



# Relativistic Envelopes and Gamma-Rays from Neutron Star Mergers

Andrei M. Beloborodov<sup>1,2</sup>, Christoffer Lundman<sup>3</sup>, and Yuri Levin<sup>1,4,5</sup><sup>1</sup> Physics Department and Columbia Astrophysics Laboratory, Columbia University, 538 West 120th Street, New York, NY 10027, USA<sup>2</sup> Max Planck Institute for Astrophysics, Karl-Schwarzschild-Str. 1, D-85741, Garching, Germany<sup>3</sup> The Oskar Klein Centre, Department of Astronomy, AlbaNova, Stockholm University, SE-106 91 Stockholm, Sweden<sup>4</sup> Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA<sup>5</sup> School of Physics and Astronomy, Monash University, Clayton VIC 3800, Australia

Received 2018 December 28; revised 2020 March 5; accepted 2020 March 25; published 2020 July 10

## Abstract

We suggest that neutron star mergers eject an ultrarelativistic envelope of mass  $m \sim 10^{-7} M_{\odot}$ , which helps explain the gamma-ray burst from GW170817. One ejection mechanism is the ablation of the neutron star surface by the burst of neutrinos in the first 30  $\mu$ s of the merger. Another, more efficient mechanism for inflating the ultrarelativistic envelope is an internal shock in the massive ejecta from the merger. A strong shock is expected if the merger product is a magnetar, which emits a centrifugally accelerated wind. The shock propagates outward through the ejecta and accelerates in its outer layers at radii  $r \sim 10^9$ – $10^{10}$  cm, launching an ultrarelativistic opaque envelope filled with  $\sim 10^4$  photons per nucleon. The Lorentz factor profile of the envelope rises outward and determines its homologous expansion, which adiabatically cools the trapped photons. Once the magnetar loses its differential rotation and collapses into a black hole, a powerful jet forms. It drives a blast wave into the envelope, chasing its outer layers, and eventually catching up with the envelope photosphere at  $r \sim 10^{12}$  cm. The ultrarelativistic photospheric breakout of the delayed blast wave emits a gamma-ray burst in a broad solid angle around the merger axis. This model explains the gamma-ray pulse from merger GW170817 with luminosity  $L_{\gamma} \sim 10^{47}$  erg s<sup>−1</sup>, duration  $\Delta t_{\text{obs}} \sim 0.5$  s, and characteristic photon energy  $\sim 100$  keV. The blast-wave Lorentz factor at the envelope photosphere is consistent with  $\Gamma \gtrsim 5$ , which we derive from the observed light curve of the burst. We suggest future tests of the model.

*Unified Astronomy Thesaurus concepts:* Gamma-ray bursts (630); Neutron stars (1108); Non-thermal radiation sources (1119); Relativistic fluid dynamics (1389); Gravitational wave sources (677); Shocks (2086)

## 1. Introduction

### 1.1. Ejecta from Neutron Star Merger GW170817

The recent detection of gravitational waves from neutron star merger GW170817 and its electromagnetic counterpart opens a new window for the study of neutron stars, cosmological gamma-ray bursts (GRBs), and the origin of heavy nuclei (Abbott et al. 2017a, 2017b; Coulter et al. 2017; Evans et al. 2017; Goldstein et al. 2017; Savchenko et al. 2017; Soares-Santos et al. 2017). The electromagnetic radiation was emitted by ejecta from the merger, viewed at an angle of  $\theta \sim 20^{\circ}$ – $30^{\circ}$  from the rotation axis of the binary. The viewing angle and the distance to the merger  $d \sim 40$  Mpc are both estimated from the observed gravitational wave signal, and its host galaxy was found at  $d \approx 40$  Mpc.

The gamma-ray counterpart, GRB 170817A, had luminosity  $L \sim 10^{47}$  erg s<sup>−1</sup> and was emitted with a delay of  $\sim 1.7$  s following the merger. It could be powered by a delayed jet from the central remnant when it breaks out from a massive cloud around the remnant (Kasliwal et al. 2017; Bromberg et al. 2018; Gottlieb et al. 2018; Pozanenko et al. 2018).

A day later, the cloud of expanding ejecta with mass  $M_{\text{ej}} \sim 5 \times 10^{-2} M_{\odot}$  and speed  $v/c \sim 0.1$ – $0.3$  emitted optical radiation with luminosity  $L \sim 10^{41}$  erg s<sup>−1</sup>—the “kilonova.” Its light curve was consistent with being powered by the decay of r-process nuclei (Drout et al. 2017; Kasen et al. 2017; Tanaka et al. 2017; Tanvir et al. 2017).

At yet later times (weeks), X-ray and radio afterglow was observed (Hallinan et al. 2017; Margutti et al. 2017; Troja et al. 2017). The unusually late rising afterglow was proposed to

result from deceleration of quasi-isotropic, moderately relativistic ejecta from the merger in an external medium (Nakar et al. 2018). It was also found to be consistent with a decelerating ultrarelativistic narrow jet with energy  $E_{\text{jet}} \sim 10^{50}$  erg (Granot et al. 2018; Lamb et al. 2018; Lazzati et al. 2018; Xie et al. 2018), and further evidence for a collimated jet came from radio imaging (Mooley et al. 2018a, 2018b). The jet is initially invisible to off-axis observers because of its strong collimation and Doppler-beaming with a very high Lorentz factor. Its emission comes into view after significant deceleration, long after the merger.

### 1.2. The Puzzling GRB 170817A

Generally, emission from GRB jets was expected to be weak and soft when viewed off-axis (e.g., Lazzati et al. 2017a, 2017b). By contrast, the gamma-ray pulse of GRB 170817A is not soft; its spectrum peaks at 100–200 keV (Goldstein et al. 2017), and an even harder spectrum was detected in a short time interval of  $\lesssim 0.1$  s (Veres et al. 2018). Furthermore, GRB jets display a strong correlation between the apparent brightness and hardness of their emission; GRB 170817A is far off this correlation (Pozanenko et al. 2018).

In addition, the simple light curve of GRB 170817A favors a single emission event, such as a blast wave from a jet, rather than variable internal dissipation typical for GRB jets. The canonical GRBs viewed on-axis are extremely bright and have diverse (usually multi-peaked) light curves (Nakar 2007). The off-axis GRB 170817A is dominated by a single weak pulse of width  $\Delta t_{\text{obs}} \sim 0.5$  s, smaller than but comparable to its delay

$t_{\text{obs}} \sim 1.7$  s. Goldstein et al. (2017) also reported an unusual transition from the gamma-ray pulse to a quasi-blackbody X-ray tail; although, the tail has a low signal-to-noise ratio, and its detailed spectral shape is uncertain.

A natural mechanism for a single, hard, gamma-ray pulse followed by a soft, thermal tail could be the breakout of a shock wave from the massive cloud around the merger. Following previous theoretical calculations (e.g., Nakar & Sari 2012), the shock breakout model proposed for GRB 170817A (Bromberg et al. 2018; Gottlieb et al. 2018; Kasliwal et al. 2017) posits the shock temperature  $kT \approx 50$  keV. It predicts the observed (Doppler-shifted) spectral peak at  $3kT \sim 150$  keV if the Lorentz factor of the shock-heated plasma is  $\Gamma \sim 1$ . The low  $\Gamma$  is, however, in conflict with observations. In Section 2, we show that the light curve of GRB 170817A requires the gamma-ray source to have  $\Gamma \gtrsim 5$ .

The Lorentz factor  $\Gamma \sim 5$  could be consistent with an off-axis component (“cocoon”) of an ultrarelativistic jet after its breakout from the massive cloud. However, this outflow is not expected to emit gamma-rays. Simulations of jet breakout and cocoon expansion show that the off-axis outflow is heated by internal shocks too early, before it becomes transparent to radiation, and this leads to reprocessed X-ray emission rather than gamma-ray emission (Lazzati et al. 2017a, 2017b).

### 1.3. This Paper

After estimating the lower limit on  $\Gamma$  in GRB 170817A in Section 2, we turn to the theory of relativistic ejecta from neutron star mergers. We find that, before jets are launched and the GRB is emitted, the merger is likely to eject an ultrarelativistic opaque envelope. The envelope quickly expands around the central massive cloud and greatly inflates the photospheric radius of the merger ejecta. In Sections 3 and 4, we describe two mechanisms for the envelope ejection and calculate its expected self-similar structure. In both cases, we find a stratified structure with four-velocity  $\gamma\beta$  growing outward and extending to  $\gamma\beta \gg 1$ . We estimate the expected mass of the ultrarelativistic envelope and its photon-to-baryon ratio.

The first mechanism of the envelope ejection is the ablation of neutron star surface at the very beginning of the merger, when it suddenly (in  $30 \mu\text{s}$ ) becomes a powerful source of neutrinos (Sections 3). The reaction  $\nu\bar{\nu} \rightarrow e^\pm$  injects heat near the surface of the merging stars, resulting in a huge energy per baryon and accelerating the surface layers to ultrarelativistic speeds. This effect is missed by the existing merger simulations (e.g., Dessart et al. 2009; Bauswein et al. 2013; Hotokezaka et al. 2013, 2018; Radice et al. 2016, 2018; Kiuchi et al. 2018), because they do not resolve the heating and dynamics of low-density surface layers. We find that an ultrarelativistic ablated mass  $m \gtrsim 10^{-8} M_\odot$  promptly escapes the vicinity of the merger, before the outflow becomes heavily polluted by baryons forming the massive cloud of slower ejecta.

The second mechanism is described in Section 4. The expanding cloud of massive ejecta can develop a strong internal shock, which propagates outward and accelerates to ultrarelativistic speeds in the outermost, low-density layers of the cloud. One appealing scenario invokes the formation of a rapidly spinning, short-lived magnetar following the merger. The fast outflow from the magnetar drives a shock into the cloud, which appears favorable for production of the “blue” kilonova (Metzger et al. 2018). We show that after the shock

crosses the cloud and accelerates in its outer layers, an ultrarelativistic envelope is inflated. We estimate its mass  $m \gtrsim 10^{-7} M_\odot$  and describe its self-similar structure.

The presence of the envelope weakly affects the bright beamed GRBs seen by on-axis observers; however, it can play a key role in off-axis GRB production. In Section 5, we discuss how a delayed launch of a jet inside the envelope leads to the production of an off-axis, single-pulse GRB, and we compare our model predictions with GRB 170817A. We find that the blast wave from the jet in the envelope can explain the observed luminosity, hardness, and light curve of the burst. Comparison with previous models and observational implications are discussed in Section 6.

## 2. Relativistic Motion in GRB 170817A

Kasliwal et al. (2017) found a lower limit for the Lorentz factor of the GRB source  $\Gamma > 2.5$ . They used the considerations of photon–photon opacity due to collisions  $\gamma\gamma \rightarrow e^+e^-$ , which occur for photons of energies above  $E_{\text{thr}} \sim \Gamma$  MeV (e.g., Lithwick & Sari 2001). It is, however, difficult to derive robust limits from photon–photon collisions. In principle, the source is allowed to be completely opaque to MeV photons, as no such photons were observed, and their number can only be guessed from extrapolations of the observed spectrum.

