Gamma-Ray Bursts: theoretical issues and developments

P. Mészáros¹,²,³

¹ Department of Astronomy & Astrophysics, Pennsylvania State University, University Park, PA 16802, USA, e-mail: nnp@psu.edu
² Department of Physics, Pennsylvania State University, University Park, PA 16802, USA
³ Center for Particle and Gravitational Astrophysics, Institute for Gravitation and the Cosmos, Pennsylvania State University, University Park, PA 16802, USA

Abstract. I discuss some aspects of the evolution of the standard GRB model, emphasizing various theoretical developments in the last decade, and review the impact of some of the most recent observational discoveries and the new challenges they pose in the expanding realm of multi-messenger astrophysics.

Key words. Gamma-Ray Bursts, Theory

1. Introduction

Fireballs in astrophysics generally refer to an optically thick plasma whose temperature exceeds the electron rest mass and which can produce $e^\pm$ pairs and photons in equilibrium with a baryonic plasma. An early study of the fireball radiation physics aimed at GRBs, leaving aside consideration of specific sources, was that of Cavallo & Rees (1978). The fireball would expand and adiabatically cool as it converts its internal into kinetic energy, and they suggested that this kinetic energy could be reconverted into radiation as it impacts the external medium, the highest efficiency (for non-relativistic expansion) being achieved when the fireball swept up an amount of external matter comparable to the fireball mass (the analog of the start of the Sedov-Taylor phase of SNe).

Paczynski (1986) proposed that a merging binary neutron star (BNS) would liberate enough energy in a short time to power a GRB at cosmological distances, and he and Goodman (1986) showed that in this case the expansion would be relativistic, the bulk Lorentz factor accelerating with radius as $\Gamma \propto r$. The initial blackbody plasma temperature would be of order a few MeV, which in the linearly expanding comoving frame would drop as $T_0 / r$, but in the observer frame this would be boosted by the bulk Lorentz factor back to its initial few MeV value, with an approximately blackbody spectrum, most of the photons escaping when the plasma became optically thin to Thompson scattering.

A different aspect of BNS mergers emphasized by Eichler et al. (1989) was their role as emitters of gravitational waves and their likely role as sources of r-process heavy elements, at the same time as being likely to appear as GRBs. A more detailed study of the properties of relativistically expanding fireballs (Paczynski, 1990; Shemi & Piran, 1990)}
showed that the bulk Lorentz factor growth $\Gamma(r) \propto r$ would saturate to a maximum value $\eta \sim E_f/M_f c^2 \gg 1$, where $E_f$ and $M_f$ are the initial energy of the explosion and the initial baryonic mass entrained in the outflow. After a saturation radius $r_{sat} \sim r_0 \eta$, where $r_0$ is the launching radius, the Lorentz factor remains constant, but since adiabatic cooling continues, the radiation energy that can escape after the photosphere becomes optically thin represents an increasingly smaller fraction of the final kinetic energy of expansion. The dynamics is similar also for neutron star-black hole (NS-BH) mergers, which would also be im-

**Fig. 1.** The standard GRB fireball shock model, e.g. from a collapsar (for compact mergers, the “collapse” region is replaced by the dynamical ejecta). Shown are the photosphere, internal shock and external shock resulting in the afterglow (Mészáros 2001).

**Fig. 2.** Left: Spectrum of a photosphere heated by \( pn \) decoupling (Beloborodov 2010). Right: Synchrotron spectra (e.g. internal or external shocks) accounting for time-dependence of transition between cooling regimes (Burgess et al. 2019).
important GRB candidates (Narayan et al. 1991, 1992), the latter mentioning briefly that reconnection, ejection of cosmic rays and their collisions might contribute non-thermal radiation in addition to the optical thick spectrum.

There were several problems with the above initial fireball models, namely, (1) for simplicity, a spherical geometry was usually tacitly assumed, and this combined with a low radiative efficiency would require excessively large explosion energies for the brighter bursts; (2) the main part of the gamma-ray spectrum predicted is approximately blackbody, whereas observed spectra are mainly non-thermal; and (3) for plausible baryon loads most of the explosion energy would be wasted on bulk kinetic energy, instead of radiation.

To address these issues the jet-like fireball shock model was developed, which in its main features is to this day the most widely used model. As a natural way to resolve the inefficiency of spherical models, Meszaros & Rees (1992b) pointed out that collimation of the fireball would be expected in the slower outflow (the dynamical ejecta) resulting from the tidal heating and the radiation from the merging BNS system. This could be powered by reconnection between their magnetospheres and collisions between their winds, as well as by neutrino-antineutrino interactions going into pairs, which would occur preferentially along the symmetry axis of the merger. This would create a hot radiation bubble, which would escape through the wind preferentially along the centrifugally rarefied axis of rotation, making a relativistic jet. For the case of NS-BH binary merger, Meszaros & Rees (1992a) discussed the increased radiative efficiency due to gravitational focusing by the BH of the neutrino-antineutrino interactions from the disrupted NS debris, giving a quantitative discussion of channeling into a jet along the axis.

To address the problem of the thermal spectrum and the radiative inefficiency at the photosphere due to most of the energy being converted into kinetic energy form, Rees & Meszaros (1992) showed that both of these issues are solved by considering the strong forward and reverse shocks produced in the deceleration of the relativistic ejecta by the external medium, which (unlike in the non-relativistic expansion) occurs when the ejecta has swept up an external mass which is \( \sim 1/\Gamma \) of its own mass, re-thermalizing about half of the bulk kinetic energy. The strong shock leads to a power-law relativistic electron spectrum via the Fermi mechanism, and via synchrotron radiation results in non-thermal power-law spectra. For a brief (impulsive) initial energy input, the effects of the deceleration are felt on a timescale \( t_{\text{dec}} \sim r_{\text{dec}}/2\Gamma c^2 \), when the initial forward shock Lorentz factor has dropped to \( \sim \Gamma_i/2 \) and the reverse shock, initially weak, has just become trans-relativistic. The results are the same whether the outflow is jet-like or spherical, for jet opening angles larger than 1/\( \Gamma \). At the photospheric radius a thermal spectrum is still emitted, but occurring above the saturation radius, adiabatic cooling makes its spectral contribution sub-dominant.

To address the problem of the thermal spectrum and the radiative inefficiency at the photosphere due to most of the energy being converted into kinetic energy form, Rees & Meszaros (1992) showed that both of these issues are solved by considering the strong forward and reverse shocks produced in the deceleration of the relativistic ejecta by the external medium, which (unlike in the non-relativistic expansion) occurs when the ejecta has swept up an external mass which is \( \sim 1/\Gamma \) of its own mass, re-thermalizing about half of the bulk kinetic energy. The strong shock leads to a power-law relativistic electron spectrum via the Fermi mechanism, and via synchrotron radiation results in non-thermal power-law spectra. For a brief (impulsive) initial energy input, the effects of the deceleration are felt on a timescale \( t_{\text{dec}} \sim r_{\text{dec}}/2\Gamma c^2 \), when the initial forward shock Lorentz factor has dropped to \( \sim \Gamma_i/2 \) and the reverse shock, initially weak, has just become trans-relativistic. The results are the same whether the outflow is jet-like or spherical, for jet opening angles larger than 1/\( \Gamma \). At the photospheric radius a thermal spectrum is still emitted, but occurring above the saturation radius, adiabatic cooling makes its spectral contribution sub-dominant.

A major motivation for introducing internal shocks arose after the launch of the Compton GRO (CGRO) spacecraft in late 1991, which found gamma-ray light curves which showed variabilities as short as \( 10^{-3} \) s. Such short variability can get smeared out in external shocks, which occur at relatively large radii. Rees & Meszaros (1994) showed that internal shocks at radii much smaller than those of the external shock can arise due to irregularly ejected gas shells of different bulk Lorentz factors. These can collide and shock at intermediate radii above the photosphere but below the external shock, leading to observable radiation whose variability is due the variability of the ejection from the central engine. Being above the photosphere, the shock radiation from synchrotron and inverse Compton is unsmeared and non-thermal.

A different power source for GRBs was proposed by Woosley (1993), in addition to BNS and NS-BH mergers. This is the collapsar model, resulting from the collapse of the core of massive stars leading to a central black hole
(or temporarily a magnetar). When the core is rotating fast enough, the mass fallback towards the BH would lead to an accretion disk powering a jet, which if fed long enough, can break out from the collapsing stellar envelope. The BH, accretion disk and jet resulting from this is similar to those expected in compact BNS or NS-BH mergers, and the shock radiation outside the envelope would have similar properties. However the accretion can last much longer, since fall-back times are long and the outer accretion radii would be larger, leading to longer total burst durations. This led to a natural explanation for the striking dichotomy between the two populations of short ($\Delta t \sim 2\ s$) and long ($2\ s \sim \Delta t \sim 10^3\ s$) GRBs identified by Kouveliotou et al. (1993).

Multi-wavelength, broadband spectra are expected in general from the external forward and reverse shocks, the reverse shock synchrotron predicting optical/UV radiation and the forward shock inverse Compton scattering of synchrotron photons reaching GeV energies Meszaros & Rees (1993). The latter provided a model (Meszaros & Rees 1994) for the long-lasting GeV emissions first seen in CGRO-EGRET data, while the former provided an explanation for the optical “prompt” optical emission first detected by Akerlof et al. (1999). Internal shocks also lead to broadband spectra Papathanassiou & Meszaros (1996), typically harder than in external shocks, due to the larger comoving magnetic fields at the smaller radii. The observation of some of the longer duration bursts, whose duration could significantly exceed the expected deceleration time $t_{\text{dec}} \sim \Delta t/2\Gamma^2c^2$, motivated a more detailed discussion of external shocks (Sari & Piran 1995, Sari 1997), distinguishing between thin shell cases (the limiting case from brief impulsive accretion) and thick shell cases (for longer accretion), where the reverse shock may be relativistic.

The long-term afterglow of a GRB, as opposed to the “prompt” emission discussed above, was first discussed quantitatively (Meszaros & Rees 1997) in a paper which appeared two weeks before the first announced detection of an X-ray afterglow from GRB 970228 with the Beppo-SAX satellite (Costa et al. 1997). Its optical afterglow was discovered by van Paradijs et al. (1997), and other afterglows soon followed, including GRB 970508 (Metzger et al. 1997) which yielded the first redshift ($z = 0.835$), proving that they were indeed cosmological. The observations confirmed in their main features the predictions of the afterglow model, including the power law time decay, spectra and timescale (Wijers et al. 1997). Synchrotron radiation, including the transition between slow and fast cooling regimes (Meszaros et al. 1998, Sari et al. 1998), provided a satisfactory fit for most of the observations in the subsequent period.

2. Standard GRB model and its evolution

The “standard” GRB fireball shock model outlined in the second half of the above section has proved extremely durable, despite a number of challenges and modifications of detail. For comprehensive reviews, see e.g., Piran (1999, 2004), Kumar & Zhang (2015), Zhang (2019). Its simplest form, most often used for interpreting observations, is in Fig. 3.
The afterglow radiation, from radio through optical, X-ray and more recently GeV is overall well fitted, with some modifications, by the external shock synchrotron emission, e.g. Zhang et al. (2006). For the "prompt" emission (broadly the typically MeV radiation within $\Delta t \approx T90$), however, an origin in terms of synchrotron has been criticized, e.g. [Preece et al. (1998)], since the low energy slope of some GRB prompt spectra is harder than the limiting synchrotron slope of -2/3 in $dN/dE$ (harder than +1/3 in $EdN/dE$).

One possible solution is that the prompt emission may be due to the optically thick photosphere, whose peak can be in the MeV range and the low energy slope is as hard as +2 (Eichler & Levinson 2000). This works but it requires an additional shock or other component to make a high energy power law (Mészáros & Rees 2000); also, if the photosphere is well above the saturation radius adiabatic cooling makes it radiatively inefficient. Radiatively efficient photospheres, however, may arise naturally if the photospheres are dissipative (Rees & Mészáros 2005), e.g. by magnetic reconnection, or subphotospheric shocks. A natural sub-photospheric dissipation mechanism is proton-neutron decoupling, which can produce efficient photospheric spectra from low energies all the way to multi-GeV (Beloborodov 2010, Fig. 2 left).

However, the earlier critiques of the synchrotron low energy slopes may be unjustified, having been obtained taking wide time bins or time-integrated spectra. If one considers the time evolution of the shock-accelerated electrons, from injection through cooling, and takes into account that electrons radiating at different frequencies have different energies and may be in different cooling regimes (fast, intermediate, slow), the spectra convolved with the detector energy resolution and response function can give various slopes, and the great majority of the observed GRB slopes can be fitted with synchrotron, e.g. Burgess et al. (2018), Ravasio et al. (2019a), Fig. 2 right.

A critique of the simple internal shocks in which only the electrons radiate (leptonic models) has been that they are generally inefficient, not dissipating enough of the mechanical energy in the relative motion between successively ejected shells, and Fermi acceleration putting much of this dissipated energy in non-radiating protons. In more realistic internal shocks, however, this radiative inefficiency can be larger, e.g. when the dissipation is largely by magnetic instabilities and reconnection, or if hadronic collisions and reacceleration of secondaries are taken into account. Thus, the early GRB paradigm based on internal plus external shocks and an inefficient photosphere (Fig. 3 top) has, since about 2005, evolved into one of an efficient photosphere and/or an efficient internal shock plus external shock (Fig. 3 bottom). An example of efficient magnetic dissipation internal shock models is the ICMART model of Zhang & Yan (2011), while an example of an efficient hadronic internal shock with secondary reacceleration is that of Murase et al. (2012).

After the launch of the Fermi Gamma-ray Space Telescope 2008, its LAT detector started to observe in a significant fraction of GRBs that the so-called Band broken power law spectrum above the MeV peak extended into the GeV range, as already found in some previous EGRET spectra. Such "extended" Band spectra can be modeled, e.g. with photospheric models, as seen in Fig. 3 (left). However, in many Fermi-LAT GRBs the GeV appeared as a second power-law component, harder than the Band $\beta$ upper branch.

The question is whether this second, harder GeV component is due to inverse Compton (IC) upscattering of the Band component, or is it due to protons being accelerated and leading to cascades with radiation from secondary leptons. Both types of models can give reasonable fits. Leptonic models where the Band spectrum arising in the photosphere is up-scattered by shocked electrons in internal shocks give reasonable results, e.g. [Toma et al. (2011)]. An alternative leptonic model considers a baryonic or magnetic photosphere producing a Band spectrum which is up-scattered in an external shock, also giving good fits (Veres & Mészáros 2012, Fig. 4 (left)).

Hadronic models, on the other hand, could in principle have substantial advantages. E.g. Murase et al. (2012) calculated an inter-
3. Some recent developments

Recently the MAGIC imaging air fluorescence telescope (IACT) announced the detection of photons in excess of 300 GeV, and perhaps up to a TeV, in the bright GRB190114C [Mirzoyan 2019], also detected at other en-
Fig. 6. GRB/GW170817A, two opposed views on the SGRB radiation from a BNS. Left: observed radiation dominated by the cocoon, for a choked jet. Right: observed radiation dominated by an emergent top-hat jet (Kasliwal et al. 2017).

Fig. 7. Left: IceCube and Antares upper limits for GRB170817 (Albert et al. 2017), compared to BNS jet EE extended emission model (Kimura et al. 2017) for various jet offset angles. Right: Internal and collimation shocks in trans-ejecta jet propagating through a BNS dynamical ejecta (Kimura et al. 2018).

Energies by Fermi, Swift, INTEGRAL and numerous other facilities. This was the first high confidence (~ 20σ) detection of a GRB with an IACT at such energies, a long awaited feat which should be easier to accomplish with the future CTA. Preliminary analyses show that the long lasting (~ 10^3 s) sub-TeV component is mostly associated with the afterglow seen at other energies, e.g. Fig. 5. The spectral slope of the sub-TeV component appears harder than the usual Band component, pending further MAGIC analysis.

Detection of ~ TeV emission from a GRB has two requirements, one being that the redshift be smallish, so that γγ absorption in the IGM external background light is not too severe, and the other being that the same absorption is absent or at least mitigated in the GRB radiation zone and its immediate neighborhood. Fermi-LAT detections of GRBs have shown source-frame emission up to several tens of GeV and in one case even ~ 100GeV, but the present ~ 300GeV can put significant
constraints on models. Much theoretical work remains to be done on this event.

The other major recent development was the short GRB 170817 detection both electromagnetically (EM) through multi-wavelength photons, and through gravitational waves (GWs). This was very exciting, being the first high significance multi-messenger detection of a transient using GWs. This was a short GRB (SGRB) with $\Delta t < 2\text{~s}$, detected by Fermi, INTEGRAL, Swift and other EM instruments. These objects were long expected to arise from BNS mergers, an interpretation for which accumulating evidence, e.g. Gehrels et al. (2009), had almost but not quite reached the 100% confidence level. In this case, slightly preceding the EM flash, the associated detection of GWs (Abbott et al. 2017), which were also expected from BNS mergers, conclusively confirmed that SGRBs where indeed BNS mergers. In addition, it also confirmed that BNSs can also produce a type of optical/IR flash known as a kilonova, surmised to be responsible also for the elements heavier than the Fe-group via the r-process, e.g. Hotokezaka et al. (2018), Kasliwal et al. (2019).

The SGRB radiation of GRB/GW170817 looked typical, except for being fainter and somewhat softer than expected for its low distance of 40 Mpc. The role played in GRBs by cocoons (Meszaros & Rees 2001) Ramirez-Ruiz et al. (2002) and choked jets (Meszaros & Waxman 2001) had been considered early on, and in the case of GRB/GW170817A a natural possibility was that its weaker $\gamma$-rays might be attributed either to a choked jet with a cocoon breakout, or to an off-axis top-hat emergent jet, e.g. Kasliwal et al. (2017), Ioka & Nakamura (2017) and others. While a cocoon interpretation may be favored over a simple top-hat jet, the observation of a superluminal jet signature (Mooley et al. 2018) and other features of the afterglow (Troja et al. 2018) indicate that either a Gaussian structured jet or a cocoon could fit the data.

The next burning question, as far as GRB multi-messenger studies, is whether GRBs can also be detectable via neutrinos. Of course both LGRBs (as core-collapse objects) and SGRBs (compact mergers involving at least one neutron star which is heated to virial temperatures) will emit a large fraction of their core binding energy in thermal (5-30 MeV) neutrinos. At these energies the neutrino-nucleon detection cross section is of order $10^{-44}\text{~cm}^2$, and at cosmological distances the flux is undetectable with current detectors. High energy neutrinos however have much higher cross sections ($\sim 10^{-35}\text{~cm}^2$ around 10 TeV), and IceCube is detecting a diffuse astrophysical flux in the 10 TeV-10 PeV range (Aartsen et al. 2013, IceCube Collaboration 2013). The total number of neutrinos so far is of order 50, distributed isotropically in the sky, with localization error circles ranging from $\sim 1''$ (for muon neutrino tracks) to $15-30''$ (for electron neutrino cascades), hence difficult to associate with individual sources. Recently, however, a high energy (multi-TeV) muon neutrino was detected, with the blazar TXS 0506+56 within its error circle, which was undergoing a $\gamma$-ray flaring episode in near time coincidence with the neutrino arrival. The region also showed other previous neutrinos in the past years, but without coincident $\gamma$-ray flares, so the total coincidence significance is $\sim 3.5\sigma$, which is interesting but not yet considered conclusive evidence (IceCube, and other Collaborations 2018, IceCube Collaboration et al. 2018).

The possibility of GRBs being high energy neutrino sources has been investigated by IceCube using classical GRBs, i.e. bright, EM-detected, mainly LGRBs. These have been disfavored by IceCube analyses, e.g. Aartsen et al. (2013), using particular models of the neutrino emission expected. The same conclusion is reached by IceCube for classical GRBs in a model-independent way using constraints based on neutrino multiplet observations (Aartsen et al. 2018), but the same study leaves unconstrained a (theoretically plausible) origin in low-luminosity or choked GRBs, e.g. Senno et al. (2016). Low luminosity and/or choked GRBs could be more numerous than classical GRBs, and at the typically high

---

1 The other previous high significance multi-messenger transient was SN 1987a, where besides photons also thermal (MeV) neutrinos were detected.
redshifts they would be electromagnetically missed or hard to detect, while their cumulative neutrino flux could add up to what IceCube sees.

SGRBs would in principle appear to be ideal objects for constraining physical models if in addition to GWs they also produced observable neutrinos. At first sight it would seem that the expected neutrino fluxes would be much lower than in LGRBs, since the SGRB prompt MeV emission is shorter and underluminous compared to LGRBs. However a large fraction of SGRBs also exhibit a longer tail (~ 100s) of softer radiation in the 50 keV range. This softer extended emission (EE) can be modeled as a late jet emission with a bulk Lorentz factor lower than the prompt, providing a higher comoving density of target photons for $p\gamma$ photo-hadronic interactions leading to neutrinos. The neutrino flux is still low at typical redshifts, but at the redshift $z \sim 0.01$ (about 40 Mpc) of GBB 170817A, it could have been detectable by IceCube, if the jet been head-on (Kimura et al. 2017); however, for a higher inclination angle $\theta_{\text{LOS}} \sim 20 - 30^\circ$ of the line of sight relative to the jet axis, as inferred from multi-wavelength observations, the lower Doppler boost in that direction implies a much lower observable flux, which falls below the IceCube sensitivity, Fig. 8 (left).

The SGRB jet and shock structure is likely to be more complicated as it is making its way through the dynamical ejecta, Fig. 8 (right). Both collimation shocks and internal shocks are expected in choked jets or before the jet emerges from the ejecta, and the internal shocks occurring in the pre-collimation jet satisfy the conditions for Fermi acceleration of charged particles, leading to neutrinos via photo-hadronic interactions (Kimura et al. 2018). One can expect from such events a few up-going neutrinos in IceCube from a merger at 40 Mpc occurring in the Northern sky, if the jet is directed at Earth. For optimistic jet parameters, a joint GW-IceCube detection might be achievable in a few years of operation, or for Ice-Cube Gen 2 this would be probable even for moderate jet parameters.

Acknowledgements. I am grateful to K. Murase, S.S. Kimura and D.B. Fox for discussions, and the Eberly Foundation for support.

References

IceCube Collaboration 2013, Science, 342, 1
IceCube, and other Collaborations 2018, Science, 361, eaat1378