A&A 636, A82 (2020)
https://doi.org/10.1051/0004-6361/201937244
© ESO 2020

Astronomy & Astrophysics

Proton–synchrotron as the radiation mechanism of the prompt emission of gamma-ray bursts?

G. Ghisellini[1], G. Ghirlanda[1], G. Oganesyan[2,3,4], S. Ascenzi[1], L. Nava[1,5,6,7], A. Celotti[8,1,6,7], O. S. Salafia[1], M. E. Ravasio[1,9], and M. Ronchi[1]

[1] INAF – Osservatorio Astronomico di Brera, Via Bianchi 46, 23807 Merate, Italy
e-mail: gabriele.ghisellini@inaf.it
[2] Gran Sasso Science Institute, Viale F. Crispi 7, 67100 L'Aquila, AQ, Italy
[3] INFN – Laboratori Nazionali del Gran Sasso, 67100 L'Aquila, AQ, Italy
[4] INAF – Osservatorio Astronomico d'Abruzzo, Via M. Maggini snc, 64100 Teramo, Italy
[5] INAF – Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, 34143 Trieste, Italy
[6] INFN – Sezione di Trieste, Via Valerio 2, 34127 Trieste, Italy
[7] IFPU – Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy
[8] SISSA, Via Bonomea 265, 34136 Trieste, Italy
[9] Université degli Studi di Milano-Bicocca, Dip. di Fisica "G. Occhialini", Piazza della Scienza 3, 20126 Milano, Italy

Received 3 December 2019 / Accepted 21 January 2020

ABSTRACT

We discuss the new surprising observational results that indicate quite convincingly that the prompt emission of gamma-ray bursts (GRBs) is due to synchrotron radiation produced by a particle distribution that has a low-energy cut-off. The evidence of this is provided by the low-energy part of the spectrum of the prompt emission, which shows the characteristic $F_\nu \propto \nu^{1/3}$ shape followed by $F_\nu \propto \nu^{-1/2}$ up to the peak frequency. This implies that although the emitting particles are in fast cooling, they do not cool completely. This poses a severe challenge to the basic ideas about how and where the emission is produced, because the incomplete cooling requires a small value of the magnetic field to limit synchrotron cooling, and a large emitting region to limit the self-Compton cooling, even considering Klein–Nishina scattering effects. Some new and fundamental ingredient is required for understanding the GRBs prompt emission. We propose proton–synchrotron as a promising mechanism to solve the incomplete cooling puzzle.

Key words. radiation mechanisms: non-thermal – gamma-ray burst: general – gamma-rays: general

1. Introduction

The radiation mechanism of the prompt emission of gamma-ray bursts (GRBs) has been debated since the very first observations. Its non-thermal appearance and the idea that shocks are responsible for accelerating particles and enhancing the magnetic field soon led to the proposal that the synchrotron process should be the dominant radiative mechanism (Katz 1994; Rees & Meszaros 1994; Tavani 1996).

The observed fast variability (down to millisecond timescales, e.g. Walker et al. 2000) requires the source to be compact, which would suggest it has large magnetic and radiation energy densities. In these conditions radiative cooling is very efficient, and the corresponding spectrum is expected to be $F_\nu \propto \nu^{-0.5}$ or softer (e.g. Ghisellini & Celotti 1999). The observed spectrum is instead much harder (see e.g. Preece et al. 1998a). When fitted with the Band function (Band et al. 1993), that is a phenomenological model composed by two smoothly connected broken power laws, the average spectrum shows a peak in the νF_ν representation, with photon spectral slopes $\alpha \sim 1$ below and $\beta \sim 2.3$ above the peak frequency $\nu_{\rm peak}$ ($\dot{N}_\nu \propto \nu^{-\alpha}, \nu^{-\beta}$; Kaneko et al. 2006; Nava et al. 2011; Goldstein et al. 2012; Gruber et al. 2014; Lien et al. 2016). This remains true when considering time-resolved spectra (for the brightest bursts, e.g. Preece et al. 1998b; Ghirlanda et al. 2002; Burgess et al. 2014; Yu et al. 2016). On a small number of occasions, the very hard low-energy spectra have been reproduced with a thermal component: in a few cases with a pure black body spectrum (Ghirlanda et al. 2004, 2013); but more often with a power law or a Band model with the addition of a black body contribution (Ryde & Pe'er 2009; Ryde et al. 2010; Guiriec et al. 2011; Burgess et al. 2014; Pe'er & Ryde 2017, but see Ghirlanda et al. 2007).

Recently, it was found that the overall spectral energy distribution (SED) could be fitted by three power laws, smoothly joining at two energies: one at the break frequency $\nu_{\rm b}$ and the other at the peak frequency $\nu_{\rm peak}$ (Oganesyan et al. 2017, 2018, 2019; Ravasio et al. 2018, 2019). Below $\nu_{\rm b}$ the photon spectral index is close to $\alpha_1 = 2/3$; between $\nu_{\rm b}$ and $\nu_{\rm peak}$ the index is approximately $\alpha_2 = 1.5$, and above $\nu_{\rm peak}$ the index β becomes (as before – Nava et al. 2011) close to 2.3 or slightly steeper ($\beta = 2.8$) when allowing for the presence of another break at low energies, possibly with an exponential cut off at high energies. This resulting typical spectrum is sketched in the two bottom panels of Fig. 1.

More physically, Oganesyan et al. (2019) also successfully reproduced GRB spectra with the synchrotron spectrum produced by a non-thermal electron energy distribution (see also Chand et al. 2019; Burgess et al. 2020; Ronchi et al. 2020). The top panel of Fig. 1 shows the particle distribution corresponding to the assumption that it emits such synchrotron radiation. It must have a low-energy cut-off at some energy $\gamma_{\rm b} = \gamma_{\rm cool}$ and particles close to $\gamma_{\rm cool}$ are responsible for the emission with the hard index α_1. The value of the index α_2 strongly suggests that the corresponding emitting particles are radiatively cooling and distributed as $N(\gamma) \propto \gamma^{-2}$. Above $\gamma_{\rm peak} = \gamma_{\rm inj}$, $N(\gamma)$ must be a relatively steep power law, $N(\gamma) \propto \gamma^{-3.6}$, to account for the observed $\beta = 2.3$.

