wind nebulae. As discussed earlier, these may accelerate particles by magnetic reconnection rather than diffusive shock acceleration, and it is quite possible (acceleration by magnetic reconnection is not nearly as well studied as shock acceleration, so one cannot be sure) that this has a higher maximum energy.

An issue with these models is that the current IceCube astrophysical neutrino data cover the energy range $10^4 - 10^{6.5}$ GeV[517], and are depressingly isotropic[523]. If we assume that neutrinos from photopion production carry off roughly 5% of the energy of the parent proton, this means that these neutrinos are tracing the proton energy range $\sim 10^{5.5} - 10^8$ GeV. In models where the knee marks the end of the Galactic component for protons, most of this range is extragalactic, and the isotropic distribution is perfectly reasonable. In models in which the Galactic component of cosmic rays continues to the ankle, most of these neutrinos come from Galactic sources—so why is there no concentration near the Galactic plane, where supernova remnants, young pulsars, and pulsar wind nebulae all live? As an example of what we might expect, consider TeVCat[257] (see figure 2.60): the extragalactic sources of TeV γ-rays (nearly all blazars) are scattered all over the sky, but the Galactic sources (mostly supernova remnants and pulsar wind nebulae) are tightly bunched on the Galactic plane, with a strong preference for the inner Galaxy (within about 60° of the direction of the Galactic centre). No such preference is apparent in the IceCube neutrino data.

The alternative is to assume that the straightforward interpretation of the rise in $\langle \ln A \rangle$ above 10^6 GeV—that it represents the maximum rigidity attainable by Galactic sources—is correct, and that extragalactic sources start to contribute to the overall cosmic ray spectrum from then on and have more or less completely taken over by 10^8 GeV. This would essentially rule out GRBs as principal sources—as discussed above, such a model violates IceCube's limits on the (lack of) association of neutrinos with GRBs—but is feasible for AGN. It is, however, surprising in this model that the "stacked" IceCube analysis[518] finds no significant association between the neutrino sample and *Fermi*–LAT blazars. It is also surprising that there is no association of neutrinos, and no significant association of UHE cosmic rays, with the most nearby AGN, particularly Cen A (there is a weak correlation with UHE cosmic rays, but no neutrino events) and M87 (there are some neutrino events, but the chance probability is 26%): as these are much closer than the typical AGN, one would have expected that they would be significant neutrino sources. As noted above, it is also more difficult to understand how a transition from one source to another could be associated with a steepening of the spectrum (but no other notable spectral features).

The strong suppression of the cosmic-ray flux at energies above a few times 10^{10} GeV, which is seen by both Auger and the Telescope Array (and earlier by HiRes), is also an important input into theories of cosmic-ray origins. It has several possible explanations[532]:

- the GZK effect—reduction of proton energies caused by pion photoproduction off CMB photons, $p\gamma \rightarrow p\pi^0$ or $n\pi^+$;
- photodisintegration of heavier nuclei (with $A \lesssim 20$), caused by excitation of various nuclear resonances;
- "turn-off" of the source.

In the first two cases, the initial cosmic-ray spectrum may extend well above 10^{11} GeV, and the observed cut-off is caused by energy losses during propagation. When the UHE composition is proton-rich, the result of this is a "pile-up" at energies just below the cut-off, because the protons will lose energy through photopion production until their energies are too low for this reaction to go, and thereafter will be able to propagate with largely unchanged energies. If the composition is heavier, so that photodisintegration is the more important mechanism, this will not happen, because the lighter nuclei resulting from the collision tend to be subject to further photodisintegration reactions (they are less tightly bound than their heavier parent) and thus there is no pile-up[532, 533]. In the third case, although the fast decline in the observed flux is presumably exacerbated by propagation effects, the maximum energy even in the absence of such effects would not be much above 10^{11} GeV, so no significant pile-up is expected even for proton-rich composition.

The GZK cut-off comes from the production of pions, which would subsequently decay. Since roughly half of the pions are charged, the result should be a diffuse flux of very-high-energy neutrinos (mostly > 100 PeV), generally referred to in the literature as **cosmogenic neutrinos**. These have not been observed: the IceCube neutrino signal does not extend beyond a few PeV, and their analysis[534] indicates that the two events in the 2010–2012 dataset with the highest deposited energies (around 1 PeV) are not consistent with a cosmogenic origin, essentially because of the lack of any signal at higher energies. The principal effect of this non-observation is to rule out models in which UHE cosmic rays are mostly protons *and* come from sources which are much more numerous at high redshift (the cosmogenic neutrino flux predicted by the local UHE cosmic-ray flux is below IceCube's current sensitivity, but if UHE cosmic rays were much more common at high redshift, the cosmogenic neutrinos they produced would still be present and detectable). This rules out a scenario in which only FR II radio galaxies are cosmic ray sources, because they fall into this category—but we have seen above that the majority of radio-loud AGN are low-excitation, low-luminosity FR I/BL Lac radio galaxies, and these have much weaker redshift evolution. The bound can also be evaded if the UHE cosmic rays are not predominantly protons, since the neutrino yield from photodisintegration of heavy nuclei is much lower.

The energy spectrum of the highest-energy cosmic rays should in principle help to discriminate between models, but in practice does not: Kotera and Olinto[533] find acceptable fits to the Auger data for essentially every model they consider, and although Harari's fit[532] appears to prefer a model in which the maximum energy essentially coincides with the GZK cut-off, the Auger energy scale error is large enough to permit a model with a higher cut-off (though the shape of the spectrum seems wrong for a higher cut-off in the case of an iron-dominated composition).

In short, it appears that, despite significant improvements in both theory—especially realism of simulations—and observation over recent years, the origin of cosmic rays above the knee remains puzzling. The lack of any evidence for point sources in the IceCube astrophysical neutrino sample is disappointing, and the lack of correlation with the Galactic plane, *Fermi*–LAT blazars, or GRBs only deepens the mystery. While the candidate sources discussed above remain the prime suspects, more neutrino data are needed to narrow down the list, and a reconciliation of the discrepant results on UHE cosmic ray composition by the Pierre Auger Observatory and the Telescope Array is also urgently needed.

4.6 Questions and Problems

1. Anomalous cosmic rays were discovered in the 1970s and so called because they presented "anomalous" energy spectra at low energies ($\sim$10 MeV per nucleon). The seven elements initially identified as anomalous were H, He, N, C, O, Ne and Ar, with the effect for carbon being substantially less than for nitrogen or oxygen. It is now generally accepted that the anomalous cosmic rays are neutral atoms from the interstellar medium which enter the heliosphere and are then ionised, "picked up" by the solar wind (hence the term "pickup ions"), transported back to the outer heliosphere and accelerated probably at the solar wind termination shock (see section 4.2.3). Explain why this is more likely to happen for these particular elements than for other common elements such as iron and silicon.

2. Aluminium-26 has a mean lifetime of 10^6 years. Its decay to ^{26}Mg yields a γ-ray line at 1809 keV which is observed by INTEGRAL–SPI[367]. The inferred mass of ^{26}Al in the Milky Way Galaxy is $(2.8 \pm 0.8)M_\odot$. ^{26}Al is produced by massive stars (mass range $\sim$10–$120M_\odot$), and theoretical calculations of the yield per star and the initial mass function for massive stars indicate that the average yield per massive star is $1.4\times10^{-4}M_\odot$, with a systematic error of $\pm 50\%$. Estimate the rate of core-collapse supernovae in units of supernovae per century, stating any assumptions that you make.

3. Figure 4.38 shows the spectral energy distribution of the supernova remnant HESS J1640–465[535]. Identify the emission mechanisms corresponding to the blue dashed, red dashed, green dashed and black solid lines on this figure, and comment on the implications for cosmic-ray acceleration in this SNR.

4. The supernova remnant G11.2–0.3 is believed to correspond to a "guest star" observed by the Chinese in AD 386. Pulsar J1811–1926, located inside this young SNR, has a period of 64.66 ms and a spin-down rate of

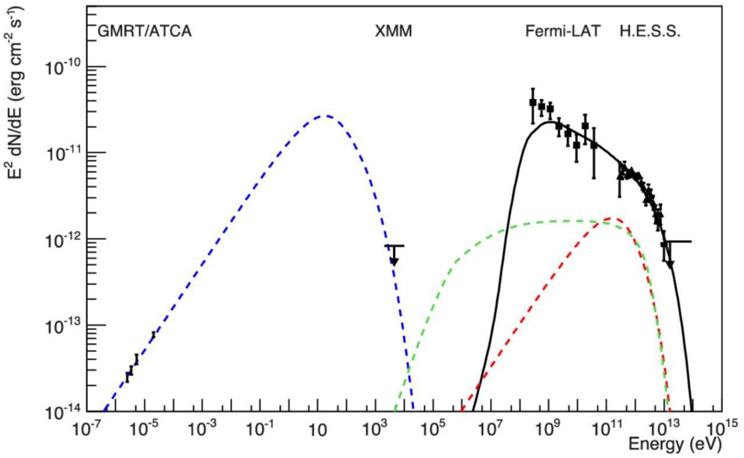


Figure 4.38: Spectral energy distribution for supernova remnant HESS J1640–465, with data from H.E.S.S., *Fermi*–LAT, and radio telescopes, and a non-detection by XMM–Newton.

6.4×10^{-13} s s^{-1}. Calculate the spin-down age of PSR J1811–1926, and comment on your result. Repeat this calculation for PSR J0205+6449, associated with supernova remnant 3C 58, which has a period of 65.69 ms and a spin-down rate of 1.93×10^{-13} s s^{-1}; this SNR is thought to correspond to a supernova observed in 1181.