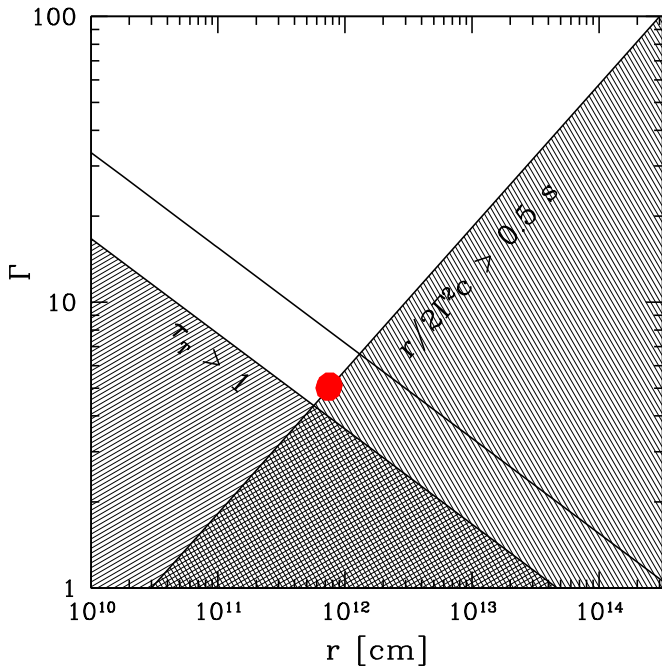
Instead, a simple lower bound on  $\Gamma$  can be obtained from the standard consideration of the scattering opacity in the source (e.g., Lithwick & Sari 2001). Let  $L \equiv 4\pi r^2 \Gamma^2 n_b m_p c^3$  be the isotropic equivalent of the kinetic power of the relativistic outflow emitting the observed gamma-rays with luminosity  $L_\gamma$  (note that  $L$  does not include  $L_\gamma$  and in general may be smaller or larger than  $L_\gamma$ ). Here  $r$  is the emission radius,  $n_b$  is the baryon number density in the flow rest frame, and  $m_p$  is the proton mass. A characteristic optical depth to Thomson scattering is given by

$$\tau_T = \frac{\zeta \sigma_T n_p r}{\Gamma} = \frac{\zeta \sigma_T L Y_e}{4\pi r m_p c^3 \Gamma^3}, \quad (1)$$

where  $n_p = Y_e n_b$  is the proton density,  $Y_e$  is the proton-to-nucleon ratio, and  $\zeta \lesssim 1$  is a numerical coefficient, which depends on the radial density profile of the outflow.

If  $L$  is approximately uniform on radial scales  $\Delta r > r/\Gamma^2$ , then one can show that  $\zeta \sim 1$ . This situation is expected if the outflow is launched on a timescale  $\Delta t > r/\Gamma^2 c$ . In particular, a ballistic outflow with  $\Gamma \approx \text{const}$  and  $L \approx \text{const}$  has a density profile  $\rho \propto r^{-2}$ . The characteristic optical depth for photons (emitted isotropically in the fluid frame) can be found by integrating the scattering coefficient  $\alpha_{\text{sc}} \propto \rho$  along the photon trajectory. This calculation gives Equation (1) with  $\zeta \sim 1$  (Abramowicz et al. 1991; Beloborodov 2011).

In the opposite limit, one can consider an outflow ejected impulsively, within  $\Delta t \approx 0$ . Then its density profile is set by radial spreading during the outflow expansion to the radius of GRB emission. This radial spreading is controlled by the distribution of four-velocity, which has a significant width for any realistic ejection mechanism. In particular, the homologous envelope described later in this paper can be idealized as an impulsive ballistic ejection with a power-law distribution of four-velocity (see Sections 5). A photon propagating in the homologous envelope will see a steeply decreasing density. In this case,  $\zeta$  can be as low as  $1/4$ .



**Figure 1.** Constraints on the radius  $r$  and Lorentz factor  $\Gamma$  of the plasma emitting GRB 170817A. The shaded regions are excluded by the observed peak duration  $\Delta t_{\text{obs}} \sim 0.5$  s and the optical depth requirement  $\tau_T \lesssim 1$ . The line of  $\tau_T = 1$  is shown for two cases:  $\zeta Y_e L / L_\gamma = 2$  (upper line) and  $1/4$  (lower line). The red circle shows the parameters of the burst model described in Section 5.

The GRB radiation can escape if  $\tau_T \lesssim 1$ , which gives a lower limit on  $\Gamma$ ,

$$\Gamma > \Gamma_* \approx 4.9 \, r_{12}^{-1/3} \left( \frac{\zeta Y_e L}{L_\gamma} \right)^{1/3} L_{\gamma, 47}^{1/3}. \quad (2)$$

Another constraint is set by the minimum dispersion of the photon arrival times  $\Delta t_{\text{disp}} \sim r / 2\Gamma^2 c$ . It applies as long as the GRB-emitting shell has a minimum angular size  $\Delta \theta_{\text{min}} \sim \Gamma^{-1}$ , which is valid for any expanding relativistic cloud accelerated by its internal pressure.<sup>6</sup> The minimum dispersion  $\Delta t_{\text{disp}}$  should not exceed the observed duration of the main peak of GRB 170817A  $\Delta t_{\text{obs}} \sim 0.5$  s, which gives the constraint

$$r \lesssim 2\Gamma^2 c \, \Delta t_{\text{obs}} = 3 \times 10^{10} \Gamma^2 \left( \frac{\Delta t_{\text{obs}}}{0.5 \text{ s}} \right) \text{ cm}. \quad (3)$$

Note that only the duration of the peak is relevant for this constraint; it is not affected by the arrival time of the peak  $t_{\text{obs}} \sim 1.7$  s.

Combining the two constraints in Equations (2) and (3), and using the observed  $L_\gamma \approx 1.6 \times 10^{47} \text{ erg s}^{-1}$  and  $\Delta t_{\text{obs}} \approx 0.5$  s (Goldstein et al. 2017), we find

$$\Gamma > 5.7 \left( \frac{L_\gamma}{1.6 \times 10^{47} \text{ erg s}^{-1}} \right)^{1/5} \left( \frac{\Delta t_{\text{obs}}}{0.5 \text{ s}} \right)^{-1/5} \left( \frac{\zeta Y_e L}{L_\gamma} \right)^{1/5}. \quad (4)$$

Figure 1 shows the constraints on the GRB source in the  $r$ - $\Gamma$  plane. If most of the observed luminosity comes from the photosphere of the explosion, the source must be located near

<sup>6</sup> The angular size of the emitting shell in GRB 170817A is likely to significantly exceed  $\Gamma^{-1}$  (below we argue that  $\Gamma \gtrsim 5$ ). The fact that we receive radiation at the polar angle  $\theta \sim 0.5$  rad, as inferred from the gravitational wave signal, suggests that  $\Delta \theta \gtrsim 0.5 > \Gamma^{-1}$ .

the line of  $\tau_T = 1$ . If it is produced by a photospheric shock breakout, then it should also be located near the line of  $r / 2\Gamma^2 c = 0.5$  s; and so, in this case, it will be near the intersection of the two lines.

A similar minimum  $\Gamma \sim 5$  for GRB 170817A may be obtained by considering the scattering opacity of  $e^\pm$  pairs created by MeV photons, using a high-energy extrapolation of the observed spectrum (Matsumoto et al. 2019).

One could also consider the possibility that the source is magnetically dominated, i.e., powered by Poynting flux from the central engine and carries a negligible amount of baryonic matter. Then the plasma emitting the GRB would be made entirely of  $e^\pm$  pairs created in photon-photon collisions. The source becomes transparent and releases radiation when its compactness parameter  $\ell = L_\gamma \sigma_T / m_e c^3 R \Gamma^3$  decreases to  $\sim 10$ , which requires  $\Gamma > 10 \, r_{11}^{-1/3}$ .

### 3. Ablation of the Neutron Star Surface

#### 3.1. Hot Sandwich at the Collision Interface

The two merging stars are strongly deformed by tidal forces, forming cusps pointing approximately toward each other (but still significantly misaligned, because of the orbital rotation of the binary). The merger begins with the tangential collision of the cusps, forming a growing interface between them (see, e.g., Figure 4 in Bauswein et al. 2013). The surface layers at the interface are shocked, compressed, and heated, forming a thin “sandwich.” Local thermodynamic equilibrium and nuclear statistical equilibrium are quickly established in the sandwich, with pressure contributions from nuclei, electrons, positrons, and Planckian radiation, at a high temperature  $T$ .

The opposite tangential velocities  $v_\parallel$  of the colliding stars imply a huge velocity shear at the interface, which immediately leads to Kelvin–Helmholtz instability (Price & Rosswog 2006; Kiuchi et al. 2018), with a growth rate comparable to the shear rate. The limited numerical resolution of the global merger simulations makes it difficult to observe the fast shear damping that develops on smallest scales, and local simulations (Zrake & MacFadyen 2013) show more details of the Kelvin–Helmholtz instability. The efficient damping of the tangential motion suggests that the collision is “sticky,” releasing its entire specific energy  $v_*^2/2 = (v_\parallel^2 + v_\perp^2)/2$ , not just the normal component  $v_\perp^2/2$ . Then the energy released in the sandwich per unit baryon rest mass is given by

$$\hat{E} \approx \frac{v_*^2}{2c^2}. \quad (5)$$

It may exceed 0.03 in mergers of massive neutron stars.

The hot compressed sandwich is bounded by two shocks propagating into the stars with speed  $v_{\text{sh}} \sim v_*$ . The energy density in the sandwich is  $U \approx \xi \rho c^2 \hat{E}$ , where  $\rho$  is the upstream (preshock) density,  $\xi = (\hat{\gamma} + 1)/(\hat{\gamma} - 1)$  is the shock compression factor, and  $\hat{\gamma}$  is the adiabatic index of the postshock matter. The sandwich pressure is

$$P = (\hat{\gamma} - 1)U \approx (\hat{\gamma} + 1)\rho c^2 \hat{E} \\ \approx 8 \times 10^{29} \rho_{10} \left( \frac{\hat{E}}{0.03} \right) \text{ erg cm}^{-3}. \quad (6)$$

As the two shocks propagate into denser subsurface layers of the colliding stars, the sandwich pressure grows,  $P \propto \rho$ .

The approximate pressure balance between the shocks implies  $P(x) \approx \text{const}$  across the sandwich, where the  $x$ -axis is chosen normal to the collision interface.<sup>7</sup> However, the nucleon density is strongly nonuniform in the  $x$ -direction—the sandwich is made of layers of stratified density. The older layers in the middle were shocked at a low pressure and later pressurized through compression. As  $P$  grows proportionally to the upstream density  $\rho$ , the old shocked layers are strongly compressed to stay in approximate pressure equilibrium with the propagating shocks. This compression implies strong adiabatic heating, which produces a low-mass, thin layer of ultrarelativistic material. The compressional heating is discussed in more detail in the [Appendix](#).