The particle distribution $N(\gamma)$ can be obtained considering particle injection and radiative cooling. First we consider the injection of relativistic particles at a rate of $Q(\gamma) \propto \gamma^{-p}$ between $\gamma_{\rm inj}$ and $\gamma_{\rm max}$ throughout an emitting source of size R, as shown by the dashed line in the top panel of Fig. 1. If the radiative cooling rate is $\propto\gamma^2$, the emitting particle distribution $N(\gamma, t)$, after one light crossing time R/c $[N(\gamma, R/c)]$ is schematically characterised by the red line in the top panel of Fig. 1 in the case of fast cooling (i.e. when $\gamma_{\rm cool} < \gamma_{\rm inj}$). Here, we obtain the following:

1. there are no particles below $\gamma_{\rm cool}$ and above $\gamma_{\rm max}$;
2. $N(\gamma) \propto \gamma^{-2}$ between $\gamma_{\rm cool}$ and $\gamma_{\rm inj}$;
3. $N(\gamma) \propto \gamma^{-(p+1)}$ between $\gamma_{\rm inj}$ and $\gamma_{\rm max}$.

Such particle distribution emits a synchrotron spectrum:

1. $F_\nu \propto \nu^{1/3}$ for $\nu < \nu_{\rm cool}$. This low-energy tail is mainly produced by particles with random Lorentz factor $\gamma_{\rm cool}$;
2. $F_\nu \propto \nu^{-1/2}$ between $\nu_{\rm cool}$ and $\nu_{\rm peak}$, radiated by particles with random Lorentz factors $\gamma_{\rm cool} < \gamma < \gamma_{\rm inj}$;
3. $F_\nu \propto \nu^{-p/2}$ in the range from $\nu_{\rm peak}$ to $\nu_{\rm max}$, owed to particles with $\gamma_{\rm inj} < \gamma < \gamma_{\rm max}$.
4. Above $\nu_{\rm max}$, the spectrum ends with an exponential cut. The emission is basically emitted by particles with $\gamma_{\rm max}$.

For $p > 2$, the νF_ν spectrum peaks at the frequency mainly produced by electrons with random Lorentz factors $\gamma_{\rm inj}$ (the example shown in Fig. 1), while for $p < 2$, the spectral peak corresponds to the frequency chiefly emitted by electrons with $\gamma_{\rm max}$.

The fact that the majority of the spectra of bright long GRBs can be fitted with the above-mentioned three-power-law model (Oganesyan et al. 2017; Ravasio et al. 2018, 2019) indeed suggests that the synchrotron process is the radiative mechanism giving rise to the prompt emission. This implies that the emitting particles do not cool completely (Daigne et al. 2011), but remain at the energy $\gamma_{\rm cool}$ for a timescale comparable to the typical time bin of the time resolved spectral analysis (∼1 s). This poses a challenge, since the prompt emission is believed to be produced in compact regions, as demonstrated by the very rapid variability of the flux that can reach values as short as one millisecond (i.e. Bhat et al. 1992). Even accounting for the relativistic Doppler time contraction, the emitting region must be small and located at a distance $R \leq ct_{\rm var}\Gamma^2/(1+z)$ from the central engine. This in turn must correspond to large energy densities, both magnetic and radiative, leading to very efficient radiative cooling due to the synchrotron and self-Compton processes. However, in the above scenario, cooling should stop when particles reach values of $\gamma_{\rm b} = \gamma_{\rm cool}$ significantly larger than unity.

As specified below, in this framework incomplete cooling of the electrons would demand low magnetic field (to avoid fast synchrotron cooling) and large radii (to avoid fast inverse Compton cooling), but the observed short variability timescales require small radii. This is the key issue we face in this work.

In Sect. 2 we reassess the synchrotron and self-Compton cooling and their relative relevance. Estimates on the expected magnetic field are revised in Sect. 3. We examine ways out within the standard scenario in Sect. 4. A proposed alternative, namely proto-synchrotron radiation, is presented in Sect. 5, and we present our conclusions in the Sect. 6.

Hereafter we adopt the notation $Q = 10^x Q_x$ and cgs units, unless otherwise noted, and a flat cosmology with $\Omega_\Lambda = h = 70$.

2. Radiative cooling

2.1. Synchrotron and self Compton cooling timescale

In this section we estimate the cooling timescale of leptons emitting synchrotron and self-Compton radiation. In general, the self-Compton process will occur partly in the Thomson and partly in the Klein–Nishina regime. As detailed below, since the latter process is less efficient, it will be approximated.

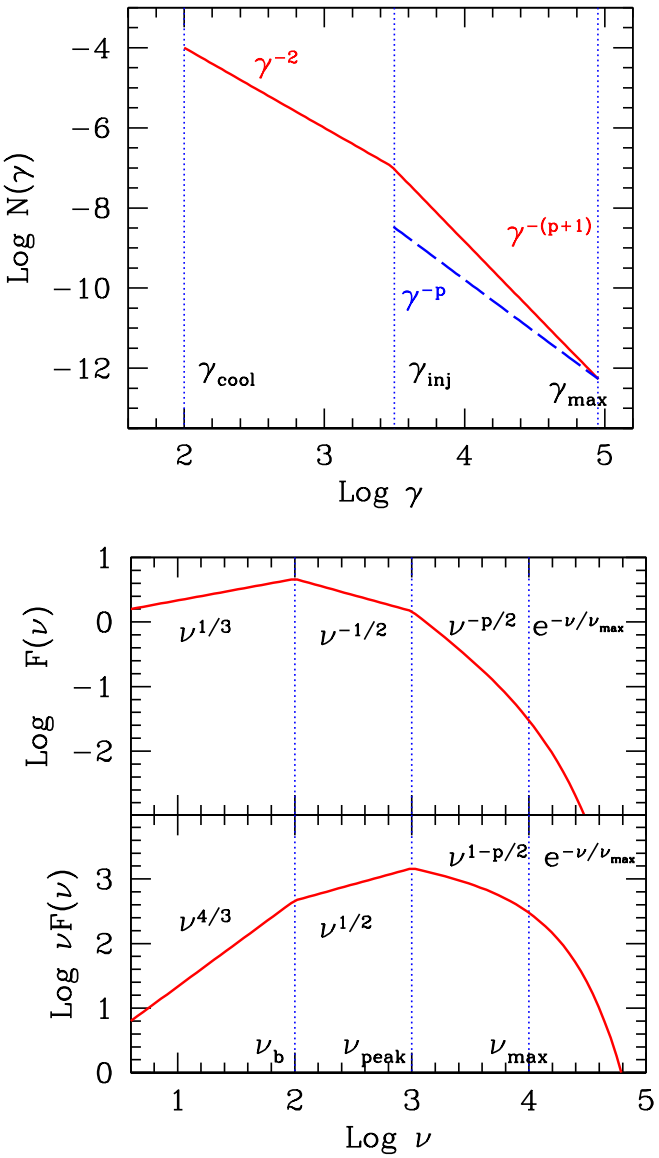

Fig. 1. *Top panel*: schematics of the particle distribution responsible for the spectra of the bottom two panels. The dashed blue line corresponds to the injected $[Q(\gamma)]$ distribution. The characteristic Lorentz factors and frequencies are labelled as described in the text. The solid and dashed lines are re–scaled by an arbitrary amount. *Bottom two panels*: sketch of synchrotron spectra as reproduced by the spectral analysis discussed in the text. The spectra show a high-energy exponential cut-off which is not always present or detectable in real data.