5. Suppose that a cosmic ray observatory at the South Pole had a sample of 500 events above 5×10^{19} eV. Assume that the observatory can observe events at angles up to 80° from the zenith, and make the unrealistic assumption that their detection efficiency does not depend on zenith angle. How many pixels 10° in radius are contained in the fraction of the sky that the observatory can see? How many events in one such pixel would constitute a 3σ signal (i) if you knew in advance which pixel should contain the signal (e.g. the pixel centred on Centaurus A); (ii) if you did not specify the "signal" pixel in advance?

   The Pierre Auger Observatory[508] conducts a search for localised cosmic ray sources by scanning over window sizes varying from 1° to 30° in radius (in steaps of 1°), with threshold energies varying from 40 to 80 EeV in 1 EeV steps. Their best "signal" (for a radius of 12° and a threshold energy of 54 EeV) is 14 events compared to an expectation of 3.23, with a statistical probability that they quote as 5.9×10^{-6} (for a simple Poisson probability I got 7.7×10^{-6}, same order of magnitude). Explain, based on your previous answer, why this is *not* a significant signal (in fact, the actual probability of obtaining a result of at least this significance is a whopping 69%).

Bibliography

[1] *Astroparticle Physics*, http://www.journals.elsevier.com/astroparticle-physics/.

[2] STFC, http://www.stfc.ac.uk/search.aspx?s=particle%20astrophysics&m=f (search results for search string "particle astrophysics" as of April 22, 2014).

[3] STFC, http://www.stfc.ac.uk/91.aspx (accessed April 22, 2014).

[4] ApPEC, http://www.appec.org/roadmap.html (2008, 2011). [Quoted text is from the 2011 document, but the text in the 2008 version is almost identical.]

[5] NJC Spooner, http://www.sheffield.ac.uk/physics/teaching/phy326/index

[6] DJ Fixsen, *ApJ* **709** (2009) 916–920.

[7] SL Cartwright, http://www.hep.shef.ac.uk/cartwright/phy306.

[8] LF Thompson, http://www.sheffield.ac.uk/physics/teaching/phy320/index.

[9] AH Guth, *Phys. Rev.* **D23** (1981) 347–356.

[10] AD Linde, *Phys. Lett.* **B100** (1981) 37–40.

[11] PAR Ade et al. (Planck Collaboration), arXiv 1303.5076 [astro-ph.CO] (2013).

[12] G Hinshaw et al. (WMAP Collaboration), arXiv 1212.5226 [astro-ph.CO] (2012).

[13] PAR Ade et al. (BICEP2 Collaboration), arXiv 1403.3985 [astro-ph.CO] (2014).

[14] There are many articles on the Higgs mechanism, most of them either too simplistic or too complicated for this level. A reasonable compromise is Matt Strassler's online articles at http://profmattstrassler.com/articles-and-posts/particle-physics-basics/how-the-higgs-field-works-with-math/ (2012).

[15] F Bezrukov, *Class. Quantum Grav.* **30** (2013) 214001.

[16] AD Linde, arXiv 1402.0526 [hep-th] (2014).

[17] The various forms of the anthropic principle have been discussed in numerous books and articles. The Wikipedia article, `http://en.wikipedia.org/wiki/Anthropic_principle`, is a reasonable introduction.

[18] AD Sakharov, *Zh. Eksp. Teor. Fiz. Pis'ma* **5** (1967) 32 (in Russian); *JETP Lett.* **91B** (1967) 24 (English trans.).

[19] G Gamow, *Phys. Rev.* **70** (1946) 572–573.

[20] G 't Hooft, *Phys. Rev. Lett.* **37** (1976) 8–11.

[21] R Saakyan, *Ann. Rev. Nucl. Part. Sci.* **63** (2013) 503–529.

[22] M. Agostini et al. (GERDA Collaboration), *Phys. Rev. Lett.* **111** (2013) 122503.

[23] A nice introduction to the physics of massive neutrinos (the numbers are well out of date but the theory isn't) is
SF King, arXiv 0712.1750 [physics.pop-ph] (2007).
For reasonably up-to-date values of parameters, consult the Particle Data Group website at `http://pdg.lbl.gov/`, especially the review article on "Neutrino mass, mixing and oscillations".

[24] W Buchmüller, RD Peccei and T Yanagida, *Ann. Rev. Nucl. Part. Sci.* **55** (2005) 311–355.

[25] J Cline, arXiv 0609145 [hep-ph] (2006).

[26] `http://arxiv.org/find/hep-ph`

[27] A convenient review of dark energy, not too out of date, is
JA Frieman, MS Turner and D Huterer, *ARAA* **46** (2008) 385–482.

[28] OE Bjaelde et al., *JCAP* **0801** (2008) 026; arXiv 0705.2018 [astro-ph].

[29] MG Aartsen et al. (IceCube Collaboration), *Science* **342** (2013) 1242856.

[30] A Goldwurm et al., arXiv:astro-ph/0102386 (2001).

[31] See list in the Wikipedia article,
`http://en.wikipedia.org/wiki/Coded_aperture`.

[32] COMPTEL web page, `http://heasarc.gsfc.nasa.gov/docs/cgro/comptel/` (2005).

[33] EGRET web page, `http://heasarc.gsfc.nasa.gov/docs/cgro/egret/` (2005).

[34] WB Atwood et al., *ApJ* **697** (2009) 1071–1102.

[35] `http://en.wikipedia.org/wiki/Cherenkov_radiation`

[36] PAMELA website, `http://pamela.roma2.infn.it/index.php`

[37] AMS-02 website, `http://ams.nasa.gov/`

[38] Auger website, `http://www.auger.org/`

[39] QR Ahmad et al. (SNO Collaboration), *Phys. Rev. Lett.* **89** (2002) 011301.

[40] S Ando and K Sato, *New J. Phys.* **6** (2004) 170.

[41] K Bays et al. (Super-Kamiokande Collaboration), *Phys. Rev.* **D85** (2012) 052007;
H. Zhang et al. (Super-Kamiokande Collaboration), arXiv:1311.3738 [hep-ex] (2013).

[42] See, for example,
Y Ashie et al. (Super-Kamiokande Collaboration), *Phys. Rev.* **D71** (2005) 112005.

[43] Y Fukuda et al. (Super-Kamiokande Collaboration), *Phys. Rev. Lett.* **81** (1998) 1562–1567.

[44] IceCube website, `https://icecube.wisc.edu/`.

[45] ANITA website, `http://www.phys.hawaii.edu/∼anita/new/index.html`.

[46] AG Vieregg (for the ANITA Collaboration), *Nucl. Phys. B (Proc. Supp.)* **229** (2012) 545.

[47] ACORNE website, `http://www.hep.shef.ac.uk/research/acorne/`.

[48] JL Feng, *ARAA* **48** (2010) 495–545.

[49] See the Wikipedia article on Goldstone bosons,
`http://en.wikipedia.org/wiki/Goldstone_boson`.

[50] GG Raffelt, *Lect. Notes Phys.* **741** (2008) 51–71.

[51] GG Raffelt and LJ Rosenberg, “Axions and other similar particles” in J Beringer et al. (Particle Data Group), *Phys. Rev.* **D86** (2012) 010001;
see `http://pdg.lbl.gov/2013/reviews/contents_sports.html`.

[52] A convenient, if US-biased, summary can be found in the Snowmass CF1 report,
D Bauer et al., arXiv 1310.8327 [hep-ex] (2013),
and its associated website,
`http://www.snowmass2013.org/`
`tiki-index.php?page=WIMP+Dark+Matter+Direct+Detection`

[53] DRIFT website, `http://driftdarkmatter.org/`;
Sheffield DRIFT web page,
`http://www.hep.shef.ac.uk/research/dm/drift.php`.

[54] DS Akerib et al. (LUX Collaboration), *Phys. Rev. Lett.* **112** (2014) 091303.

[55] J Angle et al. (XENON Collaboration) *Phys. Rev. Lett.* **107** (2011) 051301.

[56] S Adrián-Martinez et al. (ANTARES Collaboration), arXiv 1302.6516 [astro-ph.HE] (2013);
MG Aartsen et al. (IceCube Collaboration), *Phys. Rev. Lett.* **110** (2013) 131302.

[57] PAMELA: O Adriani et al., *Nature* **458** (2009) 607;
Fermi–LAT: M Ackermann et al., *Phys. Rev. Lett.* **108** (2012) 011103;
AMS-02: M Aguilar et al., *Phys. Rev. Lett.* **110** (2013) 141102.

[58] T Daylan et. al., arXiv 1402.6703 [astro-ph.HE] (2014)

[59] GA Gómez-Vargas et al., *JCAP* **1310** (2013) 029.

[60] M Ackermann et al. (*Fermi*–LAT Collaboration), *Phys. Rev.* **D89** (2014) 042001.

[61] ALEPH, DELPHI, L3, OPAL, and SLD Collaborations, and LEP Electroweak Working Group, and SLD Electroweak Group, and SLD Heavy Flavour Group, *Physics Reports* **427** (2006) 257.

[62] L Calibbi et al., *JHEP* **10** (2013) 132.

[63] ADMX website,
`http://www.phys.washington.edu/groups/admx/home.html`.

[64] CAST website, `http://cast.web.cern.ch/CAST/CAST.php`

[65] ALPS website, `https://alps.desy.de/`.

[66] *Uhuru* web page,
`http://heasarc.gsfc.nasa.gov/docs/uhuru/uhuru.html`.

[67] VF Hess, *Physik. Zeitschr.* **13** (1912) 1084; *Physik. Zeitschr.* **14** (1913) 610.
The text of these papers (in German) can be found at the Innsbruck University web page
`http://physik.uibk.ac.at/hephy/Hess/homepage/Hess_paper01.html`;
`http://physik.uibk.ac.at/hephy/Hess/homepage/Hess_paper02.html`.

[68] `http://adsabs.harvard.edu/abstract_service.html`.

[69] CS Wright, *Nature* **117** (1926) 54–56.

