1 Introduction

The journey of exploring the physical mechanism of gamma-ray bursts (GRBs) should start with an introduction about what GRBs are, how this field came to be, as well as why this subject is full of excitement.

1.1 What Are Gamma-Ray Bursts?

By definition, a *gamma-ray burst (GRB)* is a burst of γ-rays. In transient astrophysics, people usually describe electromagnetic signals in both temporal and spectral domains. The term "gamma-ray burst" clearly carries both pieces of information. To be more specific, a "burst" here means a sudden release of emission that lasts from milliseconds to thousands of seconds, and the term "gamma-ray" stands for the energy range from tens of keV to several MeV, which is the typical bandpass of spaceborne GRB detectors (e.g. Burst And Transient Source Experiment (BATSE) on board the *Compton Gamma-Ray Observatory (CGRO)*, Burst Alert Telescope (BAT) on board *The Neil Gehrels Swift Observatory* (here after *Swift*),[1] and Gamma-ray Burst Monitor (GBM) on board the *Fermi* Gamma-Ray Space Telescope). This bursty emission in the hard-X-ray/soft-γ-ray band is usually called the GRB *prompt emission*.

Although bright GRBs typically have most energy output (characterized by a parameter E_p, which is the peak of the energy spectrum of a GRB) in the sub-MeV to MeV range, it is now clear that the GRB phenomenology includes events with a wide distribution of E_p. Some less luminous GRBs are *X-ray rich GRBs* (with E_p below 50 keV). Some others are even softer, with E_p below 30 keV. In the literature these soft events are also called *X-ray flashes (XRFs)*.[2] Extensive multi-wavelength observations suggest that XRFs are not a different population from GRBs, but rather the natural extension of GRBs to the softer, less luminous regime.

For years, observations of GRBs were limited to the "burst" phase in the temporal domain, and the "γ-ray" range in the spectral domain. A breakthrough was made in 1997, when a long-lasting, multi-wavelength *afterglow* of a GRB was discovered. Currently, in the temporal domain, emission from a GRB source can be observed minutes, hours, days,

[1] The *Swift Gamma-Ray Burst Mission* was renamed as *The Neil Gehrels Swift Observatory* on January 10, 2018, in honor of the mission Principal Investigator Neil Gehrels (1952–2017).

[2] The term "X-ray burst" has already been reserved to describe bursts of X-rays from Galactic accreting neutron star systems.

weeks, months, or even years after the burst itself. In the spectral domain, GRBs have been detected in radio, millimeter (mm), infrared (IR), optical, ultraviolet (UV), X-rays, and γ-rays up to >100 GeV. Different from other astrophysical objects, which could be observed in any wavelength at any time, GRBs must be observed as soon as possible due to the rapidly fading nature of the afterglow. Throughout the history of GRB research, breakthroughs in understanding were made whenever a new temporal or spectral window opened.

Besides being strong emitters across the entire electromagnetic spectrum, GRBs are also believed to be sources of non-electromagnetic signals, including cosmic rays (in particular, ultra-high-energy cosmic rays (UHECRs) observed from Earth), neutrinos (from MeV [10^6 eV] all the way to EeV [10^{18} eV]), and gravitational waves. There have been great efforts in directly detecting high-energy neutrinos and gravitational waves from GRBs, and one direct association between a GRB and a gravitational wave event has been made.

Physically, GRBs are the most luminous explosions in the universe. With robust measurements of their distances/redshifts (from 40 Mpc to redshift 9.4) as of 2018, the typical isotropic γ-ray luminosity is $\sim 10^{51}$–10^{53} erg s^{-1}. This is the total energy released by the Sun in its entire lifetime emitted within less than 1 second! For comparison, the luminosity of the Sun is $\sim 10^{33}$ erg s^{-1}, and the total star-light luminosity of the Milky Way Galaxy is $\sim 10^{44}$ erg s^{-1}. Even the most energetic Active Galactic Nuclei (AGNs) powered by accreting super-massive black holes only have a luminosity $\sim 10^{48}$ erg s^{-1}, which is dwarfed by GRBs. Indeed, GRBs are the most luminous explosions in the universe since the Big Bang.

How are these tremendous bursts of γ-rays generated? The total energetics of the events (comparable to the supernova energy) as well as the rapid variability of time scales (as short as milliseconds) suggest that GRBs must be related to catastrophic events on the stellar scale (in contrast to AGNs, which are on the galactic scale). Multi-wavelength observations now reveal at least two distinct physical origins of cosmological GRBs. One type (typically having long durations) are believed to be associated with deaths of some special massive stars – direct evidence being that at least some of them (maybe the majority) are associated with a special type, i.e. the broad-line Type Ic, of supernovae. The second type (typically having short durations) are not associated with supernovae, and often reside in the regions in their host galaxies with little star formation. They are very likely not associated with deaths of massive stars, but rather associated with compact objects such as neutron stars and black holes. The leading scenario is mergers of two neutron stars (NS–NS) or one neutron star and one stellar-size black hole (NS–BH). In either the massive star type or compact star type, the catastrophic event leaves behind a hyper-accreting stellar-size black hole or a rapidly rotating highly magnetized neutron star (millisecond magnetar), which serves as the engine of a collimated outflow (jet) with a relativistic speed. When the jet beams towards Earth, we detect a GRB event.

As stellar-scale events located at cosmological distances, GRBs make a unique connection among various branches of astrophysics. Their high-energy, high-velocity, strong-gravity, and strong-magnetic-field environment guarantee rich physics. GRBs thus provide an important cosmic laboratory to study physics in extreme conditions.

After decades of observational and theoretical studies, one finally reaches a physical picture regarding the origin of GRBs. Even though many details remain unclear, a general theoretical framework is set up, which is found to be successful in interpreting the

multi-wavelength data. On the other hand, due to their elusive nature and the technological challenges in observing them, GRBs have not been observed in all wavelengths at all epochs. As a result, this field has been and will remain a hot subject in contemporary astrophysics.

1.2 Brief History of GRB Research

The history of GRB research has been full of struggles, surprises, and excitement. A detailed description of the bumpy journey towards understanding mysterious GRBs can be found in other books (e.g. Vedrenne and Atteia, 2009; Kouveliotou et al., 2012). Here we only briefly outline the key milestones in the development of GRB science on both the observational and theoretical fronts.

1.2.1 Observational Progress

In astronomy, especially in the field of GRBs, our understanding of phenomena usually enjoys a leap whenever new observational data flood in. In this particular field, progress has been made in a discrete manner. This is because key observational breakthroughs can be made only when new detectors/telescopes, especially those that are spaceborne, come into use. Before a new generation of instruments are put into use (which usually requires ripe new technology and, more importantly, highly competitive funding), observational progress remains at a certain "quantum" level for a while. As a result, a review of the history of the GRB research can easily be placed in a framework defined by observational facilities. Reviews on the observational progress in different eras can be found in, e.g. Fishman and Meegan (1995); van Paradijs et al. (2000); Gehrels et al. (2009).

The Discovery Era (1967–1973)

The discovery of the first GRBs was made in the late 1960s by the military satellite system *Vela*. These were a series of satellites launched by the United States to monitor compliance with the Nuclear Test Ban Treaty signed by the United States, the United Kingdom, and the Soviet Union. There were multiple launches, each placing a pair of satellites in a common circular orbit, to monitor possible denotations of nuclear bombs in space. Starting from the third launch, a more sensitive γ-ray scintillator was implemented. This led to the first detections of bursts of γ-rays which we now call GRBs. The very first GRB ever detected was caught by *Vela IVa,b* (see Fig. 1.1) on July 2, 1967. Even with poor temporal resolution, GRB 670702[3] already showed the basic features of GRBs: a duration of about 10 seconds, a lightcurve with significant structure (two emission episodes with different peak fluxes and asymmetric pulse profiles), and a peak energy around MeV. It took some time for Raymond Klebesadel and colleagues to clean up the background and confirm the

[3] The convention of naming a GRB is based on its detected year, month, and day. If multiple bursts are detected on the same day, an additional letter "A", "B", ... is added based on the sequence of their detections.

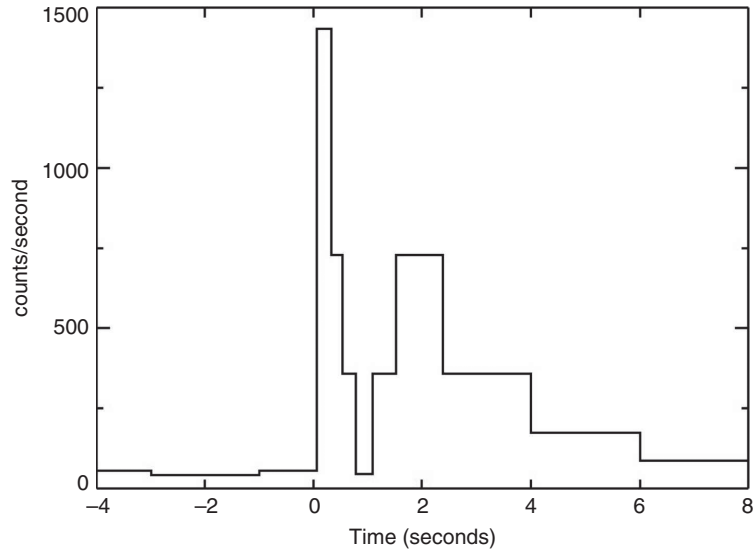

Figure 1.1 The lightcurve of the very first GRB detected on 2 July 1967 with the *Vela IVa* satellite. From Kouveliotou et al. (2012).

astrophysical origin of these events. The first paper reporting the discovery of GRBs was published almost 5 years later (Klebesadel et al., 1973). It has been commonly suspected that the authors had to wait until the data were declassified, but according to Klebesadel (Chapter 1 of Kouveliotou et al., 2012), the delay was due purely to the complicated data analysis process. The GRB data were not regarded as classified materials from the very beginning, since the observed properties (duration, spectrum, variability) were completely different from what one expected from a nuclear test in space, which would produce a millisecond duration hard X-ray flash with no significant time structure.