When treating the inverse Compton (IC) scattering it is convenient to adopt dimensionless photon energies $x \equiv h\nu/(m_{\rm e}c^2)$. In the comoving (hereafter primed) frame the scattering is described by the Klein–Nishina cross section $\sigma_{\rm KN}$ which is equal to the Thomson one ($\sigma_{\rm T}$) for $x' \ll 1/\gamma$. For simplicity, we then assume that

$$\begin{aligned} \sigma_{\rm KN} &= \sigma_{\rm T}, \quad x' \leq 1/\gamma \\ \sigma_{\rm KN} &= 0, \quad x' > 1/\gamma. \end{aligned} \tag{1}$$

This overestimates somewhat the cross section when $x' \sim 1/\gamma$ (in this case $\sigma_{\rm KN} = 0.43\sigma_{\rm T}$) and of course underestimates it at high energies. However, this approximation is reasonable when

considering scatterings between rather wide distributions of photon and electron energies, becoming more inaccurate when these are narrow.

According to such approximation, an electron of random Lorentz factor γ loses energy by scattering a fraction of the total radiation energy density $U'_{\rm r}$, given by

$$U'_{\rm r} = \frac{L'_{\rm iso}}{4\pi R^2 \Delta R'} \frac{\Delta R'}{c} = \frac{L_{\rm iso}}{4\pi R^2 c \Gamma^2}. \quad (2)$$

Not all this radiation energy density is available for scattering in the Thomson regime. The larger γ the smaller the fraction $f(\gamma)$ of scattered photons:

$$f(\gamma) = \frac{\int_0^{1/\gamma} U'_{\rm r}(x')\,{\rm d}x'}{U'_{\rm r}}. \quad (3)$$

The corresponding lepton cooling rate can be expressed as:

$$P_{\rm IC}(\gamma) = \dot\gamma_{\rm IC} m_{\rm e} c^2 = \frac{4}{3}\sigma_{\rm T} c \gamma^2 U'_{\rm r} f(\gamma), \quad (4)$$

and it is accurate enough to estimate the cooling time of electrons emitting by the synchrotron and IC process, namely:

$$t'_{\rm cool}(\gamma) = \frac{\gamma}{\dot\gamma} = \frac{3 m_{\rm e} c^2}{4\sigma_{\rm T} c \gamma \left[U'_{\rm B} + U'_{\rm r} f(\gamma)\right]}, \quad (5)$$

where $U'_{\rm B}$ is the magnetic field energy density. For a source at a redshift z whose flow is moving relativistically, the observed cooling timescale appears $t^{\rm obs}_{\rm cool} = t'_{\rm cool}(1+z)/\delta$, where $\delta = [\Gamma(1-\beta\cos\theta)]^{-1}$ is the relativistic Doppler factor and θ is the viewing angle of the flow with respect to the line of sight. Approximating $\delta \sim \Gamma$, and using

$$\nu^{\rm obs} = \frac{4}{3}\frac{eB'}{2\pi m_{\rm e} c}\gamma^2 \frac{\Gamma}{1+z} \ \text{i.e.,}\ \gamma = \left[\frac{3\pi m_{\rm e} c \nu^{\rm obs}}{2eB'}\frac{(1+z)}{\Gamma}\right]^{1/2} \quad (6)$$

we obtain

$$t^{\rm obs}_{\rm cool}(\gamma) = \frac{6\pi m_{\rm e} c^2}{\sigma_{\rm T} c B'^{3/2}}\left[\frac{2e}{3\pi m_{\rm e} c \nu^{\rm obs}}\frac{1+z}{\Gamma}\right]^{1/2}\frac{1}{\left[1+f(\gamma)U'_{\rm r}/U'_{\rm B}\right]}$$
$$= \frac{4.7\times10^{-8}\,(1+z)^{1/2}}{B'^{3/2}_6\left[\Gamma_2 \nu^{\rm obs}_{19}\right]^{1/2}} \times \frac{1}{\left[1+f(\gamma)U'_{\rm r}/U'_{\rm B}\right]}\ {\rm s}, \quad (7)$$

where the first part of Eq. (7) is the synchrotron cooling time. As reference value the random Lorentz factor of electrons emitting photons at frequency 10^{19} Hz is $\gamma = 163\,[(1+z)\nu_{19}/B'_6\Gamma_2]^{1/2}$.

Figure 2 shows the radiative cooling time for an electron of energy $\gamma_{\rm b}$, integrating the (comoving) energy density up to the Klein Nishina threshold $x'_{\rm KN} = 1/\gamma_{\rm b}$. The radiative cooling time is shown as a function of the magnetic field and for different distances R from the central engine. The black dashed line corresponds to the synchrotron cooling time only. The figure shows that for a given size a decrease in the magnetic field increases the total cooling time only slightly because the inverse Compton cooling becomes more severe. We interpret $\nu_{\rm b}$ as the cooling frequency $\nu_{\rm cool}$. In Fig. 2 the yellow horizontal line corresponds to one second, the typical integration time needed to collect enough photons for the spectral analysis. In the case of $t_{\rm cool} \sim 1$ s, the distance $R \gtrsim 10^{16}$–10^{17} cm and magnetic fields $B' \lesssim 10$ G are required.

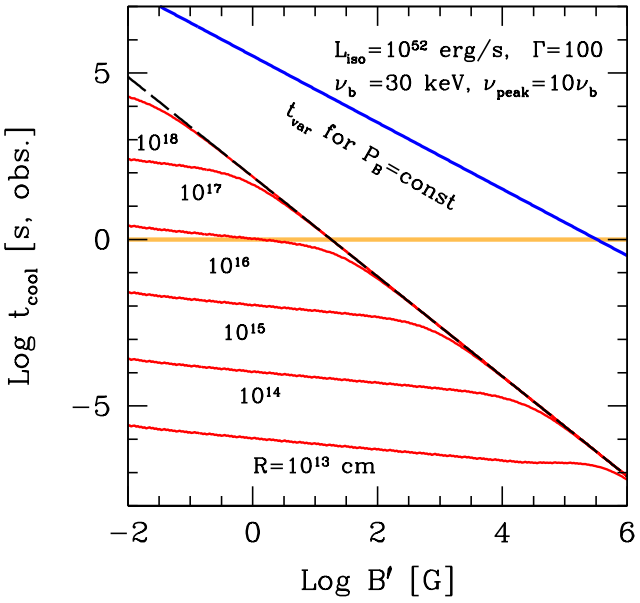

Fig. 2. Observed cooling timescale for the electrons emitting at the break frequency $\nu_{\rm b}$ as a function of the magnetic field and for different distances from the central engine. It is assumed that particles cool via synchrotron and self-Compton processes (considering, for the latter, only the fraction of the synchrotron spectrum below $h\nu/(m_{\rm e}c^2) = 1/\gamma_{\rm b}$). The black dashed line indicates the synchrotron cooling timescale only. The blue line is the minimum variability timescale found by assuming that the Poynting flux remains constant beyond the acceleration phase, leading to $B' \propto R^{-1}$ (see Sect. 3). As reference, a redshift $z = 1$ has been considered. The orange horizontal line corresponds to a typical exposure time of 1 s.