[70] RA Millikan, *PNAS* **16** (1930) 421–425.

[71] AH Compton, *Phys. Rev.* **41** (1932) 111–113.

[72] TH Johnson, *Phys. Rev.* **43** (1933) 834–835.

[73] L Alvarez and AH Compton, *Phys. Rev.* **43** (1933) 835–836.

[74] TH Johnson and JG Barry, *Phys. Rev.* **56** (1939) 219–226.

[75] CD Anderson, *Phys. Rev.* **43** (1933) 491–494.

[76] SH Neddermeyer and CD Anderson, *Phys. Rev.* **51** (1937) 884–886.

[77] JC Street and EC Stevenson, *Phys. Rev.* **52** (1937) 1003–1004.

[78] GPS Occhialini and CF Powell, *Nature* **159** (1947) 186–190; *Nature* **160** (1947) 453–456.

[79] GD Rochester and CC Butler, *Nature* **160** (1947) 855–857.

[80] JL Heilbron, RW Seidel, BR Wheaton. *Lawrence and his Laboratory*, `http://www.lbl.gov/Science-Articles/Research-Review/Magazine/1981/`, Chapter 6 (1981, 1996).
Note: none of the pictures work, and any attempt to click on them takes you to some useless commercial site. For images, use the LBL Gallery, `http://lbl.webdamdb.com/albums.php?albumId=129359`.

[81] AM Hillas, arXiv:astro-ph/0607109 (2006).

[82] CREAM website, `http://cosmicray.umd.edu/cream/`.

[83] ACE website, `www.srl.caltech.edu/ACE/ace_mission.html`.

[84] TD Sloan, *J. Phys. Conf. Series* **409** (2013) 012020.

[85] `http://stratocat.com.ar/fichas-e/2007/MCM-20071219.htm`

[86] A Andronic and JP Wessels, *NIMPA* **666** (2012) 130–147.

[87] TK Gaisser, T Stanev, S Tilav, *Front. Phys.* **8** (2013) 748–758.

[88] Telescope Array website, `http://www.telescopearray.org/`.

[89] `http://emtoolbox.nist.gov/Wavelength/Documentation.asp`

[90] SF Berezhnev et al. (Tunka–133 Collaboration), *NIMPA* **692** (2012) 98–105.

[91] B Keilhauer et al., arXiv:1210.1319 [astro-ph.HE] (2012).

[92] HiRes website, `http://hires.physics.utah.edu/reading/flyseye.html`.

[93] RA Ong, `www.astro.ucla.edu/∼rene/talks/Cronin-Fest-Ong-Writeup.pdf`

[94] A Tonachini (Pierre Auger Collaboration), arXiv:1307.5059 [astro-ph.HE] (2013) 112–115.

[95] J Abraham et al. (Pierre Auger Collaboration), *NIMPA* **620** (2010) 227–251.

[96] Pierre Auger Collaboration, *Pierre Auger Observatory Design Report* (1997), p77.

[97] KASCADE-Grande website, `https://web.ikp.kit.edu/KASCADE/`

[98] T Abu-Zayyad et al. (Telescope Array Collaboration), *ApJL* **768** (2013) L1.

[99] R Pesce (Pierre Auger Collaboration), arXiv:1107.4809 [astro-ph.HE] (2011) 13–16.

[100] A Haungs et al. (KASCADE-Grande Collaboration), arXiv:0910.4824 [astro-ph.HE] (2009).

[101] V Verzi (Pierre Auger Collaboration), arXiv:1307.5059 [astro-ph.HE] (2013) 7–10.

[102] HiRes website, `http://hires.physics.utah.edu/index.html`.

[103] T Abu-Zayyad et al. (Pierre Auger Observatory and Telescope Array Collaborations), arXiv:1310.0647 [astro-ph.HE] (2013).

[104] ES Seo et al. (CREAM Collaboration), *Adv. Sp. Sci. Res.* **33** (2004) 1777–1785.

[105] EC Stone et al. (CRIS Collaboration) *Sp. Sci. Rev.* **86** (2998) 285–356.

[106] AW Labrador et al. (CRIS), in *Proc. 28th ICRC* ed. T Kajita et al. (2003) 1773–1776.

[107] H Bichsel, DE Groom and SR Klein, "Passage of particles through matter" in J Beringer et al. (Particle Data Group), *Phys. Rev.* **D86** (2012) 010001;
see `http://pdg.lbl.gov/2013/reviews/contents_sports.html`.

[108] O Adriani et al. (PAMELA Collaboration), *ApJ* **770** (2013) 1–9.

[109] L Arruda et al. (AMS-RICH Collaboration), arXiv:astro-ph/0306224 (2003).

[110] CAPRICE web page, `http://ida1.physik.uni-siegen.de/caprice.html`.

[111] M Boezio et al. (CAPRICE), *ApJ* **518** (1999) 457–472.

[112] IceTop website, `https://icecube.wisc.edu/science/icetop`.

[113] T Abu-Zayyad et al. (Telescope Array Collaboration), arXiv:1305.7273 [astro-ph.HE] (2013).

[114] SP Swordy, *Sp. Sci. Rev.* **99** (2001) 85–94.

[115] W Hanlon, `http://www.physics.utah.edu/∼whanlon/spectrum.html` (undated, probably 2009).

[116] K Greisen, *Phys. Rev. Lett.* **16** (1966) 748–750;
GT Zatsepin and VA Kuz'min, *JETP Lett.* **4** (1966) 78–80.

[117] See, e.g., H van Pee et al. (CB-ELSA Collaboration), *Eur. Phys. J.* **A31** (2007) 61–77.

[118] K-H Kampert and P Tinyakov, arXiv:1405.0575 [astro-ph.HE] (2014).

[119] RU Abbasi et al. (HiRes Collaboration), *Phys. Rev. Lett.* **104** (2010) 161101.

[120] JS George et al. (CRIS), *ApJ* **698** (2009) 1666–1681.

[121] JZ Wang et al. (BESS Collaboration), *ApJ* **564** (2002) 244–259.

[122] K Lodders, *ApJ* **591** (2003) 1220–1247.

[123] D Maurin, F Melot and R Taillet, arXiv:1302.5525 [astro-ph.HE] (2013);
`http://lpsc.in2p3.fr/cosmic-rays-db/`

[124] AW Strong, IV Moskalenko and VS Ptuskin, *Ann. Rev. Nucl. Part. Sci.* **57** (2007) 285–327.

[125] AW Strong et al., `http://galprop.stanford.edu/`

[126] ME Wiedenbeck et al. (CRIS), *ApJ* **523** (1999) L61–L64.

[127] SM Niebur et al. (CRIS), *Proc. 27th ICRC* (2001) 1675–1678.

[128] S Chekanov for the ZEUS Collaboration, arXiv:0705.4232 [hep-ex] (2007).

[129] H Fuke et al. (BESS) *Phys. Rev. Lett.* **95** (2005) 081101.

[130] P Chardonnet, J Orloff and P Salati, *Phys. Lett.* **B409** (1997) 313–320.

[131] N Martin for the ALICE Collaboration, *J. Phys. Conf. Series* **455** (2013) 012007.

[132] IV Moskalenko et al., *ApJ* **565** (2002) 280–296.

[133] Recent examples include:
decaying WIMP dark matter, Choi, Kyae and Shin, *Phys. Rev.* **D89** (2014) 055002;
new vector lepton doublet, Tomar and Mohanty, arXiv:1403.6301 [hep-ph];
decaying dark atoms, Belotsky et al., arXiv:1403.1212 [astro-ph.CO];
Higgs portal decaying vector dark matter (whatever that is!), Baek et al., arXiv:1402.2115 [hep-ph];
local dark matter overdensity, Hector et al., *Phys. Lett.* **B728** (2014) 58–62.

[134] M Di Mauro et al., arXiv 1402.0321 [astro-ph.HE] (2014),
show that good agreement with the data can be obtained by assuming that all local pulsars contribute primary positrons, or that only a few of the most powerful ones do, or even that only a single powerful source does.

[135] See, for example,
AA Abdo et al. (*Fermi*-LAT Collaboration), *ApJ* **720** (2010) 272.

[136] O Adriani et al. (PAMELA Collaboration), *ApJ* **765** (2013) 91.

[137] Tibet–ASγ website, `http://www.icrr.u-tokyo.ac.jp/em/`.

[138] M Amenomori et al. (Tibet-ASγ Collaboration), *Science* **314** (2006) 439–443.

[139] RU Abbasi et al. (IceCube Collaboration), *ApJ* **740** (2011) 16.

[140] NA Schwadron et al. (IBEX), *Science* **343** (2014) 988–990.

[141] IBEX website, `http://www.ibex.swri.edu/`.

[142] AA Abdo et al. (MILAGRO), *Phys. Rev. Lett.* **101** (2008) 221101.

[143] B Bartoli et al. (ARGO-YBJ Collaboration), arXiv 1309.6182 [astro-ph.HE] (2013).

[144] RU Abbasi et al. (Telescope Array Collaboration), *ApJ* **790** (2014) L21.

[145] M Ángeles Pérez-García, K Kotera and J Silk, *NIMPA* **742** (2014) 237–240.

[146] R Kumar and D Eichler, *ApJ* **785** (2014) 129.

[147] P Abreu et al. (Pierre Auger Collaboration), *Ap. Phys.* **34** (2010) 314–326.

[148] *Swift*-BAT 58-month catalogue,
`http://swift.gsfc.nasa.gov/results/bs58mon/` (2012).

[149] T Abu-Zayyad et al. (Telescope Array Collaboration), *ApJ* **777** (2013) 88.

[150] M-P Véron-Cetty and P Véron, *A&A* **455** (2006) 773.
Note that this is not an ideal catalogue for this purpose: the authors state explicitly that "This catalogue should not be used for any statistical analysis as it is not complete in any sense". However, the Pierre Auger Collaboration has used it consistently, so the decreasing level of correlation reported between 2007 and 2012 is worthy of note.