In the same era, besides being seen by *Vela*, GRBs were also detected by the American solar satellite *Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI)* (Cline et al., 1973) and the Soviet space satellite Konus/*Venera* (Mazets et al., 1974).

The Dark Era (1973–1991)

Since the announcement of the discovery of GRBs and until the launch of *CGRO* in 1991, the pace of understanding of the origin of GRBs was slow. During this period, about 500 GRBs were detected using several γ-ray detectors, including *Vela*, Konus/*Venera*, *Apollo 16*, *UHURU*, and *Ginga* (Higdon and Lingenfelter, 1990). The poor localization capability of γ-ray detectors made it very difficult to discover electromagnetic counterparts of GRBs in lower frequencies. Nonetheless, some tentative clues were collected. For example, the first Konus GRB catalog showed evidence of two duration categories (long and short) for GRBs (Mazets et al., 1981a). Low-significance spectral line features were reported in some GRBs detected with the Soviet Konus instruments on board the *Venera* satellite (Mazets et al., 1981b) and with the Japanese satellite *Ginga* (Murakami et al., 1988). Even though they were not confirmed by later missions, during the pre-BATSE era, these features greatly

motivated identifying Galactic neutron stars as the sources of GRBs. On the other hand, a rough isotropy and a deviation from a uniform spatial distribution of GRBs (deficit of low flux/fluence GRBs from the nominal $N(>S) \propto S^{-3/2}$ law defined by a Euclidean geometry, see §2.5.2) had been noticed (Higdon and Lingenfelter, 1990), which pointed towards a possible cosmological origin. In any case, the small amount of, sometimes controversial, data allowed theorists to free their imaginations to propose many models (or simply scenarios) to interpret GRBs. According to a historical review paper (Nemiroff, 1994), the total number of suggested GRB models by 1994 was 118. Many of these early models were reviewed by Malvin Ruderman at the Seventh Texas Symposium on Relativistic Astrophysics (Ruderman, 1975), who summarized the status of the theoretical efforts by 1975 and stated (references are omitted, which can be found in the original article):

> ... there has been no lack of response by the theoretic community in suggesting an enormous variety of models for γ-ray bursts, such as the following: expanding supernovae shocks, neutron star formation, glitches, neutron stars in close binaries, black holes in binaries, novae, white holes, flares on "normal" stars, flares on flare stars, flares on white dwarfs, flares on neutron stars, flares in close binaries, nuclear explosions on white dwarfs, comets on neutron stars, Jupiter, antimatter on conventional stars, magnetic bottles and instabilities in the solar wind, relativistic dust, vacuum polarization instabilities near rotating charged black holes, instabilities in pulsar magnetospheres, and "ghouls".[4]

He further noted:

> For theorists who may wish to enter this broad and growing field, I should point out that there are a considerable number of combinations, for example, comets of antimatter falling onto white holes, not yet claimed.

It looked like "a theorists' heaven" due to the lack of critical data to constrain models, but it could easily become "a theorists' hell". Ruderman concluded:

> The only feature that all but one (and perhaps all) of the very many proposed models have in common is that they will not be the explanation of γ-ray bursts.

Indeed he was right. None of the above suggested models turn out to be the correct interpretations of GRBs in the modern era. The closest one is the "supernova shock" scenario proposed by Stirling Colgate (Colgate, 1974), which may be relevant to some low-luminosity long-duration GRBs. We will discuss the journey of understanding the physics of GRBs in §1.2.2 below.

The BATSE Era (1991–1997)

The *CGRO* spacecraft was launched into a low Earth orbit by NASA's Space Shuttle *Atlantis* on 5 April 1991. BATSE was one of the four instruments on board *CGRO*, which was dedicated to detecting GRBs. It carried eight Large Area Detectors (LADs) to cover

[4] According to the Merriam-Webster Dictionary, the word "ghoul" stands for "a legendary evil spirit being that robs graves and feeds on corpses". In a model proposed by Zwicky, GRBs are produced by ejected "nuclear goblins", somehow propelled from inside of a certain type of parent neutron star, which explode to produce a burst of γ-rays.

the energy range 20 keV – 1.9 MeV and eight Spectroscopy Detectors (SDs) to cover the energy range 10 keV – 100 MeV. The field of view of BATSE was all sky, with a burst detection sensitivity of 3×10^{-8} erg cm^{-2} for a 1 s burst. The *CGRO* was de-orbited on 4 June 2000 after one of its three gyroscopes failed. However, since the year 1997 marked a new era in GRB research thanks to the discovery of GRB afterglow (see below), we define the BATSE era as the time span of 1991–1997.

Through its lifetime, BATSE detected 2704 GRBs. These GRBs have large localization error boxes, with a typical (Gaussian) angular error ranging from $\sim$0.2 degrees for the strongest bursts, to $\sim$18 degrees for the weakest ones (Briggs et al., 1999b). There are numerous objects within these error boxes, so despite great efforts, no low-frequency counterpart was robustly identified for GRBs within the BATSE error boxes before 1997. Nonetheless, great progress was made during the BATSE era in understanding the nature of GRBs. A comprehensive review was presented by Fishman and Meegan (1995). The most important progress of this era was in the following three directions.

- Even though the distances of GRB sources were still subject to debate in the BATSE era, BATSE already collected important clues to suggest a cosmological origin of GRBs. The angular distribution of GRBs was found to be highly isotropic (Briggs et al., 1996), and the intensity (fluence or peak flux) distribution was found to deviate from the simple prediction of Euclidean geometry at the faint end (Meegan et al., 1992). Both facts posed great constraints on available Galactic neutron star models, but can be trivially explained if GRBs originate from cosmological distances.
- Based on the duration distribution, two categories of GRBs were firmly identified (Kouveliotou et al., 1993). A separation line is roughly 2 seconds. The *long-duration GRBs* are on average softer than the *short-duration GRBs*. Hints of a long–short dichotomy had already been collected in the pre-BATSE era (e.g. Mazets et al., 1981a; Norris et al., 1984), but the BATSE data gave a more definitive differentiation between the two classes of GRBs.
- Even though GRB lightcurves are rather irregular, spectral analyses of BATSE GRBs revealed that the GRB spectra are non-thermal, and can usually be delineated by a smoothly joined broken power-law function known as the "*Band function*" or "GRB function" (Band et al., 1993).

The *BeppoSAX*/*HETE* Era (1997–2004)

The main barrier in understanding the nature of GRBs during the first 30 years was the lack of distance information. In order to make a breakthrough, counterparts at longer wavelengths were desired. Since the optical sky is very crowded, it was essentially impossible to identify a variable counterpart in the optical band within the large error box provided by BATSE. The X-ray sky is much less crowded. A wide-field X-ray camera with a much better localization capability than γ-ray detectors held the key to catching the counterparts of GRBs.

BeppoSAX was an Italian–Dutch satellite for X-ray astronomy (Piro et al., 1995). "Beppo" was the nickname of the Italian physicist Giuseppe "Beppo" Occhialini, in whose

honor the mission was dedicated,[5] and "SAX" stands for "Satellite per Astronomia a raggi X" in Italian (i.e. "Satellite for X-ray Astronomy"). It was launched on 30 April 1996 and de-orbited on 29 April 2003. Besides a set of Narrow Field Instruments (NFIs), it also carried a set of Wide Field Instruments (WFIs), including a Gamma-Ray Burst Monitor (GRBM: 40–700 keV) and two Wide Field Cameras (WFCs: 2–30 keV). These WFCs could promptly search within the large error boxes provided by GRBM and BATSE, to allow quick identification of a possible X-ray counterpart of a GRB. The NFIs had a higher sensitivity, and could be used to confirm the transient (fading) nature of the X-ray counterpart. Even though the original mission plan was to study a wide range of Galactic and extragalactic X-ray targets, the mission turned out to be most famous for its first detections of the X-ray afterglows of GRB 970228 and GRB 970508. This enabled the discovery of optical and radio afterglows and the identification of the host galaxies, revolutionizing the field by establishing the cosmological origin of GRBs (Costa et al., 1997; van Paradijs et al., 1997; Frail et al., 1997; Metzger et al., 1997).

High Energy Transient Explorer (HETE) was an American astronomical satellite with international participation (Japan and France). The primary objective was to detect the first multi-wavelength counterpart of GRBs. Unfortunately, the first *HETE* was lost during launch on 4 November 1996. A second *HETE* satellite, *HETE-2* (Ricker et al., 2003), was launched on 9 October 2000 and continued to deliver GRB data until early 2006. It carried a French Gamma-Ray Telescope (FREGATE: 6–400 keV), a Wide Field X-ray Monitor (WXM: 2–30 keV), and a Soft X-ray Camera (SXC: 0.5–10 keV).

Before *Swift* was launched, *BeppoSAX* and *HETE-2* provided precise localizations of more than 100 GRBs, which led to detections of their afterglows and measurements of their redshifts. As a result, many great achievements were made during this era.

- The first X-ray (Costa et al., 1997) and optical (van Paradijs et al., 1997) afterglows were discovered following the *BeppoSAX* burst GRB 970228; and the first radio afterglow was discovered (Frail et al., 1997) following the *BeppoSAX* burst GRB 970508. The GRB field formally entered the multi-wavelength afterglow era. The first redshift measurement was made for GRB 970508 ($z = 0.835$) (Metzger et al., 1997), which formally established the cosmological origin of long GRBs.
- The origin of long GRBs was solved: they originate from the death of a special category of massive stars. The first tentative evidence was the discovery of SN 1998bw, a Type Ic supernova in a nearby galaxy at $z = 0.0085$, in the error box of the *BeppoSAX* burst GRB 980425 (Galama et al., 1998; Kulkarni et al., 1998). Soon afterwards, a supernova red bump was discovered in the optical lightcurves of several other GRBs (e.g. Bloom et al., 1999; Galama et al., 2000). A few years later, a robust GRB–SN association was established for the *HETE-2* burst GRB 030329/SN 2003dh (Stanek et al., 2003; Hjorth et al., 2003) at $z = 0.167$. Later, a systematic study of the host galaxies of long GRBs suggested that long GRBs typically lie in the most active star-forming regions in star-forming galaxies (Fruchter et al., 2006).