2.2. Self-Compton-to-synchrotron ratio

A generic electron of random Lorentz factor γ will cool by synchrotron and inverse Compton. The ratio of the two loss rates is

$$\frac{\dot\gamma_{\rm IC}}{\dot\gamma_{\rm Syn}} = \frac{U'_{\rm r}}{U'_{\rm B}} f(\gamma). \quad (8)$$

To find the ratio $L_{\rm C}/L_{\rm Syn}$ we must integrate over the particle energy distribution:

$$\frac{L_{\rm C}}{L_{\rm Syn}} = \frac{U'_{\rm r}}{U'_{\rm B}}\frac{\int_1^{\gamma_{\rm max}} N(\gamma)\gamma^2 f(\gamma){\rm d}\gamma}{\int_1^{\gamma_{\rm max}} N(\gamma)\gamma^2 {\rm d}\gamma}. \quad (9)$$

Therefore we must specify the shape of the particle distribution.

We assume that the typical spectrum observed in the X and γ–ray energy range during the prompt emission has the form:

$$\begin{aligned} F(\nu') &= A\nu'^{1/3}, && \nu' < \nu'_{\rm b} \\ F(\nu') &= A\nu'^{5/6}_{\rm b}\nu'^{-1/2}, && \nu'_{\rm b} < \nu' < \nu'_{\rm peak} \\ F(\nu') &= A\nu'^{5/6}_{\rm b}\nu'^{\beta-1/2}_{\rm peak}\nu'^{-\beta}, && \nu' > \nu'_{\rm peak}. \end{aligned} \quad (10)$$

The normalization constant A can be found by the observed total synchrotron flux. This spectrum is emitted by electrons distributed in energy as a broken power law:

$$\begin{aligned} N(\gamma) &= K\gamma^{-2}, && \gamma_{\rm b} \le \gamma \le \gamma_{\rm peak} \\ N(\gamma) &= K\gamma^{p-2}_{\rm peak}\gamma^{-(p+1)}, && \gamma > \gamma_{\rm peak}, \end{aligned} \quad (11)$$

where $\gamma_{\rm b}$ and $\gamma_{\rm peak}$ are the energies of the electrons emitting mainly at $\nu_{\rm b}$ and $\nu_{\rm peak}$. The slope $p = 2\beta$.

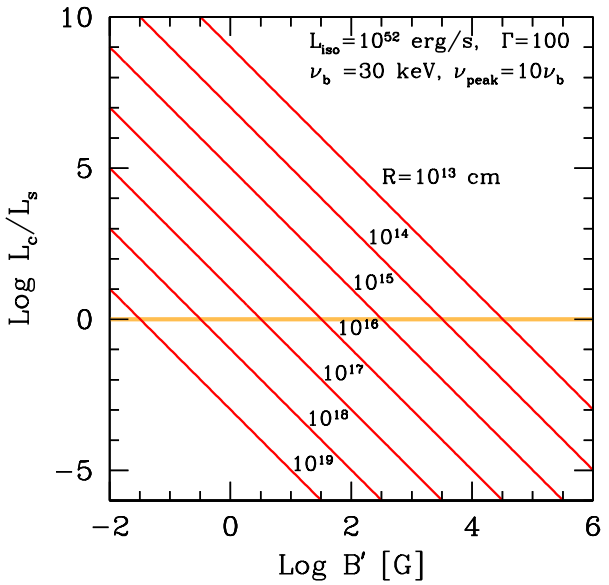

Fig. 3. Ratio of the Compton to synchrotron luminosity predicted for the same parameters as in Fig. 2. In this case we have considered the entire distribution of electrons: for each electron energy, we considered the corresponding synchrotron radiation energy that is scattered in the Thomson regime. The horizontal orange line corresponds to equal synchrotron and self-Compton luminosities.

Calculating Eq. (9) assuming the spectrum of Eq. (10), we constructed Fig. 3 showing the Compton-to-synchrotron luminosity ratio as a function of the magnetic field and for different distances from the central engine. It can be seen that to have unimportant Compton emission (i.e. a ratio smaller than unity, represented by the orange line) for magnetic fields $B' < 100$ G, the distance R must be larger than $\sim 10^{16}$ cm. This contrasts with the short (sub-second) variability timescales often seen in the prompt emission of GRBs (see e.g. MacLachlan et al. 2013; McBreen et al. 2001). In the standard scenario of shells with a spherical curvature, the minimum variability timescale is, for on-axis observers:

$$t_{\rm var} = \frac{R(1+z)}{2c\Gamma^2} \sim 17\,\frac{R_{16}(1+z)}{\Gamma_2^2}\ {\rm s}. \tag{12}$$

This should be compared with the observed variability timescales, which are much shorter. To detect short timescales we need a large effective area, and indeed the fastest (millisecond) variability was detected by BATSE onboard the Compton Gamma-Ray Observatory satellite. Golkhou & Butler (2014) reported typical variability timescales of 0.01−1 s with BAT (Burst Alert Telescope) onboard *Swift* and similar variability timescales are observed Golkhou et al. (2015) in the GRBs detected by the Gamma-ray Burst Monitor onboard *Fermi*.

3. Expected magnetic field

Most models of GRBs require a very large magnetic field at the base of the jet in order to extract the rotational energy of the black hole through the Blandford & Znajek (1977) process. Beyond the acceleration zone of the jet, the Poynting flux $P_{\rm B}$ is assumed to be constant, consistent with the adiabatic assumption. The assumption of an initially magnetically dominated fireball is not crucial for our arguments, and there could be other mechanisms capable of providing the required energetics (i.e. neutrino–antineutrino annihilation – Eichler et al. 1989; Zalamea & Beloborodov 2011). However, it is instructive to derive the value of the magnetic field in the emitting region at a distance where the fireball becomes transparent in the case of magnetic fields dominating the energetics at the start of the jet. If all the energy initially carried by the jet (i.e. close to the initial radius R_0) is magnetic, then the initial $P_{\rm B}$ should be of the same order of the total energy $P_{\rm j}$ of the jet after its acceleration. The kinetic power is increasing in the acceleration phase at the expense of $P_{\rm B}$. According to this prescription, the radial profile of $P_{\rm B}$ can be written as:

$$P_{\rm B} = \pi\psi^2 R^2 c\Gamma^2 U'_{\rm B} = P_{\rm j}\left[1 - \frac{\Gamma\beta}{\Gamma_{\rm max}\beta_{\rm max}}(1-\epsilon_{\rm B})\right], \tag{13}$$

where $\epsilon_{\rm B}$ is the fraction of the total power remaining in Pointing flux after the acceleration phase, and ψ is the semi-aperture angle of the jet. This leads to a value of the magnetic field beyond the acceleration zone of

$$B' = \left[\frac{8\epsilon_{\rm B}P_{\rm j}}{c}\right]^{1/2}\frac{1}{\psi R\Gamma_{\rm max}}. \tag{14}$$