[151] C Macolino for the Pierre Auger Collaboration, *J. Phys. Conf. Series* **375** (2012) 052002.

[152] P Abreu et al. (Pierre Auger Collaboration), *Science* **318** (2007) 938.

[153] For a convenient compilation of information on Cen A, with links to sources, see
H Steinle, `http://www.mpe.mpg.de/~hcs/Cen-A/cen-a-facts.html`.

[154] ESA/Hubble (F Granato),
`http://www.eso.org/public/unitedkingdom/images/atm_opacity/`.

[155] KG Jansky, *Pop. Ast.* **41** (1933) 548–555.

[156] G Reber, *ApJ* **91** (1940) 621–624.

[157] G Reber, *ApJ* **100** (1944) 279–287.

[158] W Baade and R Minkowski, *ApJ* **119** (1954) 215–231.

[159] HS Hudson, L Fletcher and S Krucker, arXiv:1001.1005 [astro-ph.SR] (2010).

[160] HP Warren, *ApJSS* **157** (2005) 147–173.

[161] C Wehrli,
`http://rredc.nrel.gov/solar/spectra/am0/wehrli1985.new.html`.

[162] Nobeyama Radio Observatory, `http://solar.nro.nao.ac.jp/`.

[163] K Shibasaki,
`http://solar.nro.nao.ac.jp/MicrowaveSunspot201306.pdf` (2013).

[164] See, for example,
TL Wilson, arXiv:1111.1183 [astro-ph.IM] (2011).

[165] PAR Ade et al. (Planck Collaboration), *A&A* **536** (2011) A20.

[166] A list of important radio spectral lines—maintained for the purpose of trying to prevent these frequencies being assigned for commercial use—is provided by the Committee on Radio Astronomy Frequencies at
`http://www.craf.eu/iaulist.htm` (1991, 2000).

[167] See, for example, the Diamond Light Source at the Rutherford Appleton Laboratory, `http://www.diamond.ac.uk/Home/About.html`.

[168] FR Elder, AM Gurewitsch, RV Langmuir and HC Pollock, *Phys. Rev.* **71** (1947) 829–830.

[169] Cyclotron emission is widely observed in AM Herculis binary systems, see for example
M Cropper et al., *MNRAS* **236** (1989) 29P–38P.

[170] Cyclotron emission in the transient X-ray pulsar V0332+53 is discussed in
SS Tsygankov, AA Lutovinov, EM Churazov and RA Sunyaev, *MNRAS* **371** (2005) 19–28.

[171] MS Longair, *High Energy Astrophysics* 3rd Edition (Cambridge University Press, 2011).

[172] This statement of Parseval's Theorem is often called Plancherel's Theorem. For the (straightforward) proof see, for example,
`http://mathworld.wolfram.com/PlancherelsTheorem.html`

[173] H Bethe and W Heitler, *Proc. Roy. Soc. Lon.* **A146** (1934) 83–112.

[174] LH Thomas, *Math. Proc. Cam. Phil. Soc.* **23** (1927) 542–548.

[175] E Fermi, *Rend. Accad. Naz. Lincei* **6** (1927) 602–607.

[176] A brief summary of the main features of this model can be found in the Wikipedia article,
`http://en.wikipedia.org/wiki/Thomas-Fermi_model`.

[177] RK Campbell et al., *ApJ* **678** (2008) 1304–1315.

[178] For an overview of Bessel functions, see Wikipedia,
`http://en.wikipedia.org/wiki/Bessel_function`,
or Wolfram MathWorld,
`http://mathworld.wolfram.com/ModifiedBesselFunctionoftheSecondKind.html`.
And yes, this is the same Bessel that first measured the parallax of a star: versatile chaps those 19th century astronomers.

[179] JJ Condon and SM Random,
`http://www.cv.nrao.edu/course/astr534/ERA.shtml`, section 5.

[180] JD Peterson and WR Webber, *ApJ* **575** (2002) 217–224.

[181] For a discussion of $\Gamma(z)$ see Wikipedia,
`http://en.wikipedia.org/wiki/Gamma_function`,
or Wolfram MathWorld,
`http://mathworld.wolfram.com/GammaFunction.html`.

[182] T Peters et al., *ApJ* (2010) **719** 831–843.

[183] CD Dermer et al., *Proc. 2012 Fermi Symp.*(eConf C121028); arXiv:1303.6482 [astro-ph.HE] (2013).

[184] JM Casandjian, *AIP Conf. Proc.* **1505** (2012) 37–45.

[185] CD Dermer, *ApJ* **307** (1986) 47–59.

[186] A list of X-ray satellite missions, with links, can be found at http://imagine.gsfc.nasa.gov/docs/sats_n_data/xray_missions.html. There is a more technical version at http://heasarc.gsfc.nasa.gov/docs/observatories.html.

[187] *Chandra* website, http://chandra.harvard.edu/

[188] XMM–Newton website, http://sci.esa.int/xmm-newton/

[189] Suzaku technical description, http://www.astro.isas.jaxa.jp/suzaku/doc/suzaku_td/

[190] INTEGRAL website, http://www.rssd.esa.int/index.php?project=INTEGRAL&page=About_INTEGRAL

[191] *Swift* website, http://swift.gsfc.nasa.gov/.

[192] An explanation of the operation of microchannel plates, provided by a manufacturer but not (except at the end) specific to that company's products, can be found at http://www.dmphotonics.com/MCP_MCPImageIntensifiers/microchannel_plates.htm

[193] For *Chandra* instrumentation, see http://cxc.harvard.edu/cal/.

[194] For summary of XMM–Newton instrumentation, see http://xmm.esac.esa.int/external/xmm_user_support/documentation/uhb_2.1/node14.html.

[195] A brief description of the HXD, with diagram, is given in http://heasarc.gsfc.nasa.gov/docs/suzaku/analysis/abc/node10.html

[196] There are many articles and proprietary web pages on specific phoswich detectors. A fairly general article is WH Miller and M Diaz de Leon, www.osti.gov/scitech/servlets/purl/821938 (2003) (this paper has been published, but I can't find its formal citation).

[197] For a nice example of X-ray spectral lines from clusters of galaxies, see JS Sanders et al., *MNRAS* **402** (2010) 127–144.

[198] A comprehensive overview of X-ray sources can be found in the *Chandra* physics review paper H Tananbaum et al., arXiv:1405.7847 [astro-ph.HE] (2014) (to be published in *Reports on Progress in Physics*).

[199] These numbers come from S Ettori, www.xray.mpe.mpg.de/theorie/cluster/WLworkshop08/ettori.pdf (2008).

[200] D Clowe et al., *ApJ* **648** (2006) L109–L113.

[201] S Corbel et al., *MNRAS* **428** (2013) 2500–2515.

[202] F Panessa, www.sciops.esa.int/SD/ESACFACULTY/docs/seminars/040713_Panessa.pdf (2013).

[203] R Narayan, R Mahadevan and E Quataert, "Advection-Dominated Accretion Around Black Holes" in *Theory of Black Hole Accretion Discs* eds. MA Abramowicz, G Bjornsson & JE Pringle (Cambridge University Press, 1998);
arXiv astro-ph/9803141 (1998).

[204] F Yuan and R Narayan, *ARAA* **52** (2014); arXiv:1401.0586 [astro-ph.HE].

[205] J in 't Zand, `http://astrophysics.gsfc.nasa.gov/cai/coded.html` (1992–2006).

[206] Figure from ISDC Image Gallery,
`http://www.isdc.unige.ch/gallery.cgi?ISDC`

[207] IBIS coded mask from
`http://sci.esa.int/integral/49232-ibis-coded-mask/`

[208] WFC coded mask from
`http://www.asdc.asi.it/bepposax/software/wfc_faq.html`
Beppo-SAX web page, `http://www.asdc.asi.it/bepposax/`.

[209] FJ Ballesteros et al., *Proc. 2nd INTEGRAL Workshop* eds C Winkler, TJ-L Courvoisier and P Durouchoux (1997) 357–360.

[210] For an introduction to maximum entropy methods, see
PJ Steinbach, `http://cmm.cit.nih.gov/maxent/letsgo.html`

[211] For a tutorial on maximum likelihood methods, intended for psychologists but—perhaps because psychologists aren't expected to be well-trained in mathematics—clearly explained, see
In Jae Myung, *J. Math. Psych.* **47** (2003) 90–100.

[212] FJ Ballesteros, EM Muro and B Luque, *Exp. Astron.* **11** (2001) 207–222.

[213] IBIS web page,
`http://www.rssd.esa.int/index.php?project=INTEGRAL&page=About_INTEGRAL_IBIS`

[214] *Swift*–BAT web page,
`http://swift.gsfc.nasa.gov/about_swift/bat_desc.html`

[215] P Mészáros and N Gehrels, *Res. Astron. Astroph.* **12** (2012) 1139–1161.

[216] M Ackermann et al. (*Fermi*–LAT and *Fermi*–GBM Collaborations), *ApJ* **729** (2011) 114.

[217] D Band et al. (BATSE), *ApJ* **413** (1993) 281–292.

[218] N Gehrels, E Ramirez-Ruiz and DB Fox, *ARAA* **47** (2009) 567–617.

[219] P Mészáros and MJ Rees, arXiv:1401.3012 [astro-ph.HE] (2014).

[220] J Casey for the IceCube Collaboration, contribution to ICRC13, arXiv:1309.6979 [astro-ph.HE] (2013) 5–8;
M Richman for the IceCube Collaboration, contribution to ICRC13, arXiv:1309.6979 [astro-ph.HE] (2013) 48–51.

[221] S Adrián-Martínez et al. (ANTARES Collaboration), *A&A* **559** (2013) A9.