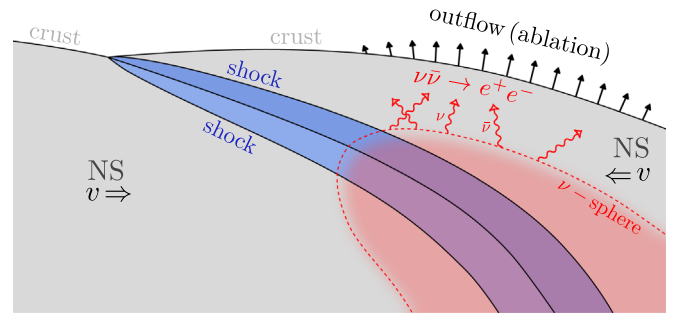
In principle, the compressed layer heated to specific enthalpy  $W > c^2$  could be partially ejected with an ultrarelativistic speed. The basic effect may be illustrated by a vessel of hot gas compressed to a small volume by external pressure. The work performed to compress the vessel is stored in the gas internal energy, and a sufficiently strong compression makes the gas relativistically hot,  $W \gg c^2$ . When the external pressure is eventually removed, the gas will explode and achieve ultrarelativistic speeds. This toy model does not, however, capture additional effects expected at the interface of the colliding stars. In particular, mixing and transport effects should suppress the relativistic ejection (see the discussion in the [Appendix](#)).

The magnetospheres of the colliding neutron stars will also be strongly compressed in collision. In an idealized model, this would create a magnetic “pillow” at the interface between the stars, with magnetic pressure in an approximate balance with the ram pressure of the two shocks propagating into the stars  $B^2/8\pi \sim P$ . As the ram pressure grows up to  $P \sim 10^{34} \rho_{14}$  erg cm<sup>−3</sup> (see Equation 6), the magnetic field in the pillow is amplified up to  $B \sim 5 \times 10^{17} \rho_{14}^{1/2}$  G, where  $\rho$  is the matter density upstream of the shocks. This implies compression of the magnetosphere by a huge factor  $B/B_0$  for reasonable pre-merger magnetic fields  $B_0$ . The resulting pillow thickness  $\delta \sim (B_0/B)R$  is many orders of magnitude smaller than the stellar radius  $R \sim 10$  km.

However, the idealized picture of a compressed magnetic pillow is destroyed by instabilities of the Raleigh–Taylor type, which will tend to mix the compressed magnetic field into dense stellar material. Note also that the magnetic field is amplified in a much thicker layer as a result of the Kelvin–Helmholtz instability.

### 3.2. Neutrino Emission from the Sandwich

When the two stars have just touched, the sandwich pressure and temperature are initially modest; at this earliest stage, most of the collision energy converts to Planckian radiation. When the shocks propagate into deep and dense layers,  $\rho \gtrsim 10^{12}$  g cm<sup>−3</sup>, the pressure becomes dominated by nucleons rather than radiation. The simplest estimate for the sandwich



**Figure 2.** Schematic picture of the shocked sandwich at the collision interface (blue) and its neutrino emission (red). The neutrinos decouple from the stellar matter (shaded in gray) and escape with a quasi-thermal spectrum at the neutrino-sphere (dotted red curve). Some of the emitted  $\nu$  and  $\bar{\nu}$  collide outside the neutrino-sphere and convert to  $e^\pm$  pairs. This process deposits energy into the cold surface layers and leads to their ablation.

temperature is given by the upper limit,

$$kT_{\text{max}} \equiv \frac{2}{3} \hat{E} m_n c^2 \approx 20 \left( \frac{\hat{E}}{0.03} \right) \text{ MeV} \quad (\text{no cooling}), \quad (7)$$

which neglects any contributions to pressure other than nucleons.

The rate of neutrino and antineutrino production by the sandwich is quickly increasing with temperature  $T$ , and becomes significant before  $T$  approaches  $T_{\text{max}}$ . The cooling becomes significant when the shocks reach crustal layers with densities  $\rho \gg 10^{10}$  g cm<sup>−3</sup>; then the cooling timescale becomes shorter than the shock age (the details will be described elsewhere). Neutrinos are mainly produced by the  $e^\pm$  capture reactions  $e^+ + n \rightarrow p + \bar{\nu}$  and  $e^- + p \rightarrow n + \nu$ . If the neutrinos escape, their energy flux can be estimated as

$$F \sim \frac{\rho v_\star^3}{2} \quad (\text{efficient cooling}). \quad (8)$$

This simply states that most of the energy released in the shock converts to the neutrino flux. The effective temperature  $T$  of escaping neutrinos is approximately related to their energy flux by  $F \approx \sigma T^4$ , where  $\sigma = ac/4$  is the Stefan–Boltzmann constant.<sup>8</sup> This gives an estimate

$$kT \approx 15 \rho_{12}^{1/4} v_{\star,10}^{3/4} \text{ MeV} \quad (\text{efficient cooling}). \quad (9)$$

It may be viewed as a lower bound on  $T$  at large densities. The upstream  $\rho$  appearing in Equations (8) and (9) is lower near the edges of the sandwich, where the shocks formed later and had less time to propagate into deep subsurface layers (Figure 2). The sandwich size measured along the collision interface grows from the initial contact point to  $\sim 10$  km on a timescale  $t \sim 3 \times 10^{-5}$  s.

As the two shocks bounding the sandwich propagate into the subsurface layers of increasing density  $\rho$ , the postshock temperature grows and so do the energies of emitted neutrinos  $\epsilon_\nu$  (in units of  $m_e c^2$ ). The cross section for neutrino interaction with matter grows as  $\epsilon_\nu^2$ , and the neutrinos eventually become absorbed near (or inside) the sandwich. This occurs when the

<sup>7</sup> Variation of  $P$  in the  $x$ -direction is small because the shocks are in hydrodynamical causal contact and not far from pressure equilibrium. At the same time,  $P(y, z)$  strongly varies along the collision interface, decreasing from the center (the initial touch point of the colliding stars) to the outer parts of the sandwich, where shocks form later (Figure 2). Initially, the collision interface area grows superluminally, faster than the shocked matter could be squeezed out from the sandwich. At a later stage, the pressure gradient in the  $y$ - $z$  plane begins to drive a fan-like “fountain” from the sandwich.

<sup>8</sup> The emitted neutrino flux is not exactly  $\sigma T^4$  for two reasons: (i) their spectrum is not exactly thermal, and (ii) even for completely thermalized neutrinos, their (fermion) statistics is different from photon statistics. The numerical factor resulting from these corrections is  $\sim 1$  and weakly affects the estimate in Equation (9).

shocks propagate sufficiently deep below the stellar surface,  $\rho \gg 10^{11} \text{ g cm}^{-3}$ . Then neutrino transport occurs in a diffusive regime, with the local neutrino density  $U_\nu$  close to local thermodynamic equilibrium. The mean energy of the thermalized neutrinos (in units of  $m_e c^2$ ) is

$$\epsilon_\nu \approx 3 \Theta \equiv 3 \frac{kT}{m_e c^2} \quad (\text{thermalized}). \quad (10)$$

Cross sections for neutrino interactions with nucleons and leptons are summarized, e.g., in Chen & Beloborodov (2007). For our estimates, it is sufficient to include one process—neutrino absorption by nucleons, which has the largest opacity  $\kappa_\nu \approx 3 \times 10^{-20} \Theta^2 \text{ cm}^2 \text{ g}^{-1}$ . The mean free path of neutrinos is

$$l_\nu = \frac{1}{\kappa_\nu \rho} \approx 3.6 \times 10^5 \rho_{11}^{-1} \left( \frac{\Theta}{30} \right)^{-2} \text{ cm}. \quad (11)$$

One can see that at densities  $\rho \gg 10^{11} \text{ g cm}^{-3}$  the neutrinos are self-absorbed and thermalized.

At the advanced stage of collision, most of the produced neutrinos are trapped in the middle of the sandwich, and the escaping neutrinos diffuse out from its outer parts (Figure 2), where density  $\rho \lesssim 10^{12} \text{ g cm}^{-3}$ . The simple estimate in Equation (8) then suggests a characteristic value for the escaping flux  $F \sim 10^{41} \text{ erg s}^{-1} \text{ cm}^{-2}$ . This estimate is, however, crude. An accurate spatial distribution of  $F$  could be found with expensive three-dimensional simulations of the collision dynamics and neutrino transport. The result will depend on the masses of the colliding stars. Massive mergers produce high  $F$  because they have a high collision speed  $v_*$ .

The flux  $F \sim 10^{41} \text{ erg s}^{-1} \text{ cm}^{-2}$  is emitted from the “neutrino-sphere”—the surface where the neutrinos decouple from matter and begin to stream freely. The temperature of the neutrino-sphere  $T_*$  satisfies the approximate relation  $F \sim \sigma T_*^4$ , which gives  $\Theta_* \sim 30$ .

### 3.3. Heating of Surface Layers by Reaction $\nu\bar{\nu} \rightarrow e^+e^-$

Some of the neutrinos and antineutrinos escaping from the neutrino-sphere can collide with each other and convert to  $e^\pm$  pairs, depositing heat. The rate of this “neutrino heating” is independent of the local matter density, and in the low-density regions near the stellar surface it injects a huge energy per nucleon, rising the local specific enthalpy

$$w \equiv \frac{U + P}{\rho c^2} \quad (12)$$

to relativistic values  $w \gtrsim 1$ . The heated surface layers will expand with relativistic speeds and leave the star. Thus, a fraction of the stellar crust will be ablated by neutrino heating.

Neutrinos collide and turn into  $e^\pm$  pairs with a significant cross section when there is a significant angle between their directions,  $\delta \gtrsim 1$ . The neutrino heating wave propagates with a speed  $v \sim c \cos \delta$ , comparable to  $c/2$ . The wave is faster than the shocks, so ablation of surface layers occurs before the shock arrival (Figure 2).

We wish to obtain a rough estimate for the mass ablated to highly relativistic speeds. The first step is to evaluate the heating rate  $\dot{Q}$  due to the reaction  $\nu\bar{\nu} \rightarrow e^+e^-$ . The cross section

for this reaction is given by (Goodman et al. 1987)

$$\sigma_{\nu\bar{\nu}} \approx 0.2 \sigma_0 \frac{c^4 (\mathbf{p}_\nu \cdot \mathbf{p}_{\bar{\nu}})^2}{(E_\nu E_{\bar{\nu}})^2} = 0.2 \sigma_0 \epsilon_\nu \epsilon_{\bar{\nu}} (1 - \cos \delta)^2, \quad (13)$$

where  $\sigma_0 \approx 1.7 \times 10^{-44} \text{ cm}^2$ ,  $\epsilon = E/m_e c^2$  (with subscripts  $\nu$  and  $\bar{\nu}$  corresponding to the colliding neutrino and antineutrino, respectively),  $\mathbf{p} = (E/c, p_x, p_y, p_z)$  is the four-momentum of the colliding neutrino/antineutrino, and  $\delta$  is the angle between their directions.