[5] Giuseppe "Beppo" Occhialini (1907–1993) was an Italian physicist who contributed to the discovery of the pion (π-meson) decay in 1947, and to the foundation of the European Space Agency.

- The abundant multi-wavelength afterglow data allowed in-depth understanding of the physics of GRBs. The power-law decay behavior of multi-wavelength afterglows was found to be consistent with the predictions of the fireball forward shock model (Rees and Mészáros, 1992; Mészáros and Rees, 1993b, 1997a; Sari et al., 1998). The early optical flash detected in GRB 990123 (Akerlof et al., 1999) was comfortably interpreted with a reverse shock model (Mészáros and Rees, 1997a; Sari and Piran, 1999b,a; Mészáros and Rees, 1999). A steepening temporal break was identified in the afterglow lightcurves of several GRBs, which was successfully attributed to collimation of GRB jets (Rhoads, 1999; Sari et al., 1999). Limited data led to the suggestion that GRBs, despite different degrees of collimation, may have a standard energy reservoir (Frail et al., 2001; Bloom et al., 2003; Berger et al., 2003b) and possibly a quasi-universal jet structure (Zhang and Mészáros, 2002b; Rossi et al., 2002; Zhang et al., 2004a).
- Diverse long GRBs were observed and studied. Besides the traditional long GRBs, softer X-ray rich GRBs and even softer X-ray flashes were regularly observed in the *BeppoSAX* and *HETE-2* era (Heise et al., 2001; Kippen et al., 2001; Sakamoto et al., 2005). These objects seem to form a continuum with traditional GRBs in the softer, less energetic regime. Based on optical follow-up observations, GRBs were found to fall into optically bright and optically dark categories. The "dark" ones made up a significant fraction of GRBs.

The *Swift* Era (2004–)

The *Swift* observatory (Gehrels et al., 2004) was launched on 20 November 2004. Built by an international team from the USA, UK, and Italy, it carries three instruments: a wide-field Burst Alert Telescope (BAT; Barthelmy et al., 2005c), a narrow-field X-Ray Telescope (XRT; Burrows et al., 2005b), and a UV-Optical Telescope (UVOT; Roming et al., 2005). The BAT (15–350 keV) is a coded aperture hard X-ray imager, with 1.4 sr field of view. The XRT has a field of view $23'.6 \times 23'.6$, which is large enough to search for an X-ray counterpart in the BAT error box (typically a few arc-minutes). With a typical slew time less than one minute, and a sensitivity $\sim 2 \times 10^{-14}$ erg cm^{-2} s^{-1} in 10^4 s, XRT can quickly catch the X-ray afterglow of the majority of detected GRBs and provide an accurate position with a point spread function (PSF) half-power diameter of $18''$. The UVOT has a 30 cm aperture, a 170–650 nm bandpass, and a field of view $17' \times 17'$. With a typical slew time less than two minutes and a sensitivity down to magnitude 23 in white light in 10^3 s, it can quickly catch the UV/optical counterpart of a GRB and provide a PSF of $1.9''$ at 350 nm. The accurate positions provided by XRT are promptly distributed to the ground-based and other spaceborne follow-up telescopes through the Gamma-ray Coordinates Network (GCN: http://gcn.gsfc.nasa.gov/gcn/), so that they can also promptly search for counterparts in other wavelengths.

Swift turned out an extremely successful GRB mission. The prompt slewing capability of XRT and UVOT allowed detections of the afterglows of the majority of detected GRBs. It enabled direct observations of the very early afterglow phase of GRBs. As a result, the field was revolutionized in many aspects:

- *Swift* made it possible to detect the faint afterglow of short-duration GRBs. This led to the identifications of the host galaxies of several short GRBs in 2005 (GRB 050509B, GRB 050709 [detected with HETE-2], and GRB 050724) and their relative locations with respect to the host. The results are very different from those of long GRBs, suggesting that short GRBs are from a different population (Gehrels et al., 2005; Bloom et al., 2006; Barthelmy et al., 2005a), likely not associated with the deaths of massive stars. Rather, they might be related to compact stars. The leading model is the coalescence of two neutron stars (NS–NS) or one neutron star and one black hole (NS–BH). Notice that *HETE-2* also contributed to the detections of short GRB afterglows (Villasenor et al., 2005; Fox et al., 2005).
- Interestingly, later observations by *Swift* suggested that the separation between the long and short populations is not clean. Two nearby "apparently" long-duration GRBs (GRB 060614 and GRB 060505) were detected by *Swift* (Gehrels et al., 2006), but deep searches of an associated supernova only placed a very stringent upper limit on the SN light (Gal-Yam et al., 2006; Fynbo et al., 2006a; Della Valle et al., 2006). Other arguments suggest that they might belong to the physical category of short-duration GRBs (Gehrels et al., 2006; Zhang et al., 2007b). On the other hand, some short or "rest-frame" short GRBs (e.g. GRB 080913, GRB 090423, and GRB 090426) were found to be more consistent with the long-duration population (e.g. Greiner et al., 2009; Levesque et al., 2010). As a result, the duration classification scheme is no longer clean, and multi-wavelength criteria are needed to diagnose the physical category of a particular GRB (Zhang et al., 2009a).
- The abundant early afterglow data, especially in the X-ray band (Nousek et al., 2006; O'Brien et al., 2006; Evans et al., 2009), allowed one to diagnose the physical processes that shape the early afterglow lightcurves (Zhang et al., 2006). A *canonical X-ray afterglow lightcurve* was identified, which displays five distinct temporal components (Zhang et al., 2006). In particular, erratic X-ray flares were discovered to follow the prompt γ-ray emission in nearly half of the GRBs, suggesting that the GRB central engine lasts longer than previously believed. Further multi-wavelength observations revealed a more complex "chromatic" behavior for at least some GRBs, suggesting more complicated afterglow physics.
- *Swift* greatly broadened the redshift range of GRBs. In the low-redshift regime, *Swift* discovered several *low-luminosity GRBs* associated with supernovae (e.g. GRB 060218/SN 2006aj, Campana et al. 2006; Pian et al. 2006; GRB 100316D/SN 2010bh, Starling et al. 2011). The results suggested that low-luminosity GRBs likely form a population distinct from high-luminosity GRBs (Liang et al., 2007a; Virgili et al., 2009; Bromberg et al., 2012). In the high-redshift regime, *Swift* continued to break the redshift record of GRBs: GRB 050904 at $z = 6.29$ (Cusumano et al., 2006; Totani et al., 2006), GRB 080913 at $z = 6.7$ (Greiner et al., 2009), GRB 090423 at $z = 8.2$ (Tanvir et al., 2009; Salvaterra et al., 2009), and GRB 090429B at $z = 9.4$ (Cucchiara et al., 2011a). Detecting GRBs in a wide redshift range allows them to be used as probes to study the evolution of the universe.
- *Swift* continues to prove that there is an unbound discovery space in transient astronomy. Every now and then, a surprising discovery is made by *Swift*. The following are a few

examples: GRB 060218 showed very different radiation signatures (e.g. long duration, smooth lightcurve, a thermal X-ray component in the time-resolved spectra, puzzling UV emission; Campana et al., 2006) from the traditional *high-luminosity GRBs*, suggesting a possible different physical origin (e.g. shock breakout, Campana et al. 2006); GRB 060614 (Gehrels et al., 2006) suggested that the simple long–short classification scheme cannot fully describe the physical origin of GRBs (Zhang, 2006); the serendipitous discovery of an X-ray outburst source associated with a Type Ic supernova, i.e. XRO 080109/SN 2008D, with the *Swift* XRT suggested that it is possible that every SN may have an associated X-ray outburst, possibly due to the breakout of the SN shock from the star (Soderberg et al., 2008); GRB 080319B (Racusin et al., 2008) suggested that a GRB can have a prompt optical flash detectable by the naked eye; the "Christmas" burst GRB 101225 was extremely long, had observational properties difficult to interpret with known GRB models (Thöne et al., 2011; Campana et al., 2011), and probably represents the prototype of a class of *ultra-long GRBs* with a possible different progenitor (e.g. Levan et al., 2014b, but see Zhang et al., 2014); the so-called "GRB 110328" (later renamed as "Swift J164449.3+573451" or simply "Sw J1644+57") was soon recognized as not a traditional GRB; rather, it signaled a new type of relativistic jet powered by tidal disruption events (TDEs) by super-massive black holes (Bloom et al., 2011; Burrows et al., 2011; Levan et al., 2011; Zauderer et al., 2011).

The *Fermi* Era (2008–)

While *Swift* continues to make new discoveries, another NASA γ-ray mission, the *Fermi* Gamma-Ray Space Telescope (FGST) was launched on 11 June 2008. It carries two main instruments: a Large Area Telescope (LAT: 20 MeV – 300 GeV), which covers 20% of the sky at any time and scans the entire sky every three hours, and a Gamma-ray Burst Monitor (GBM: 8 keV – 40 MeV), which monitors the whole sky for any burst events. The two instruments cover more than 7 orders of magnitude in energy, and have made it possible to study the broad-band spectra of GRB prompt emission in unprecedented detail.

The *Fermi* GRB data, especially the LAT high-energy data, greatly advanced our understanding of GRB emission physics.