[222] BATSE website, http://www.batse.msfc.nasa.gov/batse/

[223] E Berger, *ARAA* **52** (2014) 43–105.

[224] SE Woosley and JS Bloom, *ARAA* **44** (2006) 507–556.

[225] COS-B web page, https://heasarc.gsfc.nasa.gov/docs/cosb/cosb.html.

[226] AGILE website, http://agile.asdc.asi.it/;
M Tavani et al. (AGILE Collaboration), *A&A* **502** (2009) 995–1013.

[227] WB Atwood et al. (*Fermi*–LAT Collaboration), *ApJ* **697** (2009) 1071–1102.

[228] An explanation of the operation of a spark chamber, with videos, can be found at
http://www.ep.ph.bham.ac.uk/DiscoveringParticles/detection/spark-chamber/.

[229] P Haefner, *J. Inst.* **5** (2010) C12050.

[230] R Rando, E Charles, S Digel and L Baldini,
http://www.slac.stanford.edu/exp/glast/groups/canda/lat_Performance.htm

[231] DJ Thompson, *Rep. Prog. Phys.* **71** (2008) 116901.

[232] For a summary for the genral public, with nice pictures, see
http://www.nasa.gov/mission_pages/GLAST/news/gamma-ray-census.html;
for the catalogue itself, see
http://fermi.gsfc.nasa.gov/ssc/data/access/lat/2yr_catalog/;
for the descriptive paper, see
PL Nolan et al. (*Fermi*–LAT Collaboration), *ApJSS* **199** (2012) article id 31.

[233] Meng Su, TR Slatyer and DB Finkbeiner, *ApJ* **724** (2010) 1044–1082.

[234] K Berlöhr, http://www.mpi-hd.mpg.de/hfm/CosmicRay/Showers.html (1999).

[235] A clear explanation of the technique, with diagrams, can be found on the H.E.S.S. website,
https://www.mpi-hd.mpg.de/hfm/HESS/pages/about/telescopes/.

[236] VERITAS website, http://veritas.sao.arizona.edu/.

[237] The properties of the Dirac delta function are summarised by Wolfram MathWorld at
mathworld.wolfram.com/DeltaFunction.html.

[238] There are various derivations of this equation available on the web. A convenient one is
R Fitzpatrick,
farside.ph.utexas.edu/teaching/em/lectures/node47.html (2006).
Wikipedia also has it (see "Maxwell's equations"), and there is a nice summary of Maxwell's equations in the form of lecture notes for the CERN accelerator summer school 2010,
A Wolski, cas.web.cern.ch/cas/Denmark-2010/Lectures/Wolski-1.pdf (2010).
Fitzpatrick, Wolski and Longair[171] all work in SI units. Be aware

that most astronomers do not: they use cgs, and the form of Maxwell's equations in cgs is somewhat different.

[239] The function $\sin\theta/\theta$, which you should have met in optics (it describes diffraction from a single slit) is known as the **sinc function**, and is closely related to the Dirac delta function. It is described in Wolfram MathWorld,
`mathworld.wolfram.com/SincFunction.html`,
and the Wikipedia article,
`en.wikipedia.org/wiki/Sinc_function`.

[240] STACEE website,
`http://www.astro.ucla.edu/ stacee/staceehowitworks.html`.

[241] Kaye and Laby online: KP Birch,
`http://www.kayelaby.npl.co.uk/general_physics/2_5/2_5_7.html`.

[242] Sable Systems International,
`http://www.sablesys.com/baro-altitude.html`.

[243] H Bichsel, DE Groom and SR Klein, "Passage of particles through matter" in J Beringer et al. *Phys. Rev.* **D86** (2012) 010001; `pdg.lbl.gov`.

[244] J Linsley, *Proc. 19th ICRC* **7** (1985) 163–166.

[245] H.E.S.S. website,
`http://www.mpi-hd.mpg.de/hfm/HESS/pages/press/2012/HESS_II_first_light/`

[246] Whipple telescope page on VERITAS website,
`https://veritas.sao.arizona.edu/whipple-10m-topmenu-117`

[247] MAGIC website, `https://magic.mpp.mpg.de/`

[248] CTA website, `https://portal.cta-observatory.org/Pages/Home.aspx`

[249] See, for example, C van Eldik for the H.E.S.S. Collaboration, *J. Phys. Conf. Ser.* **110** (2008) 062003 (arXiv:0709.3729 [astro-ph]).

[250] See, for example, the SST-GATE project of the Observatoire de Paris, `http://gate.obspm.fr/`, which is investigating a novel optical design (Schwarzschild-Couder optics) for the small CTA telescope. This design, being cheaper to construct, would allow more telescopes to be built for a given budget.

[251] See, for example,
F Aharonian et al. (H.E.S.S. Collaboration), *A&A* **457** (2006) 899–915.

[252] For an introduction to boosted decision trees, see
BP Roe et al., *NIMPA* **543** (2005) 577–584.
Boosted decision trees are used in recent H.E.S.S. papers, e.g.
A Abramowski et al. (H.E.S.S. Collaboration), *MNRAS* **441** (2014) 790–799.

[253] F Aharonian et al. (H.E.S.S. Collaboration), *Phys. Rev. Lett.* **101** (2008) article id 261104.

[254] F Aharonian et al. (H.E.S.S. Collaboration), *Phys. Rev.* **D75** (2007) article id 042004.

[255] D Staszak for VERITAS, TEVPA conference 2013,
`http://www.physics.mcgill.ca/∼staszak/Staszak_TeVPA2013_v2.pdf`

[256] D Borla Tridon for MAGIC, 32nd ICRC (2011); arXiv:1110.4008 [astro-ph.HE].

[257] S Wakely and D Horan, `http://tevcat.uchicago.edu/` (version 3.400).

[258] BL Lac objects, named for the prototype BL Lacertae, are active galactic nuclei whose spectra display weak or non-existent emission lines; it is believed that our line of sight is closely aligned with the radio jet. BL Lacs and highly variable flat-spectrum radio quasars are classified together as "blazars". A brief introduction, with emphasis on γ-ray emission, can be found in
J Finke, *Proc. 2012 Fermi Symp.*(eConf C121028);
arXiv:1303.5095 [astro-ph.HE] (2013).

[259] A Abramowski et al. (H.E.S.S. collaboration), *A&A* **564** (2014) A9.

[260] F Aharonian et al. (H.E.S.S. collaboration), *Nature* **440** (2006) 1018–1021.

[261] For a recent review of solar neutrinos, see
WC Haxton, RGH Robertson and AM Serenelli, *ARAA* **51** (2013) 21–61.

[262] Literally hundreds of papers have been published on the neutrinos from SN 1987A. For a contemporary review of the supernova, including the neutrino signal, see
WD Arnett, JN Bahcall, RP Kirshner and SE Woosley, *ARAA* **27** (1989) 629–700.

[263] A full explanation of this requires the use of Clebsch-Gordan coefficients, which aren't covered in PHY304. A nice course on group theory in quantum mechanics and particle physics is
G 't Hooft, `http://www.staff.science.uu.nl/ hooft101/lectures/lieg07.pdf` (2007);
the branching ratio of the Δ is discussed in chapter 8.

[264] MG Aartsen et al. (IceCube Collaboration), *Phys. Rev.* **D89** (2014) 102004.
The predictions and data shown in the plot are from (IceCube publications unless otherwise stated):
Science **342** (2013) 1242856 ("IC-79 and IC-86");
Phys. Rev. **D83** (2011) 012001 ("IC-40");
Phys. Rev. Lett. **110** (2013) 151105 ("IC-79");
R. Abbasi et al. (AMANDA-II), *Phys. Rev.* **D79** (2009) 102005;
Y. Ashie et al. (Super-Kamiokande), *Phys. Rev.* **D71** (2005) 112005;
D Chirkin and W Rhode, *Proc. 27th ICRC, Hamburg* (2001) 1017–1020;
TK Gaisser, *Ap. Phys.* **35** (2012) 801–806 ("HKKM");
A. Schukraft, *Nucl. Phys. Proc. Suppl.* **237–238** (2013) 266–268 ("ERS");
Waxman and Bahcall see below.

[265] Particle Data Group,
`http://pdg.lbl.gov/2013/hadronic-xsections/rpp2013-gammap_total.dat`
(2013).

[266] E Waxman and JN Bahcall, *Phys. Rev.* **D59** (1999) 023002.

[267] MG Aartsen et al. (IceCube Collaboration), *Phys. Rev. Lett.* (accepted); arXiv:1405.5303 (2014).

[268] JA Formaggio and GP Zeller, *Rev. Mod. Phys.* **84** (2012) 1307.

[269] A Connolly, RS Thorne and D Waters, *Phys. Rev.* **D83** (2011) 113009.

[270] The Glashow resonance was predicted by Sheldon Glashow back in 1960 (*Phys. Rev.* **118** 316–317)—long before the invention of electroweak theory and the prediction of the W and Z (in the 1960 paper, Glashow talks about "the boson mediating muon decay", which would be the W, but assumes that its mass might be comparable to that of the kaon!). For a modern discussion in the context of neutrino telescopes, see
LA Anchordoqui, H Goldberg, F Halzen and TJ Weiler *Phys. Lett.* **B621** (2005) 18–21.

[271] There are several test programmes for acoustic detection: see for example
V Niess and V Bertin, *Ap. Phys.* **26** (2006) 243–256;
J Vandenbroucke, G Gratta, and N Lehtinen, *ApJ* **621** (2005) 301–312.

[272] Super-Kamiokande website,
`http://www-sk.icrr.u-tokyo.ac.jp/sk/index-e.html`.

[273] Lake Baikal Neutrino Telescope website, `http://baikalweb.jinr.ru/`. Note: this appears to be very out-of-date. There are more recent publications from the Baikal group. A more up-to-date status report can be found in
A Avronin et a. (Baikal Collaboration), *J. Phys. Conf. Series* **409** (2013) 012141.