As an example, for $\epsilon_{\rm B} = 0.1$, and $\Gamma_{\rm max} = 100$, $P_{\rm j} = 10^{52}\,{\rm erg\,s^{-1}}$, we have $B' = 10^7\,\psi_{-1}^{-1}R_{13}^{-1}$ G. We can relate the value of the magnetic field with the minimum variability timescale:

$$t_{\rm var} \sim \frac{R}{2\Gamma^2 c} = \left[\frac{8\epsilon_{\rm B}P_{\rm j}}{c}\right]^{1/2}\frac{1}{2\psi B'\Gamma^3 c}. \tag{15}$$

This is the blue line in Fig. 2: to have short variability timescales, the region must be small, and within our approximations (conical jet) this requires short distances from the central engines, and therefore large magnetic fields which are incompatible with relatively long cooling timescales.

4. Ways out

4.1. Continuous re-acceleration

The fast cooling rate could be halted by an acceleration mechanism that is dominant at low energies. If this is constant in time, it means that particles at low energies are heated, while particles at high energies cool. Therefore, particles accumulate at the energy for which heating and cooling balance (see e.g. Asano & Terasawa 2009, see Katz 1994 for the afterglow emission). There will be a pile up, and the emitted spectrum will disagree with the observed one. Furthermore, in the internal shock scenario of GRBs, the injected electrons are always new ones, and they are never reaccelerated. One could consider a variation of this scheme (Ghisellini & Celotti 1999) entailing a steady state between heating and cooling leading to a thermal (e.g. Maxwellian) particle distribution with a sub-relativistic temperature. In this case the main radiation mechanism is thermal Comptonization, and the observed spectral indices are unlikely to be obtained (and to be the same in different sources), because they require an ad hoc geometrical and physical setup[1].

4.2. Impulsive re-acceleration

There could be a specific acceleration mechanism that avoids the pile up of particles. Here, we assume that the particles are accelerated in a timescale shorter than their cooling time, and then radiate and cool down to the required $\gamma_{\rm cool}$. Upon reaching $\gamma_{\rm cool}$,

[1] Thermal Comptonization spectra are usually characterised by a single power law ending with an exponential cut, or by a power law, a hump (the "Wien hump"), and an exponential tail. It would be very difficult to obtain the observed three power-law segments.

they are reaccelerated back to high energies. This process avoids the pile up of particles. As an illustrative example, consider some acceleration centres throughout the source. These accelerate particles over a very short timescale. Immediately after being accelerated, the particles leave the centre and travel (more or less in random directions) and cool. After some time, they arrive at another acceleration centre, where they are reaccelerated. The minimum energy of the particles corresponds to the mean particle travel time from one acceleration centre to another. In our case, as the (comoving) cooling time is of the order of 10^{-5} s (see Eq. (7)), the average distance among the acceleration centres must be $\beta c t_{\rm cool} \sim 3\times10^5$ cm. One interesting possibility has been proposed by Sironi et al. (2016) who studied a scenario of magnetic reconnection in blobs that are accelerated within the jet by magnetic tension and can then further accelerate particles. However, in this case we again require that the spectrum is produced by the same particles that are re-used many times.

4.3. Mini-jets

Here we assume an emitting region that is at a large distance from the central engine, but is split into many mini-jets, and we are observing only one of them (e.g. Yamazaki et al. 2004; Zhang & Zhang 2014; Burgess et al. 2020, see also Giannios et al. 2009 for mini jets in blazars).

If the mini-jet is small enough, its emission could vary over a timescale compatible with what is commonly observed (10^{-3}–0.1 s). On the other hand, if this occurs then the emitting region is compact and its radiation energy density is large because all the synchrotron radiation has to be emitted in a small volume. This implies a dominating Self-Compton emission, which would inevitably imply very fast cooling of particles of all energies. This problem could be alleviated assuming that the mini-jets are emitting regions moving with a Lorentz factor $\Gamma'_{\rm mini-jet}=1$–10 as measured in the comoving frame of the outflow, itself moving with Γ with respect to the observer. This case is equivalent to mini-jets moving with a total $\Gamma_{\rm mini-jet}\sim\Gamma\Gamma'_{\rm mini-jet}$ (see e.g. Burgess et al. 2020). If mini-jets occupy a large fraction of the emitting volume, then there is no difference with the case of a unique jet moving at $\Gamma_{\rm mini-jet}$. If instead they occupy a small fraction of the available volume, they could explain fast variability, but the carried energy would be a small fraction of the total. Therefore, the efficiency (i.e. the ratio between the radiated and the total jet kinetic energy) would be smaller than what is usually assumed.

4.4. Break due to inverse Compton in Klein–Nishina regime

The hard low-energy spectrum of GRB emission could be due to the effects of inverse Compton scatterings occurring in the Klein–Nishina regime (Rees 1967) as suggested by Derishev et al. (2001), Nakar et al. (2009), and Daigne et al. (2011). These models are based on the idea that the inverse-Compton process is dominant in cooling the intermediate-energy electrons responsible for the low-energy X-rays before $\nu_{\rm peak}$. Electrons in this energy range cool at a rate γ^a, with $a<2$, hardening their energy distribution with respect to scatterings in the Thomson regime. Electrons at higher energies, which are responsible for the emission above $\nu_{\rm peak}$, cool only by synchrotron at a rate $\propto\gamma^2$.

These models work in a limited range of physical parameters because the inverse-Compton cooling is required to be reduced, but nevertheless important, before $\nu_{\rm peak}$, and negligible above. Even when this constraint is satisfied, the typical obtained spectral indices are $F_\nu\propto\nu^0$ (see e.g. Fig. 2 of Daigne et al. 2011). Harder spectra can be obtained if the adiabatic cooling is important, and they approach $F_\nu\propto\nu^{1/3}$ very rarely (see e.g. Fig. 4 of Daigne et al. 2011). In general, in this model, the inverse Compton flux must be important, while it is instead limited by the existing observations.