- According to the first *Fermi* LAT GRB catalog (Ackermann et al., 2013), LAT detected 28 GRBs above 100 MeV out of 733 GRBs detected by GBM. This is about 4%. For those detected, the LAT-band emission usually lasts longer than the GBM-band emission. This points towards an external shock origin of the observed >100 MeV emission of GRBs, at least after the prompt emission phase (GBM emission is over) (e.g. Kumar and Barniol Duran, 2009, 2010; Ghisellini et al., 2010; He et al., 2011; Liu and Wang, 2011; Maxham et al., 2011).
- GeV emission was found to have a delayed onset with respect to the MeV emission, at least in some GRBs. This was not predicted from known models, but stimulated great efforts of theoretical modeling. Such delays (or the lack of) for photons with the highest

energies place important constraints on the Lorentz Invariance Violation (LIV) (Abdo et al., 2009a,c).

- Unprecedented, detailed spectral analyses in a wide spectral window provided important information to understand the composition of GRB jets and the physical mechanisms of prompt emission. The first bright LAT GRB 080916C (Abdo et al., 2009c) showed near featureless, time-resolved spectra covering nearly 7 orders of magnitude,[6] which are in contrast to the predictions of the standard fireball internal shock model, calling for a modification of the basic theoretical framework (Zhang and Pe'er, 2009). Later observations of GRBs 090510, 090902B, and 090926A (Abdo et al., 2009a,b; Ackermann et al., 2010) revealed more complicated spectral features, suggesting that the observed GRB spectra are the superposition of at least three different spectral components (Zhang et al., 2011; Guiriec et al., 2015): besides the traditional non-thermal Band-function component, a quasi-thermal component was found in some GRBs, being either dominant (e.g. GRB 090902B, Ryde et al. 2010; Zhang et al. 2011) or sub-dominant (e.g. Guiriec et al., 2011; Axelsson et al., 2012; Guiriec et al., 2013). A third power-law component extending to high energies (and probably also to low energies) was discovered in several GRBs (e.g. GRBs 090902B, 090510, and 090926A, Abdo et al. 2009b,a; Ackermann et al. 2010). Its physical origin is a mystery.
- Photons with rest-frame energy greater than 100 GeV have been detected in several GRBs (GRB 080916C, Atwood et al. 2013; GRB 090510, Abdo et al. 2009a; and GRB 130427A, Ackermann et al. 2013). These photons posed important constraints on GRB physics, including bulk Lorentz factor, particle acceleration mechanisms in relativistic shocks, and radiation mechanisms of relativistic particles. They are also used to study the extragalactic background light (EBL), which is expected to attenuate high-energy photons through two-photon pair production (e.g. Razzaque et al., 2009).

The Multi-Messenger Era (2017–)

It has been believed that GRBs are emitters of high-energy neutrinos and gravitational waves. The high-energy neutrino telescope at the South Pole, the *IceCube* Neutrino Observatory, has been searching for coincident ∼TeV–PeV (10^{12}–10^{15}) neutrinos from GRBs. As of 2018 no positive detection has been made, and progressively stringent upper limits on the neutrino flux from GRBs have been reported (Abbasi et al., 2010, 2012; Aartsen et al., 2015, 2016, 2017a,b). The upper limits posed interesting constraints on GRB physics (e.g. Abbasi et al., 2012; He et al., 2012; Zhang and Kumar, 2013).

The detections of gravitational waves due to BH–BH mergers with the gravitational wave (GW) detector, *Advanced LIGO* (aLIGO) (http://www.ligo.caltech.edu), opened the new era of GW astronomy (Abbott et al., 2016c,b, 2017c). On 17 August 2017, a NS–NS merger event, GW170817, was detected by the *Advanced LIGO* and *Advanced Virgo* gravitational wave detectors (Abbott et al., 2017d). The event was associated with a

[6] A later re-analysis of the burst (Guiriec et al., 2015) revealed a sub-dominant thermal component in the time-resolved spectra, but its amplitude is too low to be interpreted within the standard fireball internal shock models.

low-luminosity short GRB 170817A (Abbott et al., 2017b) and a multi-wavelength counterpart detected in optical, radio, and X-ray bands (e.g. Coulter et al., 2017; Pian et al., 2017; Evans et al., 2017; Shappee et al., 2017; Smartt et al., 2017; Nicholl et al., 2017; Chornock et al., 2017) in a nearby galaxy NGC 4993 at ∼40 Mpc. With this groundbreaking discovery, the GRB field formally entered the "multi-messenger era" (Abbott et al., 2017e).

1.2.2 Theoretical Progress

Due to the lack of critical observational clues, especially the distance information, the theoretical understanding of GRBs was initially very slow. Unlike other fields,[7] the nature of GRBs was not fully unveiled until the discovery of the afterglow, which occurred 30 years after the discovery of the first GRB. As a result, there were many theoretical papers (e.g. those listed in Ruderman 1975 and Nemiroff 1994, see §1.2.1) that turned out to be wrong, not because the physics used in their analyses was wrong, but because the premise of the models, i.e. the set-up of the problem, or the initial conditions were wrong.

It is impossible and unnecessary to review all those failed attempts. In this section, I list only important theoretical insights or models that I believe have shaped the current GRB theoretical framework. This list is of course subject to personal bias, even if I have tried to be objective as much as possible. I therefore apologize to those who believe that their important work is left out or down-graded. For convenience, the theoretical progress is again grouped based on the eras defined above according to the observational progress. Reviews on the theoretical progress in different eras can be also found in, e.g. Harding (1991), Piran (1999), Mészáros (2002), Zhang and Mészáros (2004), Piran (2004), Mészáros (2006), Zhang (2007), Kumar and Zhang (2015).

The Dark Era

Shortly after the discovery of the first GRBs, Stirling Colgate proposed a model of GRBs invoking shock breakout from Type II supernovae (Colgate, 1968, 1974). He interpreted γ-ray emission as bremsstrahlung and inverse Compton emission from a supernova shock as it breaks out of the star. The estimated total energy is typically 10^{48} erg, so they are observable up to 10–30 Mpc. According to the current standard paradigm, long GRBs are indeed associated with supernovae, but only with one special type: broad-line Type Ic, not Type II as envisaged by Colgate. Typical long GRBs are much more luminous (with a typical peak luminosity 10^{51}–10^{53} erg s^{-1}), which are believed to originate from a relativistic jet that has emerged from the collapsing star. In any case, a sub-category of long GRBs, known as low-luminosity GRBs, have long durations, smooth lightcurves, and low luminosities. They are consistent with having a shock breakout origin (but again from Type Ib/c rather than Type II supernovae), as envisaged by Colgate.

[7] For example, radio pulsars were soon identified as spinning neutron stars, and quasars were identified not long after as accreting super-massive black holes.

The first cosmological model of GRBs was probably the one proposed by Prilutskii and Usov (1975), who suggested (in Russian) that GRBs are generated by collapse of the cores of active galaxies. Like all the models proposed early on, this model is now ruled out by the data. Nonetheless, this possibility made Usov and Chibisov (1975) investigate different predicted behaviors of the GRB flux distribution ($\log N$–$\log S$) within Galactic and cosmological models, and suggested that a statistical test of GRB numbers would shed light on the origin of GRBs.

When reviewing the early GRB models, Ruderman (1975) already realized that the electron–positron pair production condition would limit the achievable GRB luminosity (the so-called "compactness problem"). He suggested that the condition does not "place significant burden" on Galactic GRB models, but it would be troublesome if GRBs were cosmological. He also pointed out that relativistic motion would enlarge the emission size and, hence, alleviate the problem.

Blandford and McKee (1976) studied the self-similar solution of the deceleration of an ultra-relativistic outflow. They did not target any astrophysical object at the time, but treated it as a pure physics problem as an extension of the non-relativistic Sedov–Taylor self-similar solution. It turns out that this theory formed the basis of modern GRB afterglow models.

Cavallo and Rees (1978) first discussed the *fireball* concept, with GRBs as an example. In particular, they applied the two-photon pair production condition to set general constraints on the luminosities of GRBs. This was a more elaborate manifestation of the "compactness" constraint.

In 1986, Bohdan Paczyński and Jeremy Goodman published two influential letters side-by-side in *The Astrophysical Journal* (Paczyński, 1986; Goodman, 1986). In these two papers, the two authors established the modern cosmological fireball model of GRBs. Paczyński (1986) noticed two rough coincidences and proposed that GRBs are cosmological. The two coincidences are: placing a typical observed GRB to a typical cosmological distance, the required energy ($\sim 10^{51}$ erg) is comparable to the typical supernova energy, and emitting this energy on a time scale of seconds from a region with a radius of 10 km (the size of a neutron star) as a blackbody, the typical temperature is around MeV. Even though he did not suggest a specific progenitor, Paczyński speculated several possibilities. In particular, he wrote:

> The binary radio pulsar PSR 1913+16 will coalesce with its neutron star companion within about 10^8 yr as a result of gravitational radiation losses (Taylor and Weisberg 1982). The final stage is likely to be very violent, and again of the order of 10^{52} or 10^{53} ergs will be released.

This was probably the earliest suggestion that *NS–NS mergers*, the leading progenitor model for short GRBs, may produce GRBs. In this paper, Paczyński also calculated the dynamical evolution of a photon-pair fireball, and suggested that the observed blackbody temperature at the fireball photosphere remains the same as the central engine temperature, and the spectral shape is close to a blackbody. Goodman (1986) also studied the evolution of a photon-pair fireball. He reached a similar conclusion as Paczyński (1986), and studied the emerging spectrum of a fireball in detail.

The fireball studied by Paczyński (1986) and Goodman (1986) is idealized, and does not carry any baryons. Later, Shemi and Piran (1990) added baryons to the fireball, and found that they significantly affect the dynamics of the fireball. In particular, a significant amount of thermal energy would be converted to the kinetic energy of the outflow.

Eichler et al. (1989) first studied the *NS–NS merger model* in great detail, and proposed that the mergers can be the progenitor of a subclass of observed GRBs. They also suggested that these systems are important multi-messenger emitting sources. Besides gravitational wave emission, they suggested that NS–NS mergers are also important sources of neutrino emission and probably the dominant sources of heavy elements through the rapid neutron capture process (the r-process) of neutron-rich material ejected from the merger. This paper laid the foundation of the standard paradigm of the modern short GRB models.