Each reaction  $\nu\bar{\nu} \rightarrow e^+e^-$  creates an  $e^\pm$  pair with total energy  $E_\nu + E_{\bar{\nu}}$ . The corresponding energy deposition rate per unit volume  $\dot{Q}$  may be estimated by replacing the quasi-thermal spectrum of the neutrino-sphere by a delta-function at the average neutrino energy,  $\epsilon_\nu \approx \epsilon_{\bar{\nu}} \approx 3\Theta_*$ . Then one finds

$$\dot{Q} \sim \sigma_{\nu\bar{\nu}} c n_\nu n_{\bar{\nu}} (E_\nu + E_{\bar{\nu}}) \sim \frac{\sigma_{\nu\bar{\nu}}}{c} \frac{F_\nu}{E_\nu} \frac{F_{\bar{\nu}}}{E_{\bar{\nu}}} (E_\nu + E_{\bar{\nu}}). \quad (14)$$

We will also use the estimate  $F_\nu \approx F_{\bar{\nu}} = F \approx \sigma T_*^4$ . This gives

$$\dot{Q} \sim \sigma_0 \overline{(1 - \cos \delta)^2} \frac{\Theta_*^4 F^2}{m_e c^3} \sim 10^{35} F_{41}^{9/4} \text{ erg s}^{-1} \text{ cm}^{-3}, \quad (15)$$

where  $\overline{(1 - \cos \delta)^2} \lesssim 1$  is a numerical factor obtained after averaging over the directions of the colliding neutrinos.

It may be instructive to compare the estimate of Equation (15) with previous numerical simulations of *steady* heating by the reaction  $\nu\bar{\nu} \rightarrow e^+e^-$  by Birkel et al. (2007) and Zalamea & Beloborodov (2011), who focused on accretion disks around accreting black holes, in Kerr metric. Birkel et al. (2007) also calculated steady heating in spherical geometry, with and without gravitational bending of neutrino trajectories. These simulations gave the efficiency of converting neutrino flux to heat,  $F_h/F$ , where

$$F_h = \int \dot{Q} ds \sim H \dot{Q}, \quad (16)$$

$s > 0$  is the altitude above the neutrino source,  $F_h$  is the vertically integrated heating rate, and  $H$  is the characteristic scale-height of the heating region. In the simplest spherical model, the thermal neutrino source is described by its surface flux  $F$  and radius  $R$ . The characteristic  $H$  is comparable to  $0.1R$ . We have checked that the heating rate calculated by Birkel et al. (2007), scaled to  $R \sim 10 \text{ km}$  and  $F \sim 10^{41} \text{ erg s}^{-1} \text{ cm}^{-2}$ , is approximately consistent with the estimate in Equation (15). It gives a rough estimate of the heating efficiency  $F_h/F \sim H \dot{Q}/F$  approaching 0.1.

In a real merger, the neutrino source geometry is neither spherical nor axisymmetric, and the neutrino-sphere is not parallel to the stellar surface (Figure 2). Furthermore, an essential difference from the previous work is that here we deal with an initial-value problem rather than a steady state. The ablation of surface layers is triggered by the suddenly arising burst of neutrinos from the sandwich.

It takes a very short time  $t \sim x/v_{\text{sh}} \sim 10^{-5} \text{ s}$  for the sandwich density and temperature to reach high values so that its neutrino flux approaches  $F \sim 10^{41} \text{ erg s}^{-1} \text{ cm}^{-2}$ . As the sandwich size grows, the area of the neutrino-sphere grows on a similar timescale. Thus, the local  $\dot{Q}(t)$  measured at the stellar surface is a steeply increasing function of time, shaped by the evolution of the neutrino-sphere and the emitted flux  $F$ . This function is also slightly affected by a propagation delay: the

wave of heat injection propagates with a speed of  $v_h \sim c \cos \delta \lesssim c$  from the neutrino-sphere to the stellar surface.

### 3.4. Estimate for Relativistically Ablated Mass

Only the uppermost ablated layers, which have low densities, reach highly relativistic speeds. The relativistic ablation ends after a short time  $t_a$  when a sufficient amount of matter is lifted from the NS surface and fills the main heating region  $H \sim 1$  km. Later, the neutrino-driven outflow becomes a relatively slow quasi-steady wind, which was studied previously in detail (Qian & Woosley 1996; Thompson et al. 2001, 2004; Metzger et al. 2007; Dessart et al. 2009). The quasi-steady wind at  $t \gg t_a$  is mainly heated through neutrino absorption by baryons. By contrast, heating at the initial ablation stage  $t < t_a$  is dominated by the process  $\nu\bar{\nu} \rightarrow e^+e^-$ . There is much less matter in the initial relativistic outflow, and neutrinos are mainly absorbed in  $\nu\bar{\nu}$  collisions rather than by baryons.

Ablation may be roughly described as a two-step process: (1) enthalpy  $w$  is deposited in the heating zone  $s \lesssim H \sim 1$  km, where matter begins its acceleration, and (2) adiabatic expansion at  $s > H$  converts enthalpy to bulk kinetic energy. For instance, layers with asymptotic  $\gamma\beta \sim 1$  are still relatively slow in the heating zone  $s \lesssim H$ . Their modest characteristic speed  $v_a \sim (0.1-0.2)c$  gives them time  $t_a \sim H/v_a \gtrsim 10^{-5}$  s to accumulate enthalpy  $w \sim 1$  before leaving the heating zone.

One can estimate the mass of ultrafast ablated layers from the condition  $w \gtrsim 1$  in the heating zone. The deposited energy density

$$U \sim \dot{Q}t_a \sim 10^{30} F_{41}^{9/4} t_{-5} \text{ erg cm}^{-3} \quad (17)$$

will give dimensionless enthalpy  $w > 1$  in layers of density  $\rho < (U + P)/c^2 \sim 10^9 \text{ g cm}^{-3}$ . These layers occupy an initial volume  $V \sim A h \gtrsim 10^{16} \text{ cm}^3$ , where  $A \sim 10^{11}-10^{12} \text{ cm}^2$  is the area of the neutrino-sphere near the sandwich edge, where ablation occurs, and  $h \sim 0.3-1$  km is the characteristic hydrostatic scale-height of the stratified crust in the colliding tidal cusps of the stars. This gives the relativistically ablated mass  $\sim V \dot{Q}t_a/c^2$ .

The relativistic ablation is not isotropic and may peak in a solid angle  $\Omega_a \sim 1$ , which corresponds to a beaming factor of  $b \sim 4\pi/\Omega_a$ . The isotropic equivalent of the ablated mass  $m$  viewed within  $\Omega_a$  is given by

$$m \sim b V \frac{\dot{Q}t_a}{c^2} \sim 10^{-7} M_\odot b_1 V_{16} F_{41}^{9/4}. \quad (18)$$

We conclude that the observed isotropic equivalent of the ablated mass with  $\gamma\beta > 1$  may exceed  $10^{26}$  g, depending on the viewing angle and the precise values of the beaming factor  $b$  and the neutrino-sphere flux  $F$ .

### 3.5. Numerical Simulation

The features of relativistic ablation discussed above are illustrated by the following simplified numerical model. Let us replace the merging stars by a single sphere of radius  $R$  and choose the surface heating rate  $\dot{Q}$  in the form

$$\dot{Q}(t, s) = \dot{Q}_0(t) \left( \frac{a+s}{a} \right)^{-4} \left[ 1 - \frac{s}{(s^2 + a^2)^{1/2}} \right]^2. \quad (19)$$

Here,  $s = r - R_\star > 0$  is the distance from the neutrino-sphere  $R_\star$ , which is somewhat below the stellar surface,  $R_\star < R$ . Our simulation assumes a neutrino-sphere radius  $R_\star = 10$  km and a star radius  $R = 11$  km.

A true spherical source of neutrinos would have  $a = R_\star$  in Equation (19). However, our spherical model is designed as a proxy for the colliding stars, and we use  $a < R_\star$  to parameterize the relatively small size of the ablation region near the sandwich edge (Figure 2). The dependence of  $\dot{Q}$  on  $s$  has two parts: the power law  $(1 + s/a)^{-4}$  describes the reduction of  $n_\nu n_{\bar{\nu}}$  with distance from the neutrino-sphere, and the term in square brackets roughly describes the dependence  $\sigma_{\nu\bar{\nu}} \propto (1 - \cos \delta)^2$ , where  $\tan \delta \sim a/s$  represents the characteristic angle between the colliding neutrinos. Equation (19) implies a steep decline of  $\dot{Q}$  with  $s$ . This leads to a modest characteristic thickness of the main heating region, comparable to or smaller than 1 km.

The  $\dot{Q}_0(t)$  in Equation (19) describes the time dependence of surface heating by the neutrino wave from the sandwich. At early times,  $\dot{Q}_0(t)$  is a steep function, because it is proportional to  $F^{9/4}$ , and  $F$  is quickly increasing as the sandwich pressure grows,  $P \propto \rho$  (Equation 6). We do not know the exact shape of this function and replace it with a simple power law,

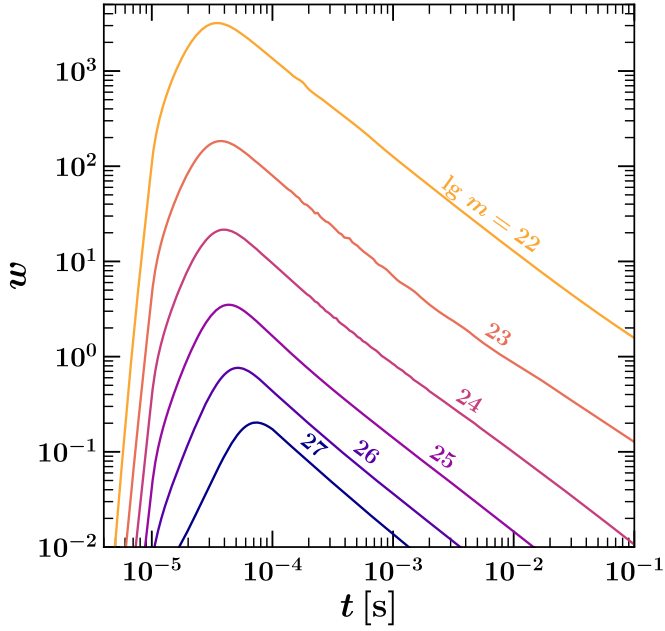
$$\dot{Q}_0(t) = 10^{35} \left( \frac{t}{t_0} \right)^\zeta \text{ erg s}^{-1} \text{ cm}^{-3}, \quad t < t_0, \quad (20)$$

followed by constant  $\dot{Q}_0(t) = \dot{Q}_0(t_0)$  at  $t > t_0$ . Our sample model will have  $\zeta = 9$  and  $t_0 = 10^{-5}$  s. This crude description of the heating onset captures its main feature: the steep rise on a short timescale, which will lead to relativistic ablation of the surface layers.