5. Proton–synchrotron

A possible solution to the problem of incomplete cooling of the emitting particles is to assume that what we see is synchrotron radiation produced by protons, not by leptons (see Gupta & Zhang 2007 for a discussion of lepton and hadronic models for the high-energy prompt emission in GRBs and Aharonian 2002 for the a proton–synchrotron model applied to blazars). Protons are accelerated efficiently in shocks, and should receive most of the shock energy, that is, more than the leptons. The typical synchrotron frequency emitted by protons is as follows, in the comoving frame:

$$\nu'_{\rm S,p}=\frac{4}{3}\frac{eB'}{2\pi m_{\rm p}c}\gamma^2\rightarrow\gamma=\left[\frac{3\pi\,\nu'_{\rm S,p}m_{\rm p}c}{2eB'}\right]^{1/2}\sim10^4\left[\frac{\nu'_{\rm S,p,keV}}{B'_6}\right]^{1/2}. \tag{16}$$

The total power emitted for a tangled magnetic field and an isotropic distribution of pitch angles is:

$$P_{\rm S,p}=\frac{4}{3}\sigma_{\rm T}c\left(\frac{m_{\rm e}}{m_{\rm p}}\right)^2\frac{B'^2}{8\pi}\gamma^2. \tag{17}$$

The synchrotron cooling time (in the observer frame) is:

$$\begin{aligned}t^{\rm obs}_{\rm cool,S,p}&=\frac{\gamma m_{\rm p}c^2}{P_{\rm S,p}}=\frac{6\pi m_{\rm p}c^2}{\sigma_{\rm T}cB'^2\gamma}\left(\frac{m_{\rm p}}{m_{\rm e}}\right)^2\frac{1+z}{\Gamma}\\&=\frac{6\pi m_{\rm e}c^2}{\sigma_{\rm T}cB'^{3/2}}\left(\frac{m_{\rm p}}{m_{\rm e}}\right)^{5/2}\left[\frac{2e}{3\pi\,c\nu^{\rm obs}_{\rm S,e}}\frac{1+z}{\Gamma}\right]^{1/2}\\&=t^{\rm obs}_{\rm cool,S,e}\left(\frac{m_{\rm p}}{m_{\rm e}}\right)^{5/2}\sim1.44\times10^8t_{\rm cool,S,e}\ \text{ for the same }\nu^{\rm obs}_{\rm S}.\end{aligned} \tag{18}$$

Comparing with the electron synchrotron cooling timescale of Eq. (7) we have values close to one second, as observed. Having an observed cooling timescale of approximately one second for particles emitting at the observed frequency of 100 keV then requires:

(A) emitting electrons: a weakly magnetised ($B'\sim1$ G, to avoid extremely fast synchrotron cooling) and very large ($R\gtrsim10^{16}$ cm, to avoid overly fast self-Compton cooling) emitting region; or

(B) emitting protons: a standard magnetic field and emitting region size, namely $B'\sim10^6$ G and $R\sim10^{13}$ cm.

5.1. Maximum frequency in proton–synchrotron

Guilbert et al. (1983) suggested that for shock-accelerated electrons, there is a maximum synchrotron frequency that can be emitted, independent of the random Lorentz factor γ and the magnetic field B. The argument was that at each shock crossing, the electrons double their energy, until their gyro-radius becomes so large that synchrotron cooling limits the maximum attainable γ.

We can repeat the original argument for protons[2]. We have that ($\beta=\sin\theta\sim1$, where θ is the pitch angle):

$$\frac{\Delta\gamma}{\gamma\Delta t}=\frac{1}{\Delta t}\rightarrow\dot\gamma_{\rm h}=\frac{\gamma}{2\pi r_{\rm L}/c}=\frac{eB}{2\pi mc}, \tag{19}$$

[2] In this subsection all quantities are considered in the comoving frame.

where we set $\Delta t = 2\pi r_{\rm L}/c$, and $r_{\rm L} = \gamma mc^2/(eB)$ is the Larmor radius. The synchrotron cooling rate is:

$$\dot{\gamma}_{\rm cool,S} = \frac{2}{3}\frac{e^4}{m^3c^5}\gamma^2 B^2. \tag{20}$$

Equating Eq. (19) with Eq. (20) we have:

$$\frac{eB}{2\pi mc} = \frac{2}{3}\frac{e^4}{m^3c^5}\gamma^2 B^2 \rightarrow \left[\gamma^2 B\right]_{\rm max} = \frac{3}{4\pi}\frac{m^2c^4}{e^3}. \tag{21}$$

Therefore, the maximum synchrotron frequency is:

$$\begin{aligned} h\nu_{\rm s,max} &= \frac{4}{3}\frac{he}{2\pi mc}\left[\gamma^2 B\right]_{\rm max} = \frac{1}{2\pi^2}\frac{hmc^3}{e^2} \\ &= 22\,{\rm MeV}, \quad \text{for electrons} \\ &= 41\,{\rm GeV}, \quad \text{for protons.} \end{aligned} \tag{22}$$

5.2. Total energy and number of emitting particles

In the standard scenario, the emitting particles are accelerated at the shocks and cool, and are not re-accelerated. Therefore, the total number of particles $\mathcal{N}_{\rm iso}$ contributing to the observed emission is:

$$\mathcal{N}_{\rm iso} \sim \frac{E_{\rm iso}}{\Gamma mc^2(\gamma_{\rm inj} - \gamma_{\rm cool})}. \tag{23}$$

This assumes that the slope of the injected distribution is $p > 2$. We now compare case A (electrons) and case B (protons) assuming in any case $\Gamma = 10^2\Gamma_2$.

Case A: electrons. From Eq. (6) the typical Lorentz factor $\gamma_{\rm cool}$ of the electrons emitting at $\nu_{\rm cool}$ is

$$\gamma_{\rm cool} = 2.5\times10^4\left[\frac{\nu^{\rm obs}_{\rm cool,keV}}{B'}\frac{(1+z)}{\Gamma_2}\right]^{1/2}. \tag{24}$$

This leads to a total number of emitting electrons:

$$\mathcal{N}_{\rm e,iso} \sim 4.9\times10^{52}\frac{E_{\rm iso,53}}{(\gamma_{\rm inj}/\gamma_{\rm cool}-1)}\left[\frac{B'}{\nu^{\rm obs}_{\rm cool,keV}\Gamma_2(1+z)}\right]^{1/2}. \tag{25}$$

Observationally, the break $\nu_{\rm b}$, interpreted as the cooling break $\nu_{\rm cool}$, is a factor of approximately ten smaller than $\nu_{\rm peak}$. This corresponds to $\gamma_{\rm inj}/\gamma_{\rm cool} \sim 3$.