[274] ANTARES website, `http://antares.in2p3.fr/`.

[275] SNO website, `http://www.sno.phy.queensu.ca/`.

[276] R Abbasi et al. (IceCube Collaboration) *Ap. Phys.* **35** (2012) 615–624.

[277] S Adrián-Martínez et al. (ANTARES Collaboration), *ApJL* **786** (2014) L5.

[278] R Abbasi et al. (IceCube Collaboration) *Nature* **484** (2012), 351–354.

[279] D Guetta et al., *Ap. Phys.* **20** (2004) 429.

[280] S Hümmer, P Baerwald and W Winter, *Phys. Rev. Lett.* **108** (2012) 231101.

[281] KM3NeT website, `http://www.km3net.org/home.php`.

[282] BM Gaensler and PO Slane, *ARAA* **44** (2006) 17–47.

[283] G Dubus, *A&Ap. Rev.* **21** (2013) 64.

[284] M Meyer, D Horns and H-S Zechlin, *A&A* **523** (2010) A2.

[285] E Amato, *Int. J. Mod. Phys. Conf. Series* **28** (2014) 1460160.

[286] O Krause et al., *Nature* **456** (2008) 617–619.

[287] F Giordano et al., *ApJL* **744** (2012) A2.

[288] DC Ellison, P Slane, DJ Patnaude and AM Bykov, *ApJ* **744** (2012) 39.

[289] Three recent discussions of blazar classification are
G Ghisellini, F Tavecchio, L Foschine and G Ghirlanda, *MNRAS* **414** (2011) 2674–2689
P Giommi et al., *MNRAS* **420** (2012) 2899–2911
G Ghisellini, in "The Innermost Regions of Relativistic Jets and Their Magnetic Fields, Granada, Spain" ed. JL Gómez; *EPJ Web of Conferences* **61** (2013) 05001; arXiv:1309.4772 [astro-ph.HE].
Most of the information in the text comes from Giommi et al.

[290] For a typical quasar spectrum in the optical and UV, see the composite created from the Sloan Digital Sky Survey,
DE Vanden Berk et al., *AJ* **122** (2001) 549–564.

[291] A summary of the Fanaroff-Riley classification can be found at
`http://ned.ipac.caltech.edu/level5/Glossary/Essay_fanaroff.html`.
The original paper is BL Fanaroff and JM Riley, *MNRAS* **167** (1974) 31P–36P.

[292] CM Urry and P Padovani, *PASP* **107** (1995) 803–845.

[293] M Ackermann et al. (*Fermi*–LAT Collaboration), *ApJ* **755** (2012) 164.

[294] M Ackermann et al. (*Fermi*–LAT Collaboration), *ApJSS* **209** (2013) 11.

[295] A Melandri et al., *A&A* **567** (2014) A29.

[296] J Hjorth, *Phil. Trans. Roy. Soc.* **A371** (2013) 20120275.

[297] J Heise, J in 't Zand, M Kippen and P Woods, *AIP Conf. Proc.* **662** (2003) 229–236.

[298] E Pian et al., *Nature* **442** (2006) 1011–1013.

[299] J Hjorth and JS Bloom, arXiv:1104.2274 [astro-ph.HE] (2011);
Chapter 9 in *Gamma-Ray Bursts*, Cambridge Astrophysics Series **51**, eds C Kouveliotou, RAMJ Wijers and S Woosley, Cambridge University Press (Cambridge, 2012).

[300] S Chakraborti et al., arXiv:1402.6336 [astro-ph.HE] (2014).

[301] SE Woosley, arXiv:1105:4193 [astro-ph.HE] (2011);
Chapter 10 in *Gamma Ray Bursts*, op. cit.

[302] See, for example, `http://www.ligo.org/science/GW-Sources.php`;
`http://www.ligo.caltech.edu/advLIGO/scripts/ref_des.shtml`.

[303] See, for example,
`http://uspas.fnal.gov/materials/14UNM/UNM_Fundamentals.shtml`

[304] E Fermi, *Phys. Rev.* **75** (1949) 1169–1174.

[305] See, for example, Wikipedia:
`en.wikipedia.org/wiki/Fick's_laws_of_diffusion`.

[306] R Blandford and D Eichler, *Phys. Rep.* **154** (1987) 1–75.

[307] JD Richardson et al., *Nature* **454** (2008) 63–66.

[308] A Balogh and RA Treumann, *Physics of Collisionless Shocks* (New York, Springer 2013), section 2.1.

[309] D Ellison, "Collisionless shocks and particle acceleration in astrophysics" (APS Colloquium, June 1, 2011)
`http://viavca.in2p3.fr/don_ellison.html`.

[310] Gauss's theorem, or the divergence theorem, states that

$$\int \nabla \cdot \mathbf{F} \mathrm{d}^3 V = \oint \mathbf{F} \cdot \mathbf{n} \mathrm{d}^2 S,$$

where V is a volume and S is the surface surrounding the volume. If $f_m \to 0$ as $v \to \infty$, the integrand on the right-hand side is zero throughout, so the integral on the left must also be zero.

[311] See, for example, Wikipedia,
`http://en.wikipedia.org/wiki/`
`Shocks_and_discontinuities_%28magnetohydrodynamics%29`

[312] See the brief summary in the introduction of
M Takamoto and JG Kirk, paper presented at the 41st EPS Conference on Plasma Physics, Berlin (2014);
`http://ocs.ciemat.es/EPS2014PAP/pdf/P1.155.pdf`.

[313] PF Winkler et al., *ApJ* **781** (2014) 65, 18.

[314] F Acero et al. (H.E.S.S. Collaboration), *A&A* **516** (2010) A62.

[315] JH Croston et al., *MNRAS* **395** (2009) 1999–2012.

[316] F Aharonian et al. (H.E.S.S. Collaboration), *ApJL* **695** (2009) L40–L44.

[317] S Wykes et al., *A&A* **558** (2013) A19.

[318] DC Ellison and R Ramaty, *ApJ* **298** (1985) 400–408.

[319] The literature on diffusive shock acceleration is huge. For a recent simulation with nice pictures, see
D Caprioli and A Spitkovsky, *ApJ* **783** (2014) 91 (Paper I); arXiv:1401.7679 [astro-ph.HE] (Paper II); arXiv:1407.2261 [astro-ph.HE] (Paper III).

[320] P Blasi, *A&A Rev.* **21** (2013) 70.

[321] J Ballet, *Adv. Sp. Res.* **37** (2006) 1902–1908.

[322] L Ball and DB Melrose, *Publ. Astr. Soc. Aus.* **18** (2001) 361–373.

[323] A convenient summary of the classification of radio bursts is prvided by the Australian Government's space weather service at
`http://www.ips.gov.au/World_Data_Centre/1/9/5`.

[324] JG Kirk and P Duffy, *J. Phys.* **G25** (1999) R163–R194.

[325] RG Blandford and CF McKee, *Phys. Fl.* **19** (1976) 1130–1138.

[326] A Achterberg, YA Gallant, JG Kirk and AW Guthmann, *MNRAS* **328** (2001) 393–408.

[327] M Lemoine and G Pelletier, arXiv:1111.7110 [astro-ph.HE] (2011).

[328] There is a nice animation of this on the Wikipedia page, `http://en.wikipedia.org/wiki/Magnetic_reconnection`.

[329] T Howard, *Coronal Mass Ejections: An Introduction* (New York; Springer, 2011), section 8.3.

[330] EM de Gouveia Dal Pino, G Kowal and A Lazarian, Proc. 33rd ICRC, 2–9 July 2013, Rio de Janeiro, Brazil; arXiv:1401.4941 [astro-ph.HE] (2014).

[331] B Cerutti, DA Uzdensky and MC Begelman, *ApJ* **746** (2012) 148.

[332] L Sironi and A Spitkovsky, *ApJL* **783** (2014) L21.

[333] The magnetohydrodynamics of this is well outside the scope of this course. For a brief description, see the Wikipedia article, `http://en.wikipedia.org/wiki/Magnetorotational_instability`.

[334] M Hoshino, *ApJ* **773** (2013) 118.

[335] E Amato, *Int. J. Mod. Phys.* **D23** (2014) 1430013; arXiv:1406.7714 [astro-ph.HE].

[336] D Caprioli, *JCAP* **1207** (2012) 038.

[337] AR Bell, *MNRAS* **182** (1978) 147–156.

[338] AM Hillas, *ARAA* **22** (1984) 425–444.

[339] G Sigl, arXiv:1202.0466 [astro-ph.HE] (2012).

[340] M Scholer, *Adv. Sp. Res.* **4** (1984) 419–421.

[341] T Terasawa, *Proc. IAU Symposium* **274** (2010) 214–219.

[342] For space weather, see for example `http://www.swpc.noaa.gov/SWN/`.

[343] For general information about the Earth's Van Allen belts, see Wikipedia: `http://en.wikipedia.org/wiki/Van_Allen_radiation_belt`.

[344] R Vainio, *Proc. IAU Symposium* **257** (2008) 413–423.

[345] SE Forbush, *Phys. Rev.* **70** (1946) 771–772.

[346] G Swarup, PH Stone and A Maxwell, *ApJ* **131** (1960) 725–738.

[347] MI Desai and D Burgess, *J. Geophys. Res.* **113** (2008) article A00B06.

[348] CT Russell, in *Collisionless Shocks in the Heliosphere: Reviews of Current Research*, Geophysical Monograph **35** (1985).

[349] RA Treumann and CH Jaroschek, arXiv:0808.1701 [astro-ph] (2008)

[350] CLUSTER website, `http://sci.esa.int/cluster/`.

[351] STEREO website, `www.nasa.gov/mission_pages/stereo/main/`.

[352] *Wind* website, `http://science.nasa.gov/missions/wind/`.