The BATSE Era

With the BATSE data showing isotropy and inhomogeneity of GRBs, the cosmological origin of GRBs became more attractive. In the early 1990s several seminal theoretical papers were published.

A baryonic fireball stores most of its energy in the kinetic form. In order to power non-thermal emission as observed in GRBs, energy dissipation is needed to re-convert kinetic energy to random particle energy and then to radiation. The most natural energy dissipation mechanism is through shocks. In a series of papers, Peter Mészáros and Martin Rees proposed the standard *fireball shock model*, which includes the main ingredients of the current GRB theoretical framework. Rees and Mészáros (1992) and Mészáros and Rees (1993b) first introduced the *external shock* of a relativistic fireball, and suggested that energy dissipation near the deceleration radius can be efficient enough to power non-thermal γ-rays to produce GRBs. *Synchrotron radiation* was invoked as the main radiation mechanism. The discussion was extended to the external *reverse shock* and *inverse Compton scattering* in Mészáros and Rees (1993a) and Mészáros et al. (1994). In 1994, Rees and Mészáros (1994) suggested that the irregularity of the central engine wind can drive internal shocks, which can lead to dissipation of kinetic energy within the flow (and hence "internal") and power GRBs via synchrotron radiation. Paczyński and Xu (1994) also discussed internal shocks with a focus on hadronic processes to power pion-induced γ-rays and neutrinos.

The dynamics of fireball evolution were also studied in detail, and consistent results were obtained from two different groups (Mészáros et al., 1993; Piran et al., 1993).

During this period of time, important progress was also made in the study of GRB *progenitor* and *central engine* models.

The NS–BH merger scenario was proposed as another possible progenitor of GRBs (Paczyński, 1991). The NS–NS merger scenario was studied in more detail (Narayan et al., 1992; Mészáros and Rees, 1992), with some basic physical processes (e.g. the BH central engine, possibility of collimated jets, jet launching mechanism, energy dissipation mechanism) sketched out.

Stan Woosley opened a new window by suggesting that not only can NS–NS or NS–BH mergers generate GRBs, but the collapse of a single Wolf–Rayet star (a type of massive star whose outer hydrogen envelope is stripped away by a stellar wind) with rapid rotation may

as well (Woosley, 1993). He even argued that this progenitor is a more appropriate interpretation of long-duration GRBs with a complex time profile. This model is now known as the standard "collapsar" model of long GRBs. In his original paper, Woosley acknowledged a difficulty of such a model: due to the large baryon contamination from the star, the outflow may not reach a highly relativistic speed, and hence may not generate a hard burst. In any case, he wrote in the abstract:

> Gamma-ray bursts or not, this sort of event should occur in nature and should have an observable counterpart.

He also named these events "failed" Type Ib supernovae, since Type Ib (no hydrogen line in the spectrum) supernovae are also believed to originate from Wolf–Rayet stars, and since a GRB progenitor may be more massive than that of SN Ib, so that inward accretion into a black hole would be more likely than an outgoing SN. Later observations showed that a GRB and a SN can co-exist. The associated SNe are of Type Ic (no hydrogen or helium in the spectrum), suggesting that the progenitor is even more stripped, i.e. besides the hydrogen envelope, the helium envelope is also lost by the time the explosion occurs.

While most modelers suggested that a *hyper-accreting black hole* (Narayan et al., 1992; Woosley, 1993) is the central engine powering a GRB, Usov (1992) suggested that a newborn, rapidly spinning, highly magnetized neutron star (*millisecond magnetar*) can also power a GRB by consuming its spin energy. The high luminosity can be sustained if the magnetic field is strong enough, and the neutron star spins down in a short period of time comparable to the burst duration. This magnetized central engine model was elaborated in Usov (1994) and Thompson (1994). Thompson (1994) also suggested that the GRB spectrum forms in a magnetically *dissipative photosphere*. He argued that a non-thermal Band-function (Band et al., 1993) spectrum would be produced from this model.

The external forward shock is long-lasting, since the fireball is expanding into an "infinite" circumburst medium. Besides generating prompt γ-ray emission (Rees and Mészáros, 1992; Mészáros et al., 1993), one naturally expects that there should be a long-lasting multi-wavelength *afterglow* in softer energy bands. Paczyński and Rhoads (1993) discussed a possible radio transient following a GRB. Katz (1994) discussed how the synchrotron peak frequency progressively passes through different energy bands at different times (even though the suggested time evolution behavior of the typical frequency and peak flux are different from the modern version). Sari and Piran (1995) studied the hydrodynamics of reverse shock propagation of a matter-dominated shell in great detail, and categorized the "thin" and "thick" shell regimes.

In 1997, two weeks before the discovery of the first X-ray and optical afterglow, Mészáros and Rees (1997a) published a seminal paper in which they systematically predicted the multi-wavelength afterglows of GRBs in a self-consistent manner. They discussed several possibilities including both the long-lasting forward shock and a short-lived reverse shock. Many predicted features of the models (power-law decaying behavior, the optical magnitudes in both forward and reverse shocks) were soon verified by observations.

During this period, the hadronic nature of GRBs and its possible implications were also discussed. In 1995, three independent papers (Waxman, 1995; Vietri, 1995; Milgrom and Usov, 1995) suggested that GRBs would be the sources of ultra-high-energy cosmic rays

(UHECRs) if they are cosmological events. In early 1997 (before the discovery of the first afterglow), Waxman and Bahcall (1997) suggested that GRB internal shocks are the site of neutrino emission in the PeV range.

The *BeppoSAX–HETE* Era

Soon after the discovery of the first afterglows, several groups independently showed that the data are generally consistent with the predictions of the fireball external shock model (Wijers et al., 1997; Vietri, 1997b,a; Tavani, 1997; Waxman, 1997b,c). In a four-page Letter to *ApJ*, Sari et al. (1998) most clearly presented the spectra and lightcurves of GRB afterglows for the simplest model (constant energy, constant medium density, and isotropic). This highly influential paper is user-friendly and serves as a standard reference for observers to quickly compare their data with the afterglow theory. More observations suggested that the simplest model cannot account for all the data. This stimulated further developments of the standard afterglow models. Mészáros et al. (1998) discussed several extensions of their earlier model (Mészáros and Rees, 1997a), including an inhomogeneous external medium and an angular structure of the outflow. Soon afterwards, these and other effects were investigated in great detail. These include: effects of stratification of the circumburst medium density in the form $n \propto R^{-k}$, especially for a stellar wind model with $k = 2$ (Mészáros et al., 1998; Dai and Lu, 1998b; Panaitescu et al., 1998; Chevalier and Li, 1999, 2000); effects of continuous energy injection into a fireball, either due to a long-lasting engine (Dai and Lu, 1998a,c; Zhang and Mészáros, 2001a) or a stratification of the ejecta Lorentz factor (Rees and Mészáros, 1998; Sari and Mészáros, 2000); effects of collimation of the ejecta (Rhoads, 1997; Panaitescu et al., 1998; Rhoads, 1999; Sari et al., 1999); and effects of the transition from the relativistic phase to the non-relativistic phase (Wijers et al., 1997; Huang et al., 1999, 2000; Livio and Waxman, 2000; Huang and Cheng, 2003). Intense afterglow modeling was carried out as growing multi-wavelength afterglow data flooded in, and model parameters were constrained from the data (Wijers and Galama, 1999; Panaitescu and Kumar, 2001, 2002; Yost et al., 2003).

Paczyński (1998) noticed that the first several afterglows (GRBs 970228, 970508, and 970828) are located close to the star-forming regions in their host galaxies, and suggested that the progenitors of these (long) GRBs are not due to compact star mergers, but are rather related to catastrophic deaths of massive stars. Similar to the "failed supernova" model of Woosley (1993), he proposed a "*hypernova*" model invoking a rapidly rotating star collapsing into a $\sim 10 M_\odot$ black hole surrounded by a thick accretion disk (or "torus"). The system magnetically launches a relativistic jet, which powers the observed GRB. MacFadyen and Woosley (1999) performed the first detailed numerical simulation of jet launching in a collapsing Wolf–Rayet star, and termed the phenomenology a "*collapsar*". More detailed simulations were carried out later by Woosley's group (MacFadyen et al., 2001; Zhang et al., 2003b, 2004b), which established the collapsar model as the standard theoretical framework of long GRBs. Many observed features were accounted for within this framework, including the associations with Type Ic SNe, collimation of the *jet*, and the existence of a less energetic "*cocoon*" surrounding the jet (see also Mészáros and Rees, 2001; Waxman and Mészáros, 2003).

As the sample of afterglows increased, one was able to attribute the optical temporal breaks of a few GRBs to jet collimation, known as *jet breaks*; and, by measuring *jet opening angle* of the bursts, one was able to measure the true energetics of the bursts. With a relatively small sample, Frail et al. (2001) surprisingly found that the total jet-corrected energy of a sample of GRBs is essentially constant.[8] In the view that the isotropic energy of GRBs varies in a wide range, this suggests that curiously different GRBs manage to collimate a *standard energy reservoir* into different jet angles (Frail et al., 2001). The suggestion was reinforced by Bloom et al. (2003) and Berger et al. (2003b). Shortly after this finding, Zhang and Mészáros (2002b) and Rossi et al. (2002) independently proposed an alternative, probably more elegant, interpretation: all GRBs probably have a *(quasi)-universal, structured jet*. Different GRBs may correspond to different viewing angles of this universal jet. The measured "jet angle" is not the true opening angle of a *uniform jet*, but is instead the viewing angle of the observer from the jet axis. This idea was further developed and extensively confronted against data before the launch of *Swift* (e.g. Perna et al., 2003; Kumar and Granot, 2003; Granot and Kumar, 2003; Lloyd-Ronning et al., 2004; Zhang et al., 2004a; Rossi et al., 2004; Nakar et al., 2004; Dai and Zhang, 2005).