The ratio between the total kinetic energy $E_{\rm K,iso,before}$ (calculated before the prompt emission) and the radiated energy $E_{\rm iso}$ is:

$$\frac{E_{\rm K,iso,before}}{E_{\rm iso}} = \frac{(\gamma_{\rm inj} + m_{\rm p}/m_{\rm e})}{\gamma_{\rm inj} - \gamma_{\rm cool}}. \tag{26}$$

This assumes that there is one cold proton per emitting electron. The same ratio after the prompt emission is:

$$\frac{E_{\rm K,iso,after}}{E_{\rm iso}} = \frac{(\gamma_{\rm cool} + m_{\rm p}/m_{\rm e})}{\gamma_{\rm inj} - \gamma_{\rm cool}}. \tag{27}$$

Case B: protons. In this case we assume $B' = 10^6 B'_6$ G. From Eq. (16) we have that protons emitting at 1 keV have $\gamma \sim 10^4$. From Eq. (23), the total number of emitting protons producing $E_{\rm iso}$ is:

$$\mathcal{N}_{\rm p,iso} \sim 6.9\times10^{49}\frac{E_{\rm iso,53}}{(\gamma_{\rm inj}/\gamma_{\rm cool}-1)}\left[\frac{B'_6}{\nu^{\rm obs}_{\rm cool,keV}\Gamma_2(1+z)}\right]^{1/2}. \tag{28}$$

In terms of total mass, this corresponds to only $M = N_{\rm p,iso}m_{\rm p} \sim 5.5\times10^{-8}\,M_\odot$.

In this case we can also calculate the ratio between the total kinetic energy (before the prompt emission) and the radiated energy $E_{\rm iso}$. Assuming that the leptonic component is unimportant we have:

$$\frac{E_{\rm K,iso,before}}{E_{\rm iso}} = \frac{1}{1-\gamma_{\rm cool}/\gamma_{\rm inj}} \sim \frac{3}{2}. \tag{29}$$

The same ratio after the prompt emission is:

$$\frac{E_{\rm K,iso,after}}{E_{\rm iso}} = \frac{1}{\gamma_{\rm inj}/\gamma_{\rm cool}-1} \sim \frac{1}{2}. \tag{30}$$

This indicates that the maximum energy emitted by the afterglow is approximately one-half of the energetics of the prompt. This implicitly assumes that the Poynting flux is not the dominant form of power that can be converted into radiation. In the opposite case, we should include the magnetic energy when calculating the fraction of the total jet power that can be converted into radiation, both in the prompt and the afterglow phases.

Since the emitting protons have $\gamma \gtrsim 10^4$, which is greater than Γ, it is not possible that they derive their energy from the conversion of bulk kinetic energy into random energy, unless only a minority of protons are accelerated at the expense of a much larger population of cold protons. This requires a not-yet-specified mechanism able to channel a fraction of the total bulk kinetic energy into a few selected protons.

Another more likely possibility could be a partial magnetic reconnection of a dominant magnetic field. In this case we would have a magnetically dominated flow with a small baryon loading, and we would expect three possible observational consequences. The first is the absence of a thermal prompt emission, the "fossil" radiation remaining after the conversion of the internal energy into bulk motion (see, e.g. Daigne & Mochkovitch 2002). The second is polarisation of the prompt emission, if part of the magnetic field, besides being dominant, is also ordered (see e.g. Lyutikov et al. 2003). The third consequence is a weak or absent reverse shock when the flow starts to decelerate (see e.g. Nakar & Piran 2004).

5.3. Electron–synchrotron versus proton–synchrotron

For illustration, we consider the case in which the number of injected electrons and protons is the same. We also consider that the observed spectrum is due to the proton–synchrotron process. We then ask if the emission produced by electrons can contribute to the observed prompt flux. Two cases are considered:

1. Electrons and protons are injected with the same random Lorenz factor distribution.
2. Electrons and protons are injected with the same energy distribution.

In case (1), the total injected power associated to the electron would be a factor $m_{\rm p}/m_{\rm e}$ smaller, making the electron–synchrotron luminosity negligible with respect to the proton–synchrotron one. Furthermore, the typical frequencies emitted by electron–synchrotron would be larger by a factor of $m_{\rm p}/m_{\rm e}$ with respect to the proton–synchrotron case.

In case (2), if a similar number of electrons and protons are injected with the same typical energies, then the two kinds of bolometric luminosities would also be equal, but the random Lorentz factors of the electrons would be $m_{\rm p}/m_{\rm e}$ times larger. The typical electron–synchrotron frequencies would be a factor $(m_{\rm p}/m_{\rm e})^3$ larger. Here we are assuming that the argument leading to a maximum synchrotron emitted frequency does

not apply, requiring an acceleration mechanism different from shocks. In this case it is likely that this extremely high-energy emission ($\sim 10^3$ TeV) would produce a pair cascade, partly inside the emitting region, and party outside. The fraction of luminosity absorbed within the emitting region would reprocess the power to smaller energies, but a detailed calculation is needed to quantify this statement. The fraction of high-energy photons that escape the source can pair-produce in the intergalactic medium interacting with the cosmic background light. In this case the luminosity initially collimated into the jet angle is dispersed because the produced pairs would be de-collimated by the intergalactic magnetic field. It is then likely that the reprocessed light would not contribute to the observed flux.

5.4. Radiative cooling and adiabatic timescales

The proton-synchrotron scenario can work because the radiative cooling timescale for protons is much longer than for leptons, and this can imprint a signature in the spectrum (the cooling break at ν_C). On the other hand, one might wonder how we can have a very fast variability (tens of milliseconds) in this scenario. The answer lies in the adiabatic timescale, $t_{ad} \sim R/(\Gamma^2 c)$, which is of the same order of the minimum variability timescale. After t_{ad}, the size of the emitting region roughly doubles, all particle energies halve, and the normalisation of the particle distribution decreases (to conserve the number of emitting particles), along with the magnetic field. As a result, the emitting flux, even if the radiative cooling is not particularly severe, is bound to decrease. The ν_C break continues to evolve (becoming smaller) but the flux decreases, making this evolution difficult to observe. In addition, when using a relatively long exposure timescale, we can see the superposition of several events, each lasting for t_{ad}. If all these events have a similar ν_C we will observe a non-evolving break frequency (as in the case of GRB 160625B discussed in Ravasio et al. 2018). Instead, if the flux is produced by a unique shell, spectra taken at different times should show an evolving ν_C, decreasing in time at least as t^{-2} (or faster, if the magnetic field is decreasing as well). This should be most visible during the decay phase of a pulse. We plan to investigate this interesting issue in a future study.

6. Conclusions

We show that the recent observations of a low-energy break in the prompt spectrum of GRBs, accompanied by observations of the slopes below and above the break, strongly suggest that the emission process is synchrotron originating from particles that cannot completely cool. This is at odds with our expectations about the properties of the emitting regions, which should be compact and then strongly magnetised. We show that the size of the emitting region should be relatively large in order to to avoid a strong self-Compton emission (and thus a severe cooling). Furthermore, the inferred lower limits on the size can start to significantly conflict with the limits posed by the onset of the afterglow.