[353] H Kucharek et al., *Annales Geophysicae* **22** (2004) 2301–2308.

[354] JY Chaufray et al., *J. Geophys. Res.* **112** (2007) E9009.

[355] EC Stone et al., *Nature* **454** (2008) 71–74.

[356] RA Treumann and CH Jaroschek, arXiv:0808.4170 [astro-ph] (2008).

[357] J Giacolone and R Decker, *ApJ* **710** (2010) 91–96.

[358] NA Schwadron, MA Lee and DJ McComas, *ApJ* **675** (2008) 1584–1600.

[359] UK Senanayake and V Florinski, *ApJ* **778** (2013) 122.

[360] JF Drake et al. *ApJ* **709** (2010) 963–974.

[361] W Baade and F Zwicky, *PNAS* **20** (1934) 254–259, 259–263.

[362] G de Vaucouleurs and HG Corwin Jr, *ApJ* **295** (1985) 287–304.

[363] Between September 1909 and October 1919, 16 novae were reported in M31 (the papers can be found by a title search in `adsabs`). Their reported magnitudes range from 15.7 to 17.9, with a mean of 17.03 ± 0.15. These are photographic (B) magnitudes; as de Vaucouleurs and Corwin conclude that S And was very red, its peak photographic magnitude was probably about 7.

[364] AV Filippenko, *ARAA* **35** (1997) 309–355.

[365] E Hubble, *ASP Leaflets* **14** (1928) 55–58.

[366] NU Mayall and JH Oort, *PASP* **54** (1942) 95–104.

[367] R Diehl et al. *Nature* **439** (2006) 45–47.

[368] EA Helder et al., *Sp. Sci. Rev.* **173** (2012) 369–431.

[369] For a general review of X-ray properties of supernova remnants, see J Vink, *A&A Rev.* **20** (2012) 49.

[370] GB Field, DW Goldsmith and HJ Habing, *ApJ* **155** (1969) L149–L154.

[371] JK Truelove and CF McKee, *ApJSS* **120** (1999) 299–326.

[372] DA Green, *A Catalogue of Galactic Supernova Remnants (2014 May version)*, arXiv:1409.0637 [astro-ph.HE]; `http://www.mrao.cam.ac.uk/surveys/snrs/` (2014).

[373] SP Reynolds, *Ap. & Sp. Sci.* **336** 257–262.

[374] EM Reynoso, JP Hughes and DA Moffett, *AJ* **145** (2013) 104.

[375] Chandra images of SN 1006, `http://chandra.harvard.edu/photo/2013/sn1006/`, credit NASA/CXC/Middlebury College/F Winkler (2013).

[376] W Reich, *Proc. 270th WE-Hereaus Seminar on Neutron Stars, Pulsars and Supernova Remnants*, eds W Becker, H Lesch and J Trümper (2002); arXiv:astro-ph/0208498.

[377] JWM Baars et al., *A&A* **61** (1977) 99–106.

[378] HJ Völk, arXiv:astro-ph/0603502 (2006).

[379] G Cassam-Chenaï et al., *ApJ* **680** (2008) 1180–1197.

[380] M Pohl, H Yan and A Lazarian, *ApJ* **626** (2005) L101–L104.

[381] G Cassam-Chenaï et al., *ApJ* **665** (2007) 315–340.

[382] V Zirakashvili, *J. Phys. Conf. Series* **409** (2013) 012012.

[383] KA Eriksen et al., *ApJ* **728** (2011) L28.
Colour image from `http://chandra.harvard.edu/photo/2011/tycho/`.

[384] JW Hewitt at al. for *Fermi*–LAT, Proc. 33rd ICRC, 2–9 July 2013, Rio de Janeiro, Brazil; arXiv:1307.6570 [astro-ph.HE] (2013).

[385] See, for example, figures 3 and 4 of
M Mandelartz and J Becker Tjus, arXiv:1301.2437 [astro-ph.GA] (2013).

[386] M Ackermann et al. (*Fermi*–LAT Collaboration), *Science* **339** (2013) 807–811.

[387] JG Kirk, Y Lyubarsky and J Petri, in *Neutron stars and pulsars, 40 years after the discovery*, ed. W Becker (Springer, 2007); arXiv:astro-ph/0703116.

[388] O Kargaltsev, B Rangelov and GG Pavlov, arXiv:1305.2552 [astro-ph.HE] (2013).

[389] DA Frail, NE Kassim, TG Cronwell and WM Goss, *ApJ* **454** (1995) L129–L132.

[390] JG Bolton and GJ Stanley, *Aus. J. Sci. Res.* **2** (1949) 139–148.

[391] DA Green and FR Stephenson, in *Supernovae and Gamma Ray Bursters*, ed. KW Weiler (Springer, 2003); arXiv:astro-ph/0301603.

[392] RB Lovelace, JM Sutton and HD Craft *IAU Circ.* **2113** (1968) 1.

[393] , *Proc. IAU Symposium* **291** (2013) 195–198.

[394] J Arons and M Tavani, *ApJSS* **90** (1994) 797–806.

[395] P Slane, in *Proc. XII Moriond Astrophys. mtg "The Gamma-Ray Universe"* eds A Goldwurm, D Neumann, and J Tran Than van (2002); arXiv:astro-ph/0205481.

[396] RA Chevalier, *Mem. SAI* **69** (1998) 977–987.

[397] G Zanardo et al., *ApJ* **796** (2014) 82.

[398] `http://chandra.harvard.edu/photo/2005/g21/` (2005); credit NASA/CXC/U. Manitoba/H. Matheson & S. Safi-Harb.

[399] `http://chandra.harvard.edu/photo/2004/snr0540/` (2004); credit NASA/CXC/SAO.

[400] A De Luca, *AIP Conf. Proc.* **983** (2008) 311–319.

[401] See `http://chandra.harvard.edu/photo/2003/b1957/` (2003).

[402] `http://www.mpi-hd.mpg.de/hfm/HESS/pages/home/som/2006/05/`

[403] `http://sci.esa.int/integral/49889-high-energy-emission-from-the-vela-pulsar-wind-nebula/`.

[404] J Pétri and Y Lyubarsky, *A&A* **473** (2007) 683–700.

[405] J van Paradijs et al., *Nature* **386** (1997) 686–689.

[406] P Kumar and B Zhang, *Phys. Rep.* **561** (2015) 1–109.

[407] E Nakar, T Piran and J Granot, *ApJ* **579** (2002) 699–705.

[408] LSST website, `http://www.lsst.org/lsst/`.

[409] JT Bonnell (NASA/GSFC),
`https://heasarc.gsfc.nasa.gov/docs/objects/grbs/grb_profiles.html`

[410] Ying Qin et al., *ApJ* **763** (2013) 15.

[411] N Tanvir et al., *Nature* **461** (2009) 1254–1257.

[412] Hou-jun Lü et al., *ApJ* **725** (2010) 1965–1970.

[413] G Cavallo and MJ Rees, *MNRAS* **183** (1978) 359–365;
B Paczýnski, *ApJ* **308** (1986) L43–L46;
J Goodman, *ApJ* **308** (1986) L47–L50.

[414] P Mészáros, *Rep. Prog. Phys.* **69** (2006) 2259–2322.

[415] B Zhang, *Int. J. Mod. Phys.* **D23** (2014) 1430002.

[416] N Gehrels and S Razzaque, *Front. Phys.* **8** (2013) 661–678.

[417] K Asano, S Guiriec and P Mészáros, *ApJ* **705** (2009) L191.

[418] R Narayan, T Piran and P Kumar, *ApJ* **557** (2001) 949–957.

[419] I Zalamea and AM Beloborodov, *MNRAS* **410** (2011) 2302–2308.

[420] RD Blandford and RL Znajek, *MNRAS* **179** (1977) 433–456.

[421] S Terrade,
`hmf.enseeiht.fr/travaux/CD0001/travaux/optmfn/hi/01pa/hyb72/kh/kh_theo.htm`
(2001).

[422] H-J Lü and B Zhang, *ApJ* **785** (2014) 74.

[423] BD Metzger et al., *MNRAS* **413** (2011) 2031–2056.

[424] SB Cenko et al., *ApJ* **711** (2010) 641–654.

[425] G Chincarini et al., *MNRAS* **406** (2010) 2113–2148.

[426] T Sakamoto et al., *ApJ* **679** (2008) 570–586.

[427] J Greiner et al., *Nature* **523** (2015) 189–192.

[428] K Wiersema et al., *GCN* **13276** (2012) 1.

[429] A Melandri et al., *A&A* **547** (2012) A87.

[430] Z Cano et al., *A&A* **568** (2014) A14.

[431] D Xu et al., *ApJ* **776** (2013) 98.

[432] S Schulze et al., *GCN* **14994** (2013) 1.

[433] V D'Elia et al., *A&A* **577** (2015) A116.

[434] S Klose et al., *GCN* **15320** (2013) 1.

[435] AJ Levan et al., *ApJ* **781** (2014) 13.

[436] SD Barthelmy, *Phil. Trans. Roy. Soc. A* **365** (2007) 1281–1291.

[437] MHPM van Putten et al., *MNRAS* **444** (2014) L58–L62.

[438] FJ Virgili et al., *ApJ* **778** (2013) 54.

[439] JPU Fynbo, D Malesani and P Jakobsson, chapter 13 (pp 269–290) of *Gamma-Ray Bursts*, eds C Kouvellotou, RAMJ Wijers and SE Woosley, Cambridge Astrophysics Series **51** (Cambridge University Press, 2012); arXiv:1301.4908 [astro-ph.CO].

[440] T Krühler et al., *A&A* in press (2015); arXiv:1505.06743 [astro-ph.GA].

[441] S Schulze et al., *ApJ* in press (2015); arXiv:1503.04246 [astro-ph.GA].