While most theoretical studies in this era focused on afterglows, the investigations of the mechanism of GRB prompt emission continued. First, although the external shock model for GRB prompt emission was further developed (e.g. Dermer and Mitman, 1999), the requirements of producing both rapid variability and relatively high efficiency of prompt emission made the internal shock model more preferred (Kobayashi et al., 1997; Sari and Piran, 1997). In the meantime, it was realized that the radiative efficiency of the internal shock model is also not large, typically a few percent (Kumar, 1999; Panaitescu et al., 1999), unless some special settings of the central engine wind are envisaged (e.g. Beloborodov, 2000; Guetta et al., 2001; Kobayashi and Sari, 2001).

In view of the observations of a very hard low-energy photon index in some GRBs that exceeds the so-called $F_\nu \propto \nu^{1/3}$ "synchrotron line of death" (Preece et al., 1998), Mészáros and Rees (2000b) suggested that the fireball photosphere is an important emission site, whose emission can outshine the internal shock synchrotron component and dominate the observed GRB spectra in some GRBs. This triggered a wave of investigations into GRB photospheric emission (e.g. Mészáros et al., 2002; Kobayashi et al., 2002; Daigne and Mochkovitch, 2002). In 2005, Rees and Mészáros (2005) proposed that a *dissipative photosphere* may be the dominant emission component, and the photosphere temperature defines E_p of the GRB spectra – a revival of the earlier proposal of Thompson (1994). This suggestion soon became popular and was echoed by many authors, especially in the *Fermi* era (e.g. Pe'er et al., 2006; Thompson, 2006; Thompson et al., 2007; Giannios, 2008; Beloborodov, 2010; Ioka, 2010; Lazzati and Begelman, 2010; Toma et al., 2011; Lundman et al., 2013; Lazzati et al., 2013). See below in "The *Fermi* Era" for more discussion. In the meantime, synchrotron and synchrotron self-Compton (SSC) remained possible mechanisms to power GRBs, and Zhang and Mészáros (2002a) performed a systematic, comparative study of the predicted E_p properties of various models.

[8] Later investigations (e.g. Liang et al., 2008a; Racusin et al., 2009; Wang et al., 2015b) suggested that even though such a trend exists, the jet-corrected GRB energy still has a wide range of distribution.

Fitting the early optical lightcurves of several GRBs using the reverse/forward shock model led to the interesting finding that the reverse shock is usually more magnetized than the forward shock (Fan et al., 2002; Zhang et al., 2003a; Kumar and Panaitescu, 2003). The implication was that the GRB outflow may carry a magnetic field, suggesting that the GRB central engine is highly magnetized. In a long pre-print posted to arXiv:astro-ph, Lyutikov and Blandford (2003) proposed an electromagnetic model of GRBs. Instead of invoking a matter-dominated "fireball", they proposed that the GRB outflow is *Poynting flux dominated* from the central engine all the way to the deceleration radius. According to this model, GRB prompt emission is triggered when this electromagnetic bubble is decelerated by the ambient medium. This model invokes an extremely high value of the magnetization parameter, e.g. $\sigma \sim 10^6$ even at the deceleration radius, which is usually regarded as unrealistic. Nonetheless, it pushes the idea to another extreme direction. Around the same time, GRB models invoking an intermediate regime with moderate magnetization were also discussed (e.g. Spruit et al., 2001; Drenkhahn, 2002; Drenkhahn and Spruit, 2002). The reverse shock properties of an outflow with an arbitrary magnetization parameter were systematically studied (Zhang and Kobayashi, 2005; Mimica et al., 2009; Mizuno et al., 2009).

The *Swift* Era

Swift observations of the early afterglow phase of GRBs brought several surprises. Instead of decaying with a single power law from the beginning, as predicted by the theory, a large fraction of X-ray afterglows show a peculiar broken power-law decay lightcurve, which is known as the *canonical X-ray afterglow lightcurve* (Zhang et al., 2006; Nousek et al., 2006). Besides the *normal decay phase* and the *post-jet-break phase* well known in the pre-*Swift* era, the early afterglows show an early *steep decay phase* connected to the prompt emission (Tagliaferri et al., 2005; Barthelmy et al., 2005c), and a *shallow decay phase* (or plateau) before the normal decay phase kicks in (Campana et al., 2005; Vaughan et al., 2006). In nearly half of all GRBs, bright X-ray flares (Burrows et al., 2005a; Romano et al., 2006; Falcone et al., 2006) are detected. All these challenged the standard external shock afterglow model. Confronting data with theory, Zhang et al. (2006) suggested there are multiple physical processes operating during the early afterglow phase to shape the observed X-ray lightcurves: the steep decay phase is the tail of prompt emission, which is likely due to emission from the high latitudes with respect to the observer's line of sight when the prompt emission is over (Kumar and Panaitescu, 2000a); the shallow decay phase is likely the external shock emission with continuous energy injection, either from a long-lasting central engine, or from a Lorentz-factor-stratified ejecta; X-ray flares are internal emission due to late central engine activities, through a mechanism similar to that producing prompt γ-ray emission. Within such a picture, the so-called afterglow is a superposition of the traditional afterglow due to the ejecta–medium interaction and a long-lasting central-engine-driven afterglow.

While most GRBs can be understood within this theoretical framework, some others showed even more complicated afterglow behavior. In particular, some GRBs show the so-called *chromatic afterglow* behavior (e.g. Panaitescu et al., 2006a; Liang et al., 2007b),

with the optical lightcurve showing no break at the X-ray break time, or vice versa. More curiously, there is no associated spectral variation across the X-ray temporal breaks (Liang et al., 2007b, 2008a). This essentially ruled out the possibility of interpreting a multi-wavelength afterglow within the standard external shock framework at least in some GRBs. Many suggestions were made (e.g. Zhang 2007 for a review), but they were not fully successful. The origin of the early afterglow of some GRBs remains a puzzle. Some suggestions even attribute the entire X-ray afterglow to late central engine activities (e.g. Ghisellini et al., 2007; Kumar et al., 2008b).

Even though no consensus has been reached in interpreting the broad-band afterglow, various arguments suggest that X-ray flares must invoke delayed, intermittent central engine activities (Burrows et al., 2005a; Fan and Wei, 2005; Zhang et al., 2006; Liang et al., 2006b; Lazzati and Perna, 2007; Maxham and Zhang, 2009). How to restart the central engine becomes a pressing question. Within the black hole–torus central engine model, various suggestions were made, which include fragmentation of the collapsing star (King et al., 2005), fragmentation of the accretion disk due to gravitational instability (Perna et al., 2006), and modulation of the accretion flow by a dynamical magnetic barrier (Proga and Zhang, 2006). Alternatively, the magnetic activity of a rapidly spinning neutron star central engine may also account for X-ray flares (Dai et al., 2006; Metzger et al., 2008).

The existence of the X-ray plateau seems to favor a millisecond magnetar central engine (Zhang and Mészáros, 2001a). Later observations revealed a mysterious X-ray plateau followed by extremely rapid decay, in some (both long, Troja et al. 2007; Liang et al. 2007b, and short, Rowlinson et al. 2010, 2013) GRBs. These features, known as *internal plateaus*, are best understood as emission from a supra-massive millisecond magnetar, which survived the GRB itself, but later collapsed into a black hole. Extensive investigations of the magnetar central engine models (Bucciantini et al., 2007, 2009; Metzger et al., 2011) and their possible observational signatures have been carried out.

Swift data suggested that (at least some) short GRBs likely form a distinct population apart from long GRBs. This further reinforced the significant interest in studying NS–NS merger and NS–BH merger progenitor models. That these systems are also sources of gravitational waves adds additional motivation for these investigations. Many numerical simulations of NS–NS and NS–BH mergers have been carried out, with results focusing on different aspects of the problem, including jet launching (e.g. Rezzolla et al., 2011), ejecta mass distribution (e.g. Hotokezaka et al., 2013; Rosswog et al., 2013), evolution of magnetic field configuration (e.g. Siegel et al., 2014), and the properties of the final merger products (e.g. Giacomazzo and Perna, 2013).

The *Fermi* Era

Fermi opened the spectral window to a much wider bandpass. For the GRBs that are detected by both LAT and GBM, the spectral coverage is 6–7 orders of magnitude. This provides invaluable information about the GRB prompt emission physics.

Shortly after the discovery of >100 MeV emission (or simplified as "GeV" emission) in several GRBs by LAT (Abdo et al., 2009c,a; Ackermann et al., 2010), Kumar and Barniol Duran (2009, 2010) and Ghisellini et al. (2010) suggested that it comes from the external

shock. The main observational evidence is that the GeV emission lasts much longer than the MeV (GBM band) emission, and that it typically decays as a power law (Ghisellini et al., 2010; Zhang et al., 2011). Soon it was realized that this applies to GeV emission after the prompt emission phase, and that the GeV emission during the prompt emission phase is still of an internal origin (Maxham et al., 2011; He et al., 2011; Liu and Wang, 2011; Ackermann et al., 2013).

Another interesting observational fact is that, at least in some GRBs, GeV emission has a delayed onset with respect to the MeV emission. The origin of such a delay is still subject to debate. Several mechanisms have been proposed (e.g. Ghisellini et al., 2010; Razzaque et al., 2010; Asano and Mészáros, 2012; Mészáros and Rees, 2011; Bošnjak and Kumar, 2012; Beloborodov, 2013).

Great theoretical efforts have been made in the *Fermi* era for developing advanced models to interpret the prompt emission of GRBs. After the *Fermi* team published their first bright LAT burst GRB 080916C (Abdo et al., 2009c), Zhang and Pe'er (2009) pointed out that the data were not consistent with the prediction of the standard fireball photosphere–internal-shock model. They argued that the GRB central engine is strongly magnetized, so that most of the energy is initially carried in magnetic fields rather than in a hot outflow entrained with copious photons, and hence the bright photosphere component is suppressed (Daigne and Mochkovitch, 2002; Zhang and Mészáros, 2002a).

The apparent conflict with the standard model triggered a stream of theoretical investigations. Theorists' views in accounting for the same set of data could not be more diverse since the establishment of the cosmological origin of GRBs. Along the argument of Zhang and Pe'er (2009), Zhang and Yan (2011) proposed the *Internal-Collision-induced MAgnetic Reconnection and Turbulence (ICMART)* model of GRB prompt emission, which invokes a moderately *Poynting-flux-dominated outflow* in the emission region, so that turbulent magnetic reconnection in a $\sigma > 1$ flow plays the role of accelerating electrons and radiating γ-ray photons via synchrotron radiation. Further studies were carried out to account for other observational properties of GRB prompt emission (Zhang and Zhang, 2014; Uhm and Zhang, 2014b, 2016b; Deng et al., 2015, 2016). The connection between magnetic reconnection physics in the high-σ regime and GRB phenomenology is gaining growing attention (e.g. McKinney and Uzdensky, 2012; Kumar and Crumley, 2015; Beniamini and Granot, 2016; Guo et al., 2016; Lazarian et al., 2018).

In the meantime, some proposals were suggested to modify the fireball paradigm to accommodate the *Fermi* data. The first proposal (Daigne et al., 2011; Hascoët et al., 2013) admitted that the GRB central engine is highly magnetized, so that the photosphere emission component is suppressed. However, the magnetic energy is assumed to be quickly converted to the kinetic energy of the outflow, so that internal shocks are developed to power the observed γ-ray emission.

The second proposal to modify the fireball paradigm is to interpret the GRB Band-function spectrum as quasi-thermal emission from a dissipative photosphere (Beloborodov, 2010; Lazzati and Begelman, 2010; Ioka, 2010; Toma et al., 2011; Pe'er and Ryde, 2011; Pe'er, 2012; Mizuta et al., 2011; Lazzati et al., 2013; Lundman et al., 2013; Thompson and Gill, 2014). The internal shock component is assumed to be significantly suppressed, probably due to its low radiative efficiency.

Whereas a hot debate regarding the origin of GRB prompt emission is still going on as of the writing of this book, it is possible or even likely that the jet composition and energy dissipation mechanism of GRBs may differ from case to case. Different physical processes discussed in the literature may all play a certain role in producing GRBs.

Multi-Messenger Aspects

Over the years, the multi-messenger aspects of GRBs have been widely studied theoretically.

Back in 1995, three groups (Waxman, 1995; Vietri, 1995; Milgrom and Usov, 1995) independently suggested GRBs as a dominant source of ultra-high-energy cosmic rays (UHECRs). The foci of the three papers were different. Waxman (1995) proposed an internal shock origin for UHECRs, while Vietri (1995) proposed an external shock origin. Milgrom and Usov (1995), on the other hand, noticed two possible coincident events between UHECRs and GRBs, and suggested an association. The suggestions were revisited over the years (Waxman, 2004; Vietri et al., 2003), and it was argued that the cases were strengthened by further GRB observations. The ever stringent upper limits of the PeV neutrino flux from GRBs set by the *IceCube* collaboration imposed important constraints on the GRB–UHECR association models (Abbasi et al., 2012), even though the possibility of the association is not ruled out. It was suggested that low-luminosity GRBs (Murase et al., 2006) and engine-driven relativistic supernovae (Wang et al., 2007b; Chakraborti et al., 2011) could also be the sources of UHECRs.

Cosmic rays accelerated in GRBs interact with background photons or other baryons through hadronic ($p\gamma$) processes to produce high-energy neutrinos. Waxman and Bahcall (1997) suggested that PeV neutrinos can be produced from internal shocks through $p\gamma$ interactions at the Δ-resonance, with the observed sub-MeV γ-ray emission as the target photons. The predicted neutrino flux depends on several unknown parameters (Murase et al., 2008; He et al., 2012), and is also model dependent (Zhang and Kumar, 2013). The current flux upper limit set by *IceCube* has posed interesting constraints on GRB models. Neutrinos with different energies can be generated in a GRB from different emission sites. When applying the similar $p\gamma$ mechanism to the external shock, the typical neutrino energy shifts to the EeV (10^{18} eV) range (Waxman and Bahcall, 2000; Dai and Lu, 2001). As a GRB jet penetrates through the progenitor star, internal shocks may develop inside the star, from which protons may be accelerated and interact with X-ray photons to produce TeV (10^{12} eV) neutrinos (Mészáros and Waxman, 2001; Razzaque et al., 2003a). The process may be suppressed in radiation-mediated shocks, so it may be more relevant to low-power GRBs (Murase and Ioka, 2013). During fireball acceleration, inelastic collision between protons and neutrons may happen, which powers GeV (10^9 eV) neutrinos (Bahcall and Mészáros, 2000). Finally, $p\gamma$ interactions in low-luminosity GRBs (Murase et al., 2006; Gupta and Zhang, 2007a) and X-ray flares (Murase and Nagataki, 2006) also contribute to $\sim$EeV neutrinos.

Compact star mergers (NS–NS, NS–BH, BH–BH) have been well known as gravitational wave (GW) emitters (Taylor and Weisberg, 1989). As of the time of finishing this book, Advanced LIGO had already detected a few BH–BH merger events (Abbott et al.,

2016c,b, 2017c) and one NS–NS merger event. In particular, since NS–NS and NS–BH mergers are the top candidates for short GRB progenitors, a joint detection of a GRB and a GW source was expected (e.g. Kochanek and Piran, 1993; Bartos et al., 2013). The joint detection of GW170817 and the low-luminosity short GRB 170817A (Abbott et al., 2017b; Goldstein et al., 2017) in 2017 robustly confirmed such an expectation.

The detection of an electromagnetic counterpart of a GW source is of great interest (Kochanek and Piran, 1993; Finn et al., 1999). Besides short GRBs, some other electromagnetic counterparts of GW sources due to compact star mergers have been suggested in the literature. These include a radioactive "r-process" powered optical/infrared transient dubbed "macronova", "kilonova", or "mergernova" by various authors (Li and Paczyński, 1998; Kulkarni, 2005; Metzger et al., 2008; Yu et al., 2013; Metzger and Piro, 2014) (which was discovered to be associated with GW170817, e.g. Pian et al. 2017; Nicholl et al. 2017; Chornock et al. 2017), a faint radio afterglow (also called a radio flare) as this ejecta interacts with the ambient medium (Nakar and Piran, 2011; Piran et al., 2013; Gao et al., 2013b), and an X-ray counterpart due to magnetic dissipation if the NS–NS merger product is a millisecond magnetar (Zhang, 2013; Siegel and Ciolfi, 2016a; Sun et al., 2017) or a black hole (Kisaka and Ioka, 2015). With a lot of uncertainties, it is suspected that collapsars may also make strong GW burst emission (Kobayashi and Mészáros, 2003; Ott, 2009), making long GRBs and core-collapse hypernovae another possible multi-messenger target. The possible existence of a rapidly rotating, deformed magnetar at the central engine of these core-collapse events would also enhance the chance of detecting GWs associated with them (e.g. Corsi and Mészáros, 2009).

1.3 GRBs in Astrophysics

The GRB field is almost unique in astrophysics in its multi-disciplinary nature. Involving stellar-scale events located at cosmological distances, GRBs bridge several main branches of contemporary astronomy: stellar astronomy, interstellar medium (ISM) astronomy, galactic astronomy, and cosmology (Fig. 1.2).

Within the stellar context, the GRB physics is closely connected to many fundamental stellar astrophysics problems. In order to understand the progenitor of GRBs, one should understand the structure and evolution of massive stars, role of rotation, metallicity, and magnetic fields, as well as the final fates of stellar evolution: type(s) of supernova and the remaining remnant – a BH or NS. GRB progenitor(s) may invoke binary systems. This is likely relevant to most short-duration GRBs, and may be relevant to long GRBs as well. One therefore needs to study complicated stellar evolution channels invoking binaries (e.g. mass transfer between the members in the binary system, common envelope physics, as well as mergers of various binary systems: BH–He core, BH–WD, NS–NS, NS–BH), as well as the global distributions of these systems through stellar population synthesis. In order to understand how GRBs are generated, one needs to understand how a relativistic jet is launched from the central engine, either a hyper-accreting BH or a rapidly rotating, highly magnetized neutron star. This requires understanding the physics of BH accretion

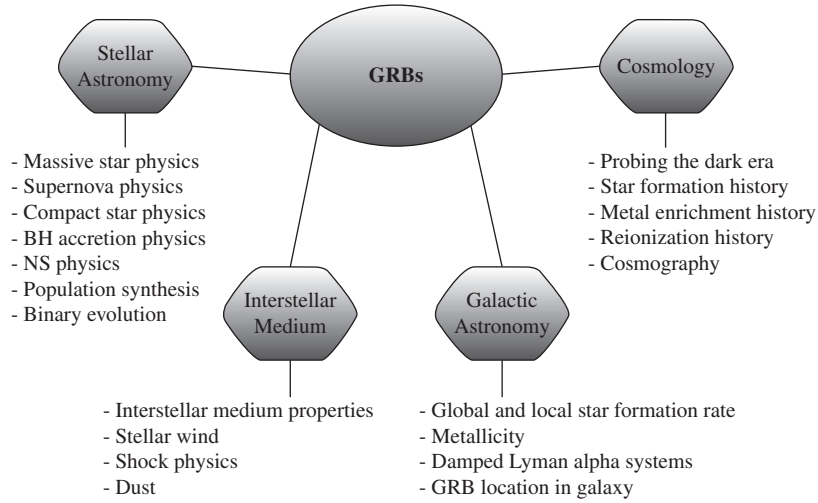

Figure 1.2 A flowchart showing connections between the GRB field and other major branches in astrophysics. Figure courtesy Jared Rice.

and evolution of nascent, rapidly spinning, strongly magnetized NSs. As a long GRB jet emerges from the progenitor star, one needs to study the interaction between the jet and the stellar envelope. As the jet is decelerated to produce afterglow emission, one also needs to study the interaction between the jet and a pre-explosion stellar wind or ejecta of the progenitor star. Since long GRBs are observed to be associated with supernovae, a close marriage between the GRB community and supernova (SN) community is also well established.

In the ISM context, GRBs define a unique case of jet–ISM interaction. A GRB afterglow is the relativistic version of a supernova remnant. By studying the temporal and spectral properties of the afterglow, one may learn detailed information regarding the density profile and clumpiness of the ISM. Absorption lines in the afterglow spectra can be used to diagnose chemical composition and abundances of the medium along the line of sight. Highly extinct afterglow and optically dark GRBs can probe the interstellar dust in the GRB environment.

On the galactic scale, GRB host galaxies define a unique cosmological galaxy sample. The morphology, star formation rate, and specific star formation rate, as well as the location of the GRB inside the host galaxy, carry rich information about galaxy evolution, properties of GRB progenitors, and their cosmological evolution. Spectroscopic observations can reveal metallicity (via metal lines) and local neutral hydrogen column density (via damped Lyman-α systems) of the host galaxy.

Finally, since long GRBs are found in a wide range of redshift (from $z = 0.0085$ to $z = 9.4$), they are ideal cosmological beacons to probe the evolution of the universe, in particular, the history of star formation, metal enrichment, and reionization during the "dark ages" shortly after the recombination epoch. Some GRB correlations can serve as a complementary tool to the traditional SN Type Ia "standard candles" to measure the cosmological parameters of the universe.

Historically, GRBs were discovered around the same time as quasars (or, broadly speaking, active galactic nuclei or AGNs) and pulsars, two other important discoveries in the 1960s. The nature of those two classes of objects was unveiled soon after their discoveries. AGNs are extragalactic sources believed to be powered by gigantic black holes, while pulsars are compact neutron stars located in our Galaxy. On the other hand, the lack of GRB observational breakthroughs hampered progress in the GRB field, and scientists from both the AGN and the pulsar communities brought the collective wisdom from each of their own fields to tackle the GRB problem. The GRB neutron star models were developed to fairly sophisticated levels (Harding, 1991), (mis-)motivated by the reported detections of absorption and emission features in some GRB spectra (e.g. Mazets et al., 1981b; Murakami et al., 1988). It turned out that the "classical" GRBs are of a cosmological origin, so that the wisdom borrowed from the AGN community (especially that for blazars, the most energetic type of AGNs) finally bore fruit. Nonetheless, one neutron star model eventually turned out to be partially correct. This model (Duncan and Thompson, 1992) invokes ultra-strong magnetic fields, or magnetars, as sources of the so-called "Soft Gamma-ray Repeaters" (SGRs). These SGRs, initially confused as a prototype of GRBs but later identified as a separate class of Galactic sources, are now confirmed to be slow-rotating magnetars in our Milky Way Galaxy (Thompson and Duncan, 1995). How strong magnetic fields are generated in these magnetars is still a mystery. One possibility, proposed by Thompson and Duncan (1993), is that magnetars are born as millisecond rotators. Strong convection during the rapidly, differentially rotating phase of these neutron stars would amplify magnetic fields via a dynamo mechanism. If this is the case, then millisecond magnetars should exist, which would be a possible central engine of GRBs (Usov, 1992). Recent observational data and theoretical modeling suggest that millisecond magnetars are indeed a viable engine of GRBs (Zhang and Mészáros, 2001a; Troja et al., 2007; Rowlinson et al., 2010; Metzger et al., 2011; Rowlinson et al., 2013; Lü and Zhang, 2014).

As the physical nature of GRBs is gradually unveiled, the wisdom gained in understanding GRBs is also applied to study other newly discovered phenomena, such as supernova shock breakout events (Campana et al., 2006; Soderberg et al., 2008), jets launched from tidal disruption of stars by super-massive black holes (Burrows et al., 2011; Bloom et al., 2011), and fast radio bursts (Lorimer et al., 2007; Thornton et al., 2013).

1.4 GRBs and Physics

The twentieth century saw two fundamental revolutions in physics – the development of relativity and quantum mechanics. GRBs are Nature's laboratory of extreme physics (Fig. 1.3). Macroscopically, GRBs are the fastest objects in the universe in bulk motion, with a measured Lorentz factor of a few hundreds and even up to $\sim$1000. Special relativity is pervasive in every aspect of GRB theory. At the central engine, a stellar-size black hole or a highly compact magnetar is at play, which distorts space-time in the vicinity, so that general relativity is demanded in correctly delineating the central engine physics. Microscopically, GRB phenomena invoke leptons and hadrons with extremely high energies, so

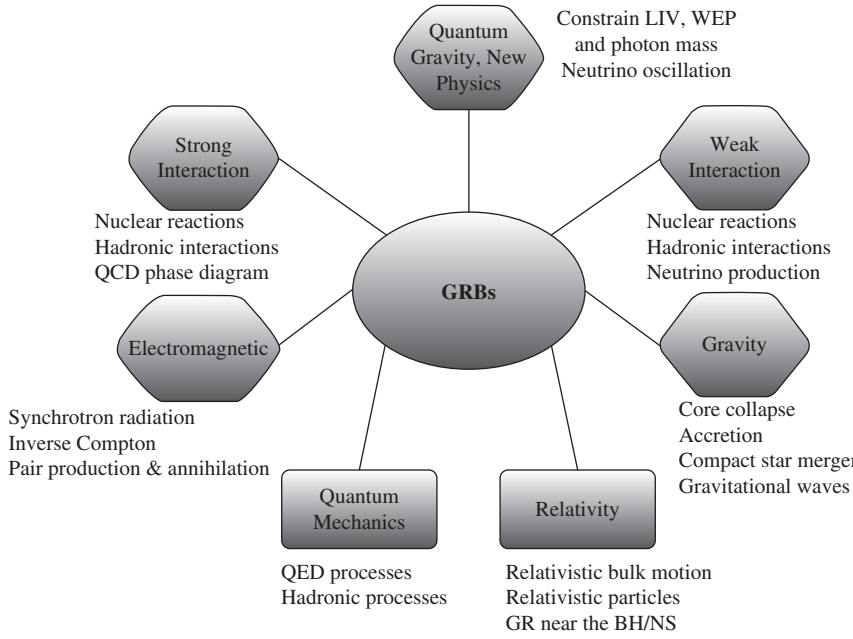

Figure 1.3 A flowchart showing connections between GRB astrophysics and different branches of physics. Figure courtesy Jared Rice.

that many high-energy processes invoking quantum electrodynamics (QED) and hadronic interactions are destined to take action.

One can easily find all four fundamental forces in operation in GRBs.

Gravity is at the heart of GRB physics, since GRBs involve catastrophic events controlled by gravity, either through core collapse in massive stars or coalescence of two compact stars. After this cataclysmic event, gravity again plays an essential role in powering a jet through accretion. In most progenitor and central engine models of GRBs, gravitational wave signals are predicted to escape from the source, which could be detected by an Earth observer along with the electromagnetic signals.

GRBs radiate electromagnetic radiation in the full band. Emission from GRBs has been detected from GHz to 100 GeV, which covers near 15 orders of magnitude. To understand GRB emission, one needs to understand generation and propagation of this broad-band emission, which invokes the physics of radiation mechanisms (e.g. synchrotron and inverse Compton scattering), pair production and annihilation, and the acceleration of charged particles in shocks or magnetic reconnection sites. In some GRB models, magnetic fields even play a dynamically dominant role, so that electromagnetic theory has to be applied to the GRB jet itself.

In the high-luminosity, high-energy, high-compactness environment of a GRB, strong and weak interactions are everywhere. In the extremely hot accretion disk at the central engine (a black hole or a proto-neutron star), many nuclear processes (such as nuclear photo-disintegration, neutralization, and β decay) continue to occur. Within the jet, inelastic collisions among protons and neutrons, and hadronic interactions between protons

and photons would generate pions, which subsequently decay to produce neutrinos, electrons/positrons, and photons. Nuclear synthesis may occur in the disk and even in the fireball. For NS–NS and NS–BH mergers, a small fraction of neutron-rich material is ejected from the system before the merger, which rapidly synthesizes heavy elements through a rapid neutron capture process (r-process). In the core of a newborn proto-neutron star, a QCD phase transition may even occur. All these processes would leave observational imprints in the photon and neutrino signals from GRBs.

Finally, GRBs can be used to place observational constraints on the physics models beyond the standard model. For example, the arrival time difference of photons of different energies from GRBs can be used to constrain particle physics models invoking Lorentz Invariance Violation (LIV), Einstein's Weak Equivalence Principle (WEP), as well as the photon rest mass. Flavor oscillations of GRB neutrinos, if detected, may bring clues to beyond-standard-model particle physics.

1.5 Broader Connections

GRBs are also discussed beyond the context of astrophysics and physics. For example, as the most violent explosions in the universe, GRBs are often discussed as one possible astrophysical source that may cause mass extinctions throughout Earth's history. Indeed, studies have shown that the intense γ-ray flux of a nearby GRB in the Milky Way Galaxy could destroy the ozone layer of the atmosphere, which would cause fatal DNA damage to life forms on Earth. Some even suspect that GRBs could have caused the extinction of the dinosaurs.

GRBs often appear in the headlines of various social media. This field is full of surprises, often beyond the imagination of the public and even us GRB researchers. For example, a GRB with a bright optical counterpart visible (in principle) to the naked eye was discovered on 19 March 2008 (Racusin et al., 2008). At a redshift $z = 0.937$ (about 7.5 billion light years away), this burst is the most distant object visible to the human eye. Imagine a massive star dying 7.5 billion years ago. It launches a very narrow jet, and somehow this jet is aimed squarely at Earth. After a long trip across about half of the observable universe, the photons released from the GRB would be seen by a human living on Earth 7.5 billion years later. How amazing!