In a leptonic scenario we find no simple solution to this problem. We consider this to be serious and to require an alternative explanation to the common and standard scenario we have considered up to now (i.e. emitting region located just beyond the transparency radius, with strong magnetic field, very small cooling times, and limited importance of the self–Compton emission).

One possibility that is capable of preserving the standard scenario is to assume that the radiation we observe is still synchrotron but produced by protons. Since their random energy exceeds the bulk one, this possibility likely requires that the dominant form of energy carried by the jet is magnetic. If the magnetic field is also ordered, then we expect a largely polarised prompt emission. A magnetically dominated jet should also imply a limited importance of any thermal component in the prompt emission, as well as a weak (or null) reverse shock at the start of the deceleration phase. The first simple estimates concerning the presence of emitting, ultra-relativistic protons are very promising, and we plan to further investigate their consequences in the near future.

Acknowledgements. We would like to thank Fabrizio Tavecchio for discussion and the anonymous referee for comments. We thank a ASI–*NuSTAR* grant and we acknowledge financial contribution from the agreement ASI-INAF n.2017-14-H.0, and from the PRIN–INAF 2016 and the PRIN-MIUR "FIGARO" grants.

References

Aharonian, F. A. 2002, MNRAS, 332, 215
Asano, K., & Terasawa, T. 2009, ApJ, 705, 1714
Band, D., Matteson, J., Ford, L., et al. 1993, ApJ, 413, 281
Bhat, P. N., Fishman, G. J., Meegan, C. A., et al. 1992, Nature, 359, 217
Blandford, R. D., & Znajek, R. L. 1977, MNRAS, 179, 433
Burgess, J. M., Preece, R. D., Connaughton, V., et al. 2014, ApJ, 784, 17
Burgess, J. M., Begue, D., Greiner, J., et al. 2020, Nat. Astron., 4, 174
Chand, V., Chattopadhyay, T., Oganesyan, G., et al. 2019, ApJ, 874, 70
Daigne, F., & Mochkovitch, R. 2002, MNRAS, 336, 1271
Daigne, F., Bosnjak, Z., & Dubus, G. 2011, A&A, 526, A110
Derishev, E. V., Kocharovsky, V. V., & Kocharovsky, VI, V. 2001, A&A, 372, 1071
Eichler, D., Livio, M., Piran, T., & Schramm, D. N. 1989, Nature, 340, 126
Ghirlanda, G., Celotti, A., & Ghisellini, G. 2002, A&A, 393, 409
Ghirlanda, G., Celotti, A., & Ghisellini, G. 2004, A&A, 422, L55
Ghirlanda, G., Bosnjak, Z., Ghisellini, G., Tavecchio, F., & Firmani, C. 2007, MNRAS, 379, 73
Ghirlanda, G., Pescalli, A., & Ghisellini, G. 2013, MNRAS, 432, 3237
Ghisellini, G., & Celotti, A. 1999, ApJ, 511, L93
Giannios, D., Uzdensky, D. A., & Begelman, M. C. 2009, MNRAS, 395, L29
Goldstein, A., Burgess, J. M., Preece, R. D., et al. 2012, ApJS, 199, 1
Golkhou, V. Z., & Butler, N. R. 2014, ApJ, 787, 90
Golkhou, V. Z., Butler, N. R., & Littlejohns, O. M. 2015, ApJ, 811, 93
Gruber, D., Goldstein, A., Weller von Ahlefeld, V., et al. 2014, ApJS, 211, 1
Guilbert, P., Fabian, A. C., & Rees, M. 1983, MNRAS, 205, 593
Guiriec, S., Connaughton, V., Briggs, M. S., et al. 2011, ApJ, 727, L2
Gupta, N., & Zhang, B. 2007, MNRAS, 380, 78
Kaneko, Y., Preece, R. D., Briggs, M. S., et al. 2006, ApJS, 166, 298
Katz, J. I. 1994, ApJ, 432, L107
Lien, A., Sakamoto, T., Barthelmy, S. D., et al. 2016, ApJ, 829, 1
Lyutikov, M., Pariev, V. I., & Blandford, R. D. 2003, ApJ, 597, 998
MacLachlan, G. A., Shenoy, A., Sonbas, E., et al. 2013, MNRAS, 432, 857
McBreen, S., Quilligan, F., McBreen, B., Hanlon, L., & Watson, D. 2001, A&A, 380, L31
Nakar, E., & Piran, T. 2004, MNRAS, 353, 647
Nakar, E., Ando, S., & Sari, R. 2009, ApJ, 703, 675
Nava, L., Ghirlanda, G., Ghisellini, G., & Celotti, A. 2011, A&A, 530, A21
Oganesyan, G., Nava, L., Ghirlanda, G., & Celotti, A. 2017, ApJ, 846, 137
Oganesyan, G., Nava, L., Ghirlanda, G., & Celotti, A. 2018, A&A, 616, A138
Oganesyan, G., Nava, L., Ghirlanda, G., et al. 2019, A&A, 628, A59
Pe'er, A., & Ryde, F. 2017, Int. J. Mod. Phys. D, 26, 1730018-296
Preece, R. D., Briggs, M. S., Mallozzi, R. S., et al. 1998a, ApJ, 506, L23
Preece, R. D., Pendleton, G. N., Briggs, M. S., et al. 1998b, ApJ, 496, 849
Ravasio, M. E., Oganesyan, G., Ghirlanda, G., et al. 2018, A&A, 613, A16
Ravasio, M. E., Ghirlanda, G., Nava, L., & Ghisellini, G. 2019, A&A, 625, A60
Rees, M. J. 1967, MNRAS, 137, 429
Rees, M. J., & Meszaros, P. 1994, ApJ, 430, L93
Ronchi, M., Fumagalli, F., Ravasio, M. E., et al. 2020, A&A, 636, A55
Ryde, F., & Pe'er, A. 2009, ApJ, 702, 1211
Ryde, F., Axelsson, M., Zhang, B. B., et al. 2010, ApJ, 709, L172
Sironi, L., Giannios, D., & Petropoulou, M. 2016, MNRAS, 462, 48
Tavani, M. 1996, ApJ, 466, 768
Walker, K. C., Schaefer, B. E., & Fenimore, E. E. 2000, ApJ, 537, 264
Yamazaki, R., Ioka, K., & Nakamura, T. 2004, ApJ, 607, L103
Yu, H.-F., Preece, R. D., Greiner, J., et al. 2016, A&A, 588, A135
Zalamea, I., & Beloborodov, A. M. 2011, MNRAS, 410, 2302
Zhang, B., & Zhang, B. 2014, ApJ, 782, 92