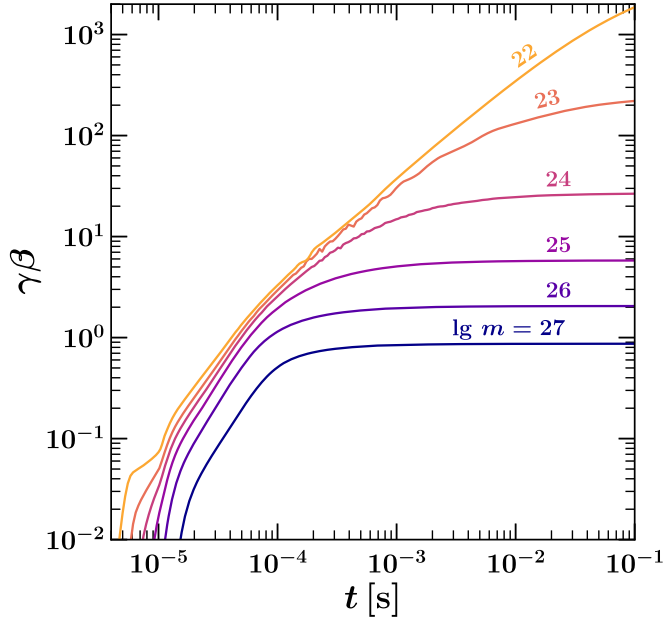
The simulation tracks the dynamics of the outer crust with total mass  $M_{\text{sim}} = 10^{28}$  g. The crust initially occupies a spherical shell with the outer radius  $R = 11$  km. It is initially static and stratified in hydrostatic equilibrium with a power-law index  $q$ :  $\rho \propto x^q$  where  $x = R - r$  is the depth below the stellar surface at a radius  $r < R$ . We have run simulations with  $q = 5$  and  $q = 3$ . Our simulations are performed using a relativistic Lagrangian hydrodynamic code described in Lundman et al. (2018), with some modifications. In particular, we use a nonuniform discretization in the mass coordinate  $0 < m < M_{\text{sim}}$ , which allows us to resolve well the dynamics of the low-density layers near the stellar surface. The simulated mass  $M_{\text{sim}}$  is discretized into  $10^4$  subshells.

The results are shown in Figures 3–5. One can see that the heat is deposited in the surface layers on the timescale  $t \sim (3-7) \times 10^{-5}$  s. Then, the layers leave the main heating zone, expand, and cool, converting the accumulated heat to bulk kinetic energy. The relativistically ablated matter approaches its final (asymptotic) momentum  $\gamma\beta \gg 1$  after a longer time, when it has lost its enthalpy through adiabatic cooling.

Figure 5 shows how much mass escapes with  $\beta\gamma$  larger than a given value. The result is sensitive to the geometric parameter  $a$ . For instance, in the most “optimistic” case of  $a = 4$  km, mass  $m \approx 10^{26}$  g is ejected with  $\gamma\beta > 2$ , and  $m \approx 10^{25}$  g is ejected with  $\gamma\beta > 4$ . One can also see that the distribution  $\gamma\beta(m)$  extends to very high values of  $\gamma\beta \sim 10^3$ . This is expected, as at the onset of ablation, the neutrino wave deposits comparable energy everywhere in the heating zone in the upper crust (or even above it), regardless of the local density of



**Figure 3.** Heating history of ablated surface layers. The dimensionless enthalpy  $w = 4P/\rho c^2$  is shown as a function of time  $t$  for several layers with different Lagrangian mass coordinates  $m$ ;  $m$  is measured inward from the preablation stellar surface. This sample ablation model assumes the initial hydrostatic stratification index  $q = 5$  and the geometric parameter of the neutrino-sphere  $a = 4$  km.

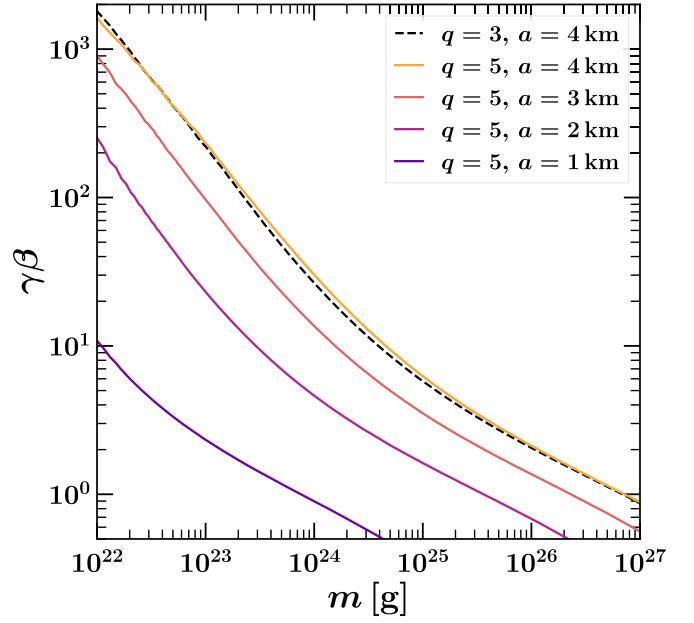


**Figure 4.** Evolution of the dimensionless momentum  $\beta\gamma$  of the same layers as in Figure 3.

matter. As a result, the outermost layers of the neutrino-driven outflow form an ultrarelativistic fireball that freely expands with acceleration by the radiation pressure  $\gamma \propto r$ . The fireball Lorentz factor is limited to  $\gamma \sim 10^3$ , because at higher  $\gamma$  the fireball becomes transparent to radiation, and acceleration becomes inefficient.

#### 4. Early Internal Shock in the Merger Ejecta

This section describes the second mechanism for generating a self-similar ultrarelativistic envelope. It invokes a shock wave



**Figure 5.** Asymptotic four-velocity  $\beta\gamma$  of the relativistically ablated layers as a function of the Lagrangian coordinate mass  $m$  measured inward (i.e.,  $m = 0$  at the outer boundary of the crustal material). Our ablation simulation included only upper layers  $m < 10^{28}$  g, which are heated mainly via the reaction  $\nu\bar{\nu} \rightarrow e^+e^-$  (later ejecta emerge from deeper and denser layers heated mainly via neutrino absorption by nucleons). The results are shown for several models that have heating rates  $\dot{Q}$  (see the text) with different choices of the parameter  $a$  that depends on the shape of the sandwich neutrino-sphere. The results are sensitive to  $a$  and weakly depend on the initial stratification of the surface layers  $\rho \propto x^q$  in the relevant range of  $q$ .

accelerating in the outer layers of the massive cloud around the merger remnant.

This mechanism is similar to shock breakout in supernovae, see, e.g., detailed calculations in Tan et al. (2001) for spherical shocks, which are parallel to the surface of the supernova progenitor. These calculations were applied to merging neutron stars by Kyutoku et al. (2014). We note that the merger shocks propagate obliquely to the neutron star surface. When the oblique shock approaches the upper crust, the sudden density drop will make the shock perpendicular to the surface, rather than parallel, and this effect introduces an upper cutoff on the velocity of the ejected material (Matzner et al. 2013). Therefore, it is unclear if a sufficient amount of ultrarelativistic ejecta can be produced by the merger shocks launched inside the neutron stars. Instead, below we focus on internal shocks in the large cloud around the merger remnant. Shock acceleration in the expanding cloud is more capable of ejecting a relativistic envelope.

Clouds around the merger remnants were predicted to have masses  $M_{\text{ej}} \sim 10^{-2} M_{\odot}$  and expansion speeds  $v_{\text{ej}}/c \sim 0.1\text{--}0.3$  (e.g., Bauswein et al. 2013; Siegel & Metzger 2017). The cloud inferred from the kilonova in GW170817 is somewhat more massive,  $M_{\text{ej}} \sim 0.05 M_{\odot}$ . An internal shock must develop in the cloud if matter ejected at later times has a higher speed. In particular, this is expected if the merger produces a magnetar—a massive, differentially rotating neutron star that generates ultrastrong magnetic fields. Although the field dynamics in the merger is not fully understood, it is plausible that an ultrastrong field is generated and then rises to the stellar surface due to magnetic buoyancy (Kluźniak & Ruderman 1998). Then, a strong magnetosphere should form, and the neutrino-driven

wind from the massive neutron star becomes centrifugally accelerated (Mestel & Spruit 1987; Metzger et al. 2018). This fast wind will drive a shock wave into the earlier, slower ejecta. Metzger et al. (2018) argued that such a wind could help explain the “blue” part of the kilonova emission in GW170817.

A shock with a velocity jump  $\delta v$  crossing the cloud of mass  $M_{\text{ej}}$  will dissipate energy

$$\mathcal{E} \sim \frac{M_{\text{ej}}(\delta v)^2}{2} \approx 10^{51} \left( \frac{M_{\text{ej}}}{10^{-2} M_{\odot}} \right) \left( \frac{\delta v}{10^{10} \text{ cm s}^{-1}} \right)^2 \text{ erg.} \quad (21)$$

As long as the shock propagates deep inside the cloud, its speed remains mildly relativistic. When it reaches the outer layers of the cloud, where density is lower, the shock will accelerate and bring the outer layers to ultrarelativistic speeds. This process of shock breakout will inflate a self-similar relativistic envelope around the merger remnant.

Below, we estimate the mass of the relativistic envelope and its structure, and we illustrate it with a hydrodynamical simulation. Then, we estimate the photon-to-baryon ratio in the inflated envelope.

#### 4.1. Previous Results for Shock Breakout in Static Clouds

Shock acceleration in nonrelativistic hydrodynamics was studied in detail six decades ago (see Zel’dovich & Raizer 1967 and references therein). Its relativistic version was proposed as a possible mechanism for outflows in GRBs (Paczynski 1998).

Tan et al. (2001) provided analytical fits for relativistic mass ejection by shock breakout, which were tested against hydrodynamical simulations. In their simulations, a shock emerges from an initially static star with density  $\rho = 0$  at the stellar surface and a polytropic mass stratification with depth  $x$  below the surface,  $\rho \propto x^n$ , with a typical  $n = 3$ . The form of their analytical approximation is motivated by earlier results (obtained in the relativistic and nonrelativistic limits), which are valid for more general density profiles. Therefore, similar ejecta are expected for shock breakout in clouds with different density distributions, as long as  $\rho$  steeply drops in the outer layers of the cloud.

The main dimensionless parameter of the problem is the ratio  $\mathcal{E}/M_{\text{ej}}c^2$ . For strong shocks in the merger ejecta, we expect

$$\tilde{\mathcal{E}} \equiv \frac{\mathcal{E}}{M_{\text{ej}}c^2} \sim 0.01 - 0.1, \quad (22)$$

which corresponds to  $\delta v \sim (0.1-0.5)c$  during the shock propagation inside the cloud, before the breakout.

The breakout problem has two parts: shock dynamics and subsequent expansion of the shock-heated fluid, with adiabatic cooling and bulk acceleration. The growth of the shock speed  $\beta_s = v_s/c$  with decreasing density  $\rho \ll M_{\text{ej}}/r^3$  is approximately described by

$$\gamma_s \beta_s \approx A \tilde{\mathcal{E}}^{1/2} \left( \frac{M_{\text{ej}}}{\rho r^3} \right)^{\alpha}, \quad (23)$$

where  $\gamma_s = (1 - \beta_s^2)^{-1/2}$ , and  $A \sim 1$  is a numerical factor. The power-law index  $\alpha \approx 0.2$ ; its more accurate value is 0.187 when  $\gamma_s \beta_s \ll 1$  (Zel’dovich & Raizer 1967) and  $\alpha = \sqrt{3} - 3/2 \approx 0.23$  when  $\gamma_s \beta_s \gg 1$  (Johnson & McKee 1971; Pan & Sari 2006). As  $\beta_s$  grows with decreasing  $\rho$ , the dissipated energy per unit mass increases; however, the

energy density decreases. Most of the energy  $\mathcal{E}$  is dissipated in the dense, heavy part of the ejecta, and only a fraction of  $\mathcal{E}$  is delivered to the outer, low-density layers that eventually develop highly relativistic motion  $\gamma\beta > 1$ .

Figure 6 in Tan et al. (2001) shows how the ejecta energy is distributed over the asymptotic  $\gamma\beta$  for several choices of  $\tilde{\mathcal{E}}$ . In particular, for  $\tilde{\mathcal{E}} \sim 0.03$ , they find that the ejecta with asymptotic  $\gamma\beta > 1$  carry the energy of  $\mathcal{E}_1 \approx 6 \times 10^{-5} M_{\text{ej}}c^2$ , and these ejecta have mass  $m_1 \sim 10^{-4} M_{\text{ej}}$ . When a similar estimate is applied to the merger cloud, it gives  $m_1 \sim 10^{-6} M_{\odot} (M_{\text{ej}}/0.01 M_{\odot})$ . Ejecta with yet higher  $\gamma\beta \gg 1$  have a significantly smaller mass, e.g.,  $m \sim 10^{-6} M_{\text{ej}}$  for  $\gamma\beta > 3$ .

These estimates are sensitive to  $\tilde{\mathcal{E}}$ . One can see from Figure 6 in Tan et al. (2001) that a change of  $\tilde{\mathcal{E}}$  from 0.03 by a factor of three (in either direction) changes  $m_1$  by approximately two orders of magnitude. A crude estimate in the relevant parameter range may be written as

$$m_1 \equiv m(\beta\gamma > 1) \sim 10^{-6} M_{\odot} \left( \frac{\tilde{\mathcal{E}}}{0.03} \right)^4 \left( \frac{M_{\text{ej}}}{0.01 M_{\odot}} \right), \quad (24)$$

$$\frac{m(\beta\gamma > 3)}{m(\beta\gamma > 1)} \sim 3 \times 10^{-3}. \quad (25)$$

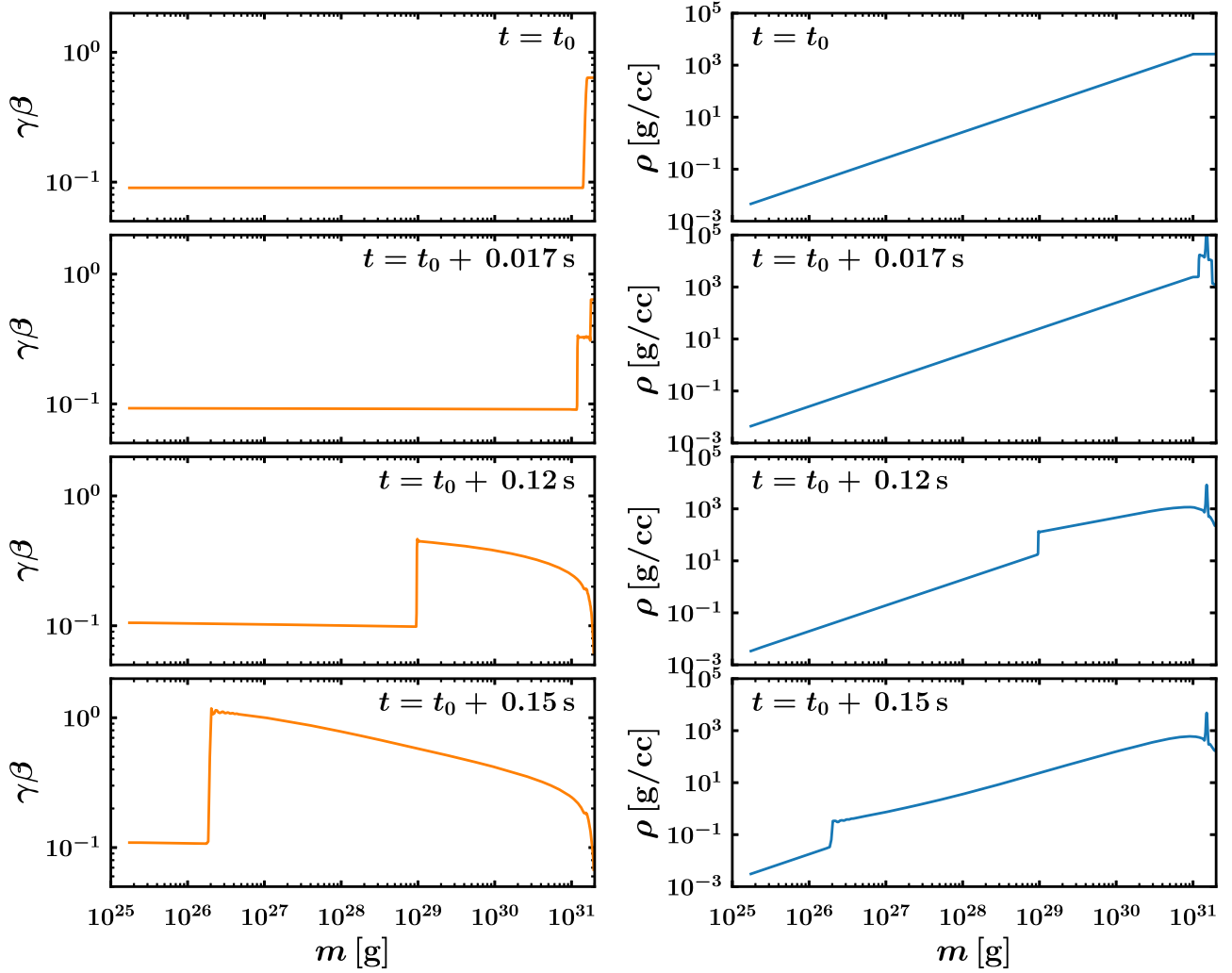
While Tan et al. (2001) considered a static star with a certain density profile, similar order-of-magnitude estimates apply to shocks in expanding clouds. In Section 4.2, we perform a detailed calculation for a sample cloud model and find the accurate distribution of the asymptotic four-velocity in the ejected envelope.

#### 4.2. Simulation of Early Shock Breakout from an Expanding Cloud

At the start of the simulation (time  $t = t_0$ ), we specify the cloud parameters as follows. We place a spherical shell of mass  $M_{\text{ej}} = 10^{-2} M_{\odot}$  with an outer radius  $R_c = 2 \times 10^9$  cm and an inner radius of  $10^9$  cm. The outer half of the shell is expanding with a uniform speed  $v_0 = 0.1c$ , and the inner half is expanding with  $v_0 + \delta v$ , where  $\delta v/c = (2\tilde{\mathcal{E}})^{1/2} \approx 0.45$  corresponds to  $\tilde{\mathcal{E}} = 0.1$ .

We assume that the cloud was adiabatically cooled as it expanded from the merger remnant of radius  $\sim 10^6$  cm, and we give the shell a low (insignificant) enthalpy  $w = (U + P)/\rho c^2 = 10^{-3}$ . The density profile of the shell is flat except near its outer edge, where the density falls off exponentially on a scale  $\Delta r = 0.3R_c$ . We choose the moderate  $\Delta r/R_c$  keeping in mind that ejecta from neutron star mergers are initially hot, and there is significant pressure in the cloud until it strongly expands and cools adiabatically. Even if the cloud was initially ejected with a sharp edge, the pressure drop in the outermost layers will accelerate them, creating a positive velocity gradient in the radial direction. It leads to stretching of  $\Delta r$  and makes the density decline at the edge smooth and gradual. Therefore, freely expanding warm clouds in general cannot have sharp edges. They are also generally expected to have a positive gradient of  $v_0(r)$ , so our assumption of  $v_0(r) = \text{const}$  in the outer layers is a rather crude simplification of the expansion velocity profile.

Our initial condition with the velocity jump  $\delta v$  inside the cloud of radius  $\sim 10^9$  cm roughly corresponds to shock



**Figure 6.** Evolution of the fluid density profile  $\rho(m)$  (right panels) and the four-velocity profile  $\gamma\beta(m)$  (left panels). The Lagrangian coordinate  $m$  is measured inward;  $m = 0$  at the outer boundary of the simulated cloud and  $m = 2 \times 10^{31}$  g at the inner boundary.

launching from the central object at time  $\sim 1/2$  s after the merger. In the simulation, the jump immediately launches a forward shock in the middle of the cloud. There is also a reverse shock; however, we are mainly interested in the dynamics of the forward shock, which propagates outward, reaches the low-density layers, and accelerates. Similar to the ablation simulation in Section 3, we use the Lagrangian mass coordinate  $m$  counted from outside and employ the Lagrangian relativistic hydrodynamics code of Lundman et al. (2018). The nonuniform discretization in  $m$  allows us to track the evolution of the entire massive shell  $M_{\text{ej}} = 0.01 M_{\odot}$  while resolving the dynamics of the outer, low- $m$  layers, where the forward shock reaches highly relativistic speeds.

Figure 6 shows snapshots of fluid density  $\rho(m)$  and four-velocity  $\gamma\beta(m)$  in our simulation. It demonstrates the shock evolution and its effect on the structure of the cloud. As the forward shock enters the outer low-density layers and accelerates, it loses causal contact with the inner massive part of the cloud. The shock completely crosses the outer half of the cloud in  $\sim 1/4$  s, and after this, the shocked layers continue to expand with acceleration, converting heat to bulk kinetic energy. Then, the ejecta become cold and ballistic.

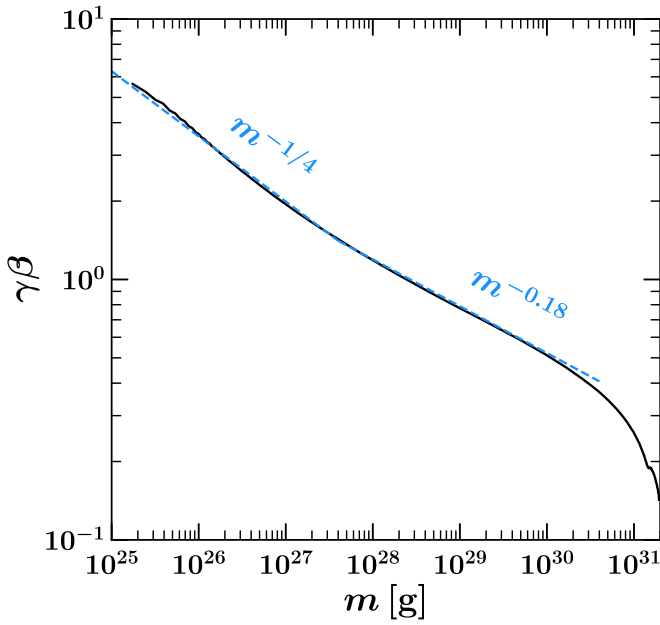
The final profile of  $\gamma\beta$  as a function of the Lagrangian mass coordinate  $m$  is shown in Figure 7. It may be approximately

described by power laws with different slopes  $\psi$  in the regions  $\gamma\beta \lesssim 1$  and  $\gamma\beta \gg 1$ ,

$$\gamma\beta \approx \left(\frac{m}{m_1}\right)^{-\psi}. \quad (26)$$

In our sample simulation with  $\tilde{E} = 0.1$ , we find  $m_1 \approx 10^{-5} M_{\odot}$ . The power-law slope is  $\psi \approx 0.18$  at  $m \gtrsim m_1$  and  $\psi \approx 1/4$  at  $m \ll m_1$ .

In this sample simulation, we assumed a flat preshock velocity profile  $\beta_0 = 0.1$ . More realistic expanding clouds have an increasing profile  $v_0(r)$ , shaped during the cloud formation near the central object. Changing the shape of  $v_0(r)$  slightly changes the results, as long as  $v_0 \sim 0.1$ . We also run models with a faster preshock speed  $v_0 = 0.1\text{--}0.3$ ; then, the profile of  $v_0(r)$  significantly affects the final distribution of  $\gamma\beta$  after shock breakout. Furthermore, the detailed shape of this distribution is affected by the initial density profile in the outer layers and the initial enthalpy in the cloud. However, in all runs, we found the final  $\gamma\beta(m)$  qualitatively similar to that shown in Figure 7: a shallow power law at  $\gamma\beta \lesssim 1$ , and a steeper power law at  $\gamma\beta \gg 1$ , with  $\psi$  close to  $1/4$ .



**Figure 7.** The final distribution of the four-velocity  $\gamma\beta$  in the envelope launched by the internal shock breaking out from the cloud of massive ejecta. The Lagrangian coordinate  $m$  is measured inward ( $m = 0$  at the outer boundary of the envelope). The cloud mass in the simulation was  $M_{\text{ej}} = 0.01 M_{\odot}$ , and its preshock expansion speed was  $v_0 = 0.1c$ . The shock was created by depositing an additional energy of  $10^{51}$  erg in the inner half of the cloud (see the text and Figure 6). The black curve shows the result of the numerical simulation, and the blue dashed line shows a fit by the broken power law, with the break at  $\gamma\beta = 4/3$ .

#### 4.3. Photon-to-baryon Ratio in the Envelope

The early shock breakout described above occurs at a modest distance  $R_{\text{sh}}$  from the merger. This distance depends on the timescale of launching the shock,  $t_{\text{sh}}$ , which may be related to the magnetar formation in the center; we assume  $t_{\text{sh}} < 1$  s. During this time, the ejecta expands to a radius  $r \sim v_0 t_{\text{sh}} < 10^{10}$  cm, and the shock breaks out at a similar radius. It emits a burst of radiation when it reaches the ejecta photosphere; however, this burst is too weak to be observed, as discussed below. Instead, the main radiative effect of the early internal shock is an increase of the photon number trapped in the ejecta. The large photon number will play a role later, when a new explosion from the merger remnant energizes the envelope at large radii  $r \gtrsim 10^{12}$  cm, and a detectable GRB is emitted (Section 5).

First, let us estimate the photon number carried by the massive cloud without an internal shock. The ejection of mass  $M_{\text{ej}}$  in time  $t_{\text{ej}}$  implies a characteristic mass outflow rate  $\dot{M} \sim M_{\text{ej}}/t_{\text{ej}} = 2 \times 10^{31} (M_{\text{ej}}/0.01 M_{\odot}) (t_{\text{ej}}/1 \text{ s})^{-1}$ . The ejection radius  $R_0 \sim 20$  km (the size of the merger remnant) and expansion speed  $\beta = v/c \sim 0.1\text{--}0.3$  determine a characteristic density  $\rho_0 \sim \dot{M}/4\pi R_0^2 v \sim 10^8 \dot{M}_{31}/\beta_{-1} \text{ g cm}^{-3}$ . Note that the expansion speed requires an initial enthalpy per unit rest mass  $w_0 \sim \beta^2/2$  and, hence, a thermal energy density at the base of the outflow  $U_0 = (3/4)w_0\rho_0 c^2 \sim 10^{27} \dot{M}_{27}\beta_{-1} \text{ erg cm}^{-3}$ . This rough, order-of-magnitude estimate is sufficient to evaluate a characteristic temperature from  $aT_0^4 \sim U_0$ :  $kT_0 \sim 1 \dot{M}_{31}^{1/4} \beta_{-1}^{1/4} \text{ MeV}$ . A more accurate estimate takes into account that at temperatures  $\sim 1$  MeV,  $e^{\pm}$  pairs make a contribution to  $U_0$  comparable to that of photons; then,  $T_0$  is reduced by the factor of  $(11/4)^{-1/4} \approx 0.8$ .

The photon number density is  $n_{\gamma} = U_0/2.7kT_0$ , and the baryon number density is  $n_b = \rho_0/m_p$ . The photon-to-baryon ratio in the massive cloud is then given by

$$\frac{n_{\gamma}}{n_b} \approx \frac{(3/4)w_0 m_p c^2}{2.7kT_0} \sim 10 v_{10}^{7/4} \dot{M}_{31}^{-1/4} \quad (\text{without shock}). \quad (27)$$

The internal energy and entropy of the cloud are dominated by radiation, and during its adiabatic expansion, the photon number remains approximately constant, because it is proportional to entropy. The conservation of the photon and baryon numbers implies that their ratio remains unchanged as the cloud expands.

The simple model of adiabatic expansion can be refined by including two effects. First, photon number is increased by a moderate factor of  $\sim 2$  when the temperature decreases below  $m_e c^2 = 511$  keV and the  $e^{\pm}$  pairs annihilate into photons. Second, when the cloud temperature drops to  $kT_{\text{rec}} \sim 150\text{--}200$  keV, free nucleons recombine into  $\alpha$ -particles, releasing 28 MeV per  $\alpha$ -particle (see Beloborodov 2003 for a discussion of  $T_{\text{rec}}$  in GRBs). The number of photons generated by the recombination may be estimated as

$$\frac{n_{\gamma}}{n_b} \sim 2Y_e \frac{7 \text{ MeV}}{2.7kT_{\text{rec}}} \sim 10, \quad (28)$$

where  $Y_e < 0.5$  is the proton-to-baryon ratio in the cloud. The produced photon number is small, comparable to that in Equation (27) and orders of magnitude smaller than  $n_{\gamma}/n_b$  generated by the shock discussed below.

Deep inside the massive cloud, the photon number may be calculated taking into account the synthesis of heavy neutron-rich elements and their  $\beta$ -decay at a later stage. However, nucleosynthesis of elements heavier than helium is inefficient in the outermost, low-density layers of main interest for us, and so, we do not include this effect here.

Let us now consider the production of photon number by an internal shock in the cloud. The shock boosts  $n_{\gamma}/n_b$ , because it generates entropy and radiation at radii  $r \sim v_0 t_{\text{sh}} \gg R_0$ . The jump conditions for a shock with speed  $v_s = \beta_s c$  give the generated energy density  $U \approx 2\gamma_s^2 \beta_s^2 \rho c^2$ , where  $\rho$  is the fluid density ahead of the shock. We are interested in the outer layers of the cloud,  $m \ll M_{\text{ej}}$ , where the shock accelerates to  $\gamma_s \beta_s \gtrsim 1$ . The cloud density  $\rho(m)$  may be estimated as follows. Each layer of the cloud at radii of interest expands ballistically, with some positive velocity gradient, and the expansion is homologous, with  $\rho(m, t) \propto t^{-3} \propto r^{-3}$ . The parameter  $\xi = r/\Delta r$  (which describes the sharpness of the density decline at the outer edge) freezes for homologous ballistic expansion, and so, the density of the preshock layers may be written as

$$\rho(m, t) \approx \frac{\xi m}{4\pi r^3(t)}, \quad \xi = \frac{r}{\Delta r} \gtrsim 1. \quad (29)$$

Using this equation for  $\rho$  and Equation (23) for  $\gamma_s \beta_s$ , one can express  $U \approx 2\gamma_s^2 \beta_s^2 \rho c^2$  as a function of the mass coordinate of the propagating shock,  $m$ . This gives the following estimate for the postshock energy density:

$$U(m) \sim 2\gamma_s^2 v_s^2(m) \rho(m) \sim \frac{\tilde{\xi} c^2}{r^3} \left( \frac{\xi m}{4\pi} \right)^{0.6} M_{\text{ej}}^{0.4}. \quad (30)$$

Here,  $r \sim R_{\text{sh}}$  is the shock breakout radius. Assuming that the energy density  $U$  is quickly thermalized, we can estimate the blackbody radiation density  $aT^4 \approx U$ , which gives the temperature

$$kT \sim 40 m_{26}^{0.15} r_9^{-3/4} \left( \frac{M_{\text{ej}}}{0.01 M_{\odot}} \right)^{0.1} \left( \frac{\tilde{\epsilon}}{0.1} \right)^{1/4} \text{ keV}. \quad (31)$$

The generated photon number per nucleon in the region where the shock accelerates to  $\gamma_s \gg 1$  is given by

$$\frac{n_{\gamma}}{n_b} \sim \frac{\gamma_s m_p c^2}{\sqrt{2} 2.7 kT} \sim 10^4 m_{26}^{-0.35} r_9^{3/4} \quad (\text{after shock}). \quad (32)$$

In this last estimate, we have omitted the weak dependence on  $M_{\text{ej}}$  and  $\tilde{\epsilon}$ .

The assumption of quick thermalization of radiation in a heated flow can be verified as follows (e.g., Levinson 2012; Beloborodov 2013). Radiation must relax to a Planckian spectrum with the photon density  $n_{\gamma} = U/2.7kT$  if the plasma efficiently emits photons. The two main processes of photon production by the thermal plasma are double Compton scattering and bremsstrahlung. In particular, double Compton scattering occurs with a rate  $\dot{n}_{\text{DC}} \approx 0.1 n_{\gamma} n \sigma_T c \Theta^2$ , where  $n$  is the electron/positron number density and  $\Theta = kT/m_e c^2$ . The number of photons produced during the expansion timescale for the postshock plasma is  $n_{\text{DC}} \approx \dot{n}_{\text{DC}} r/c$ . It becomes exponentially large, ensuring thermalization, if

$$\frac{n_{\text{DC}}}{n_{\gamma}} \approx 0.1 n \sigma_T r \Theta^2 = 0.1 \xi \tau_T \Theta^2 \gg 1. \quad (33)$$

Here,  $\tau_T \approx n \sigma_T \Delta r$  is the scattering optical depth of the ejecta outside the current shock radius. It is related to the mass coordinate  $m$  by

$$\tau_T = \frac{\sigma_T Y_e m}{4\pi r^2 m_p} \approx 3 \times 10^4 Y_e m_{26} r_{10}^{-2}, \quad (34)$$

where  $\sigma_T$  is the Thomson cross section, and  $Y_e = n_p/n_b$  is the proton-to-baryon ratio. For simplicity, we neglected the possible presence of  $e^{\pm}$  pairs (which could only increase the photon production rate). One can see that the condition of Equation (33) is satisfied, and thus, the shock-generated radiation is thermalized in the shells of interest  $m > 10^{25}$  g, as long as the shock crosses the cloud before it expands to  $r \sim 10^{10}$  cm.

#### 4.4. Free Neutrons

An important feature of the merger ejecta is their neutron-rich composition  $Y_e < 0.5$ . The baryons at the base of the outflow are initially free nucleons, predominantly neutrons. As the matter expands and cools, the nucleons recombine into  $\alpha$ -particles, and the neutron excess implies some leftover free neutrons. Deep inside the massive cloud, most of the free neutrons become locked into heavy, neutron-rich nuclei after  $\sim 1$  s of expansion (e.g., Metzger et al. 2010). However, this process is less efficient in the outer layers of main interest here,  $m \sim 10^{25}$ – $10^{27}$  g, because their density is well below the typical density inside the massive cloud.

The free neutrons and ions are still well coupled by frequent nuclear collisions; so to a first approximation, one can treat them as a single fluid. However, this approximation is not valid

on small scales comparable to the shock thickness, and the drift of neutrons relative to the ions changes the shock dissipation mechanism (Beloborodov 2017). In the absence of free neutrons, the shock is mediated by radiation and has a thickness comparable to the photon free path. In the presence of free neutrons, the shock is partially mediated by neutrons, which have much longer free paths. In addition, the neutron–ion collisions around the shock cause spallation of  $\alpha$ -particles (Belyanin et al. 2001; Beloborodov 2003).

When the shock becomes highly relativistic,  $\gamma_s \beta_s \gtrsim 1$ , the neutron–ion collisions in the shock become inelastic and generate pions. The pions immediately decay into ultrarelativistic leptons and generate a nonthermal inverse Compton cascade. In the presence of a magnetic field, the cascade would be capable of producing a significant photon number through synchrotron emission (Beloborodov 2010; Vurm & Beloborodov 2016). However, the envelope magnetization is likely low—both mechanisms of the envelope ejection (Sections 3 and 4) suggest that it expands outside the magnetic fields generated by the merger. This suggests weak synchrotron emission by the cascade from neutron collisions.

#### 4.5. Photospheric Emergence of the Early Shock

We also note that the early shock breakout radiates little energy when it reaches the photosphere of the cloud. It does not produce a detectable burst, because of the modest emission radius  $r \sim R_{\text{sh}} < 10^{10}$  cm. The photosphere is located in the outermost layers with mass

$$m_{\star} \sim \frac{4\pi r^2 m_p}{Y_e \sigma_T} \approx 3 \times 10^{21} \frac{r_{10}^2}{Y_e} \text{ g}, \quad (35)$$

which we estimated from Equation (34) by setting  $\tau_T \sim 1$ .<sup>9</sup> Since  $m_{\star}$  is so small, the internal shock must accelerate to a high Lorentz factor  $\gamma_s(m_{\star}) \sim 10$  as it reaches the photosphere. The energy of the shocked photospheric layers is given by

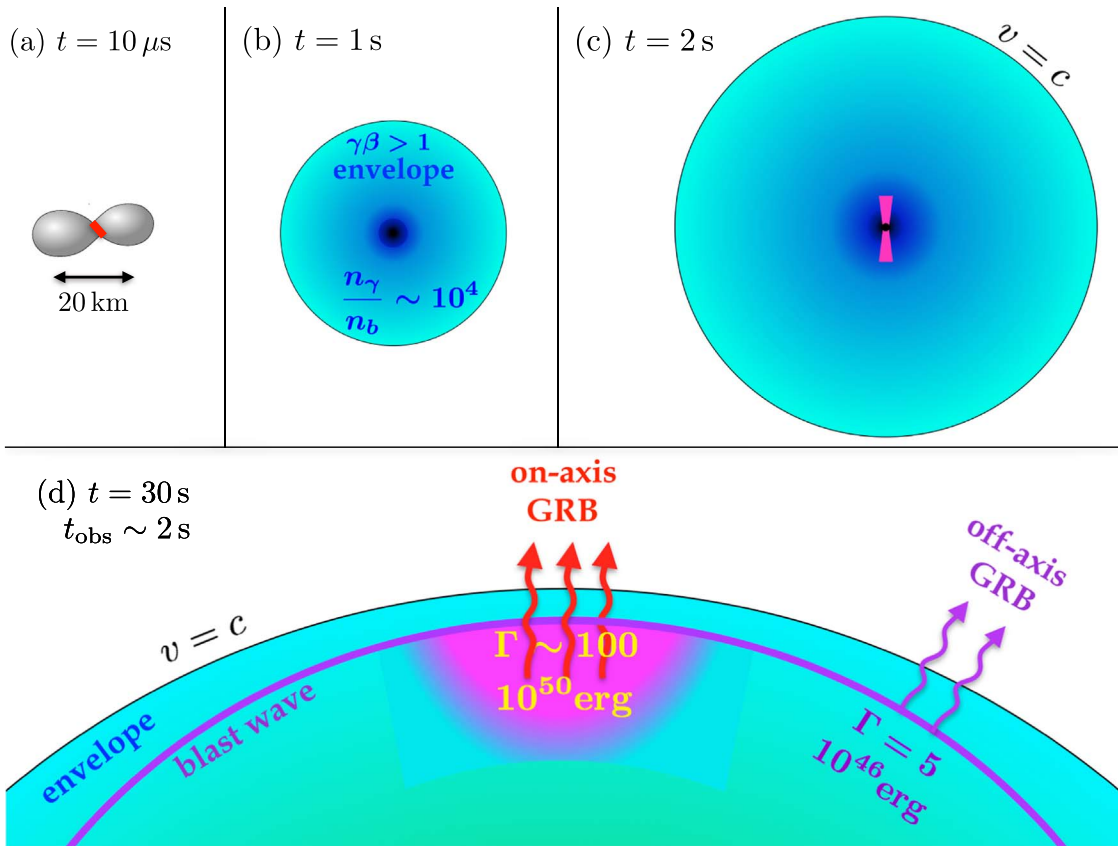
$$\mathcal{E}_{\star} \sim \gamma_s^2(m_{\star}) m_{\star} c^2. \quad (36)$$

It is 2–3 orders of magnitude smaller than the energy of GRB 170817A, and thus, the early shock breakout is hardly capable of emitting detectable radiation. The large number of photons produced by the shock inside the cloud remain trapped by the huge optical depth and experience strong adiabatic cooling.

### 5. GRB Production in the Envelope

The main conclusion from the preceding sections is that the merger GW170817 likely ejected a low-mass, opaque envelope expanding with a stratified Lorentz factor  $\gamma(m) \gg 1$ . The cold ballistic envelope becomes capable of emitting a GRB only if it is reheated by some dissipation process. A simple way to accomplish this is to drive a new shock wave. Therefore, below we consider a scenario where the merger remnant produces a delayed explosion (e.g., Gottlieb et al. 2018). In particular, if the remnant is a supermassive neutron star with a limited lifetime (Lipunova & Lipunov 1998), the explosion may be associated with its collapse. The collapse is promoted by the generation of ultrastrong magnetic fields and loss of differential

<sup>9</sup> Density of the photospheric layers is low compared with the inner parts of the cloud. Therefore, heavy nuclei are not synthesized in these layers, and there is no bound-free absorption of photons, so we assume Thomson opacity.



**Figure 8.** Schematic summary of the explosion model. Panel (a): two neutron stars begin to merge ( $t \sim 10^{-5}$  s). At this stage, a hot sandwich (red) forms between the stars, and its powerful neutrino emission ablates the neutron star crust. Panel (b): the merger remnant (a magnetar) is engulfed by its ejecta, shown at time  $t = 1$  s. The ejecta are pictured schematically as a spherical cloud; real ejecta may not be spherical. The dense cloud of mass  $M_{\text{ej}} \gtrsim 10^{-2} M_{\odot}$  (dark blue) is expanding with speed  $v \sim (0.1\text{--}0.3)c$ . It will emit the kilonova a day later. The dense cloud is surrounded by the large ultrarelativistic envelope ( $\gamma\beta > 1$ ) of mass  $m \sim 10^{-7} M_{\odot}$  (cyan). It was ejected by an internal shock wave that broke out of the cloud at  $t < 1$  s. The shock also loaded the envelope with a large number of photons  $n_{\gamma}/n_b \sim 10^4$ , which are trapped and adiabatically cooled. The envelope will continue to expand ballistically and homologously. The profile of its Lorentz factor  $\gamma(m)$  diverges toward the outer edge, and the radial scale of its density variation is compressed as  $\Delta r \sim r/\gamma^2$ . Panel (c): the magnetar has collapsed into a black hole at  $t \lesssim 2$  s, and a pair of powerful jets (bright magenta) have been launched by the accreting black hole. The jets will chase the outermost, ultrarelativistic layers of the envelope for a long time. Panel (d): the blast wave from the jet has expanded far into the ultrarelativistic envelope at all polar angles  $\theta$ . The radius of its photospheric emergence depends on  $\theta$ , because its energy and Lorentz factor  $\Gamma$  decrease with  $\theta$ . By  $t \sim 30$  s ( $t_{\text{obs}} \sim t/2\gamma^2 \sim 2$  s), the blast wave has reached the photosphere at  $\theta \sim 0.5$ , the viewing angle of GW170817. The indicated characteristic energies of  $10^{50}$  erg and  $10^{46}$  erg are the outflow energies at  $\theta \sim 0.1$  (the jet core) and  $\theta \sim 0.5$ . The corresponding isotropic equivalent of the energy is larger by a factor of  $\sim(1 - \cos\theta)^{-1}$ , which gives  $\sim 10^{52}$  erg for the jet core and  $\sim 10^{47}$  erg at  $\theta \sim 0.5$ . An ultrabright on-axis burst is emitted by the jet plasma (bright magenta), which is more energetic than the blast wave ahead of it. The on-axis burst is powered by variable dissipation inside the jet and may have a complicated light curve. The off-axis burst is emitted by the blast wave in the envelope; its  $\gamma$ -ray light curve has a single peak and is followed by a soft X-ray tail.

rotation, as well as by cooling due to neutrino emission, on a timescale of a few seconds. Then, the nascent spinning black hole launches powerful, ultrarelativistic, magnetized jets.

Compared with the precollapse massive neutron star (the magnetar), the black hole is much more capable of launching the jets. During the collapse, the source of the baryonic wind polluting the magnetosphere of the magnetar disappears behind the event horizon. At the same time, the accretion disk of the merger debris continues to sustain a strong magnetic field threading the black hole. The Poynting flux from the black hole (of radius  $\sim 5$  km) may exceed that from the magnetar, because it is more compact than the magnetar and is spinning faster. These conditions are favorable for the formation of an ultrarelativistic jet, which is collimated by the surrounding slower ejecta.

Our proposed model for GRB production is schematically summarized in Figure 8. The jets first propagate inside the massive cloud and then expand into the large ultrarelativistic envelope. The forward shock from the jet (and its cocoon in the cloud) forms a blast wave, which initially expands forward and

sideways around the jet. Simulations by Duffell et al. (2018) suggest that the blast wave will be launched into the outer envelope if the jet itself is successful, i.e., if it exits the massive cloud. At later times, the blast-wave shape becomes nearly spherical as it travels with almost the speed of light and has a radius  $r \approx ct$ . The jet must be strongly collimated, as required by the late afterglow observations