[442] D Maoz, F Mannucci and G Neiemans, *ARAA* **52** (2014) 107–170.

[443] BMS Hansen and ES Phinney, *MNRAS* **291** (1997) 569–577.

[444] J Abadie et al. (LIGO Collaboration), *Class. Quant. Grav.* **27** (2010) 173001.

[445] J Camp et al., *Exp. Ast.* **36** (2013) 505–522.

[446] NR Tanvir et al., *Nature* **500** (2013) 547–549.

[447] B Yang et al., *Nature Comm.* **6** (2015) 7323.

[448] Z-P Jin et al., arXiv:1507.07206 [astro-ph.HE] (2015).

[449] MG Aartsen et al. (IceCube Collaboration) *ApJ* **805** (2015) L5.

[450] P Baerwald, M Bustamente and W Winter, *Ap. Phys.* **62** (2015) 66–91.

[451] V Berezinsky, *EPJ Web of Conferences* **53** (2013) 01003; arXiv:1307.4043 [astro-ph.HE].

[452] M Ackermann et al. (*Fermi*–LAT), *ApJ* **810** (2015) 14.

[453] The first instances of the term in an ADSABS title search are VA Ambartsumian, *Proc. 13th Conf. Phys. Univ. Brussels*, ed. R Stoops (1965) 1, and VA Ambartsumian, *Proc. IAU Symp.* **29** (1968) 11–20; the latter is in Russian, and I can't find a copy of the former, so what he actually said about activity in galactic nuclei I do not know.

[454] WN Brandt and G Hasinger, *ARAA* **43** (2005) 827–859.

[455] R Mushotzky, *Supermassive Black Holes in the Distant Universe*, ed. AJ Barger (Kluwer Academic Publishers, 2004), Chapter 2; arXiv:astro-ph/0405144.

[456] PD Barthel, *ApJ* **336** (1989) 606–611.

[457] R Antonucci and R Barvainis, *ApJ* **363** L17–L20.

[458] R Antonucci, *ARAA* **31** (1993) 473–521.

[459] See, e.g.,
MJ Hardcastle, DA Evans and JH Croston, *MNRAS* **376** (2007) 1849–1856,
PN Best and TM Heckman, *MNRAS* **421** (2012) 1569–1582.

[460] TM Heckman and PN Best, *ARAA* **52** (2014) 589–660.

[461] H Netzer, *ARAA* **53** (2015) 365–408.

[462] See, e.g., C Tadhunter, *New Astr. Rev.* **52** (2008) 227–239.

[463] BL Fanaroff and JM Riley, *MNRAS* **167** (1974) 31P–35P.

[464] RA Laing et al., *ASP Conf. Series* **54** (1994) 201–208.

[465] RG Hine and MS Longair, *MNRAS* **188** (1979) 111–130.

[466] MJ Hardcastle, *MNRAS* **405** (2010) 2810–2816.

[467] S van Velzen et al., *A&A* **544** (2012) A18.

[468] J Kormendy and LC Ho, *ARAA* **51** (2013) 511–653.

[469] See, for example, EE Salpeter, *ApJ* **140** (1964) 796–800.

[470] J Frank, A King and DJ Raine, *Accretion Power in Astrophysics*, 3rd edition (Cambridge University Press, 2002).

[471] JE Pringle and AR King, *Astrophysical Flows* (Cambridge University Press, 2007).

[472] A Merloni et al., *MNRAS* **437** (2014) 3550–3567.

[473] RH Becker et al., `http://sundog.stsc.edu` (2014).

[474] JJ Condon et al., `nttp://www.cv.nrao.edu/nvss` (2012).

[475] See, e.g., R Singh et al., *A&A* **558** (2013) A43.

[476] RJ Janssen et al., *A&A* **541** (2012) A62.

[477] B Mingo et al., *MNRAS* **440** (2014) 269–297.

[478] CS Reynolds, *Sp. Sci. Rev.* **183** (2014) 277–294.

[479] C Tadhunter at al., *MNRAS* **445** (2014) L51–L55.
See also C Ramos Almeida et al., *MNRAS* **419** (2012) 687–705.

[480] See, e.g., RP Fender, *Lect. Notes Phys.* **794** (2010) 115–142.

[481] RJH Dunn et al., *MNRAS* **403** (2010) 61–82.

[482] R Narayan, IV Igumenschchev and MA Abramowicz, *Publ. Astro. Soc. Japan* **55** (2003) L69–L72.

[483] M Sikora and MC Begelman, *ApJ* **764** (2013) L24.

[484] R Moll, *A&A* **507** (2009) 1203–1210.

[485] See, for example,
RA Laing and AH Bridle, *MNRAS* **336** (2002) 1161–1180 (modelling 3C 31) and
S Wykes et al., *A&A* **558** (2013) A19 (modelling Cen A).

[486] See, e.g., G Ghisellini, F Tavecchio and M Chiaberge, *A&A* **432** (2005) 401–410.

[487] DC Gabuzda, AR Reichstein and EL O'Neill, *MNRAS* **444** (2014) 172–184.

[488] AP Lobanov, *A&A* **330** (1998) 76–89.

[489] M Zamaninasab et al., *Nature* **510** (2014) 126–128.

[490] G Cavallo, *A&A* **65** (1978) 415–419.

[491] SS Doeleman et al., *Science* **338** (2012) 355–358.

[492] F Aharonian et al. (H.E.S.S. Collaboration), *ApJ* **695** (2009) L40–L44.

[493] M Kachelrieß, S Ostapchenko and R Tomàs, *New J. Phys.* **11** (2009) 065017.

[494] M Kachelrieß, S Ostapchenko and R Tomàs, *Pub. Ast. Soc. Aus.* **27** (2010) 482–489.

[495] JL Goodger et al., *ApJ* **708** (2010) 675–697.

[496] AA Abdo et al. (*Fermi*–LAT Collaboration) *Science* **328** (2010) 725–729.

[497] Ł Stawarz et al., *ApJ* **766** (2013) 48.

[498] NRAO, `http://www.nrao.edu/pr/1999/m87big/layout.jpg` (1999).

[499] *Chandra*, `http://chandra.harvard.edu/photo/2001/0134/more.html` (2001).

[500] SG Jorstad et al., *AJ* **130** (2005) 1416–1465.

[501] The NASA/IPAC Extragalactic Database, `https://ned.ipac.caltech.edu`.

[502] SG Jorstad et al., *ApJ* **773** (2013) 147.

[503] R Schopper, H Lesch and GT Birk, *A&A* **335** (1998) 26–32.

[504] D Giannios, *MNRAS* **408** (2010) L46–L50.

[505] G Kowal, EM de Gouviea Dal Pino and A Lazarian, *ApJ* **735** (2011) 102.

[506] L Sironi, M Petropoulou and D Giannios, *MNRAS* **450** (2015) 183–191.

[507] MJ Hardcastle et al., *MNRAS* **393** (2009) 1041–1053.

[508] A Aab et al. (Pierre Auger Observatory), *Contributions to the 34th International Cosmic Ray Conference*, arXiv:1509.03732 [astro-ph.HE] (2015) 18–25.

[509] GR Farrar et al., *JCAP* **1301** (2013) 023.

[510] A Aab et al. (Pierre Auger Observatory), *Contributions to the 34th International Cosmic Ray Conference*, arXiv:1509.03732 [astro-ph.HE] (2015) 41–53.

[511] RU Abbasi et al. (Telescope Array), *Ap. Phys.* **64** (2015) 49–62.

[512] C Konar and MJ Hardcastle, *MNRAS* **436** (2013) 1595–1614.

[513] JH Croston et al., *ApJ* **626** (2005) 733–747.

[514] AA Abdo et al. (*Fermi*–LAT), *ApJ* **719** (2010) 1433–1444.

[515] S Sahu, B Zhang and N Fraija, *Phys. Rev.* **D85** (2012) 043012.

[516] JC Joshi and N Gupta, *Phys. Rev.* **D87** (2013) 023002.

[517] MG Aartsen et al. (IceCube Collaboration), *ApJ* **796** (2014) 109.

[518] T Gl usenkamp for the IceCube Collaboration, talk given at the Roma International Conference on Astroparticle Physics: arXiv:1502.03104 [astro-ph.HE] (2015).

[519] IB Jacobsen et al., *MNRAS* **451** (2015) 3649–3663.

[520] HBJ Koers and P Tinyakov, *Phys. Rev.* **D78** (2008) 083009.

[521] JK Becker and PL Biermann, *Ap. Phys.* **31** (2009) 138–148.

[522] S Yoshida and H Takami, *Phys. Rev.* **D90** (2014) 123012.

[523] MG Aartsen et al. (IceCube Collaboration), *ApJ* **809** (2015) 98.

[524] DR Bergman and JW Belz, *J. Phys. G* **34** (2007) R359–R400.

[525] R Johnson and GR Farrar, *ApJ* **757** (2012) 14.

[526] P Blasi and E Amato, *JCAP* **1201** (2012) 010.

[527] G Giacinti, M Kachelrießand DV Semikoz, *Phys. Rev.* **D91** (2014) 083009.

[528] G Giacinti et al., arXiv:1507.07534 [astro-ph.HE] (2015).

[529] D Allard, O Parizot and AV Olinto, *Ap. Phys.* **27** (2007) 61–75.

[530] L O'C Drury, *Ap. Phys.* **39–40** (2012) 52–60.

[531] P Blasi, *C.R. Phys. 15* (2014) 329–338.

[532] D Harari, *C.R. Phys. 15* (2014) 376–383.

[533] K Kotera and AV Olinto, *ARAA* **49** (2011) 119–153.

[534] MG Aartsen et al. (IceCube Collaboration), *Phys. Rev.* **D88** (2013) 112008.

[535] A Abramowski et al. (H.E.S.S. collaboration), *MNRAS* **439** (2014) 2828–2836.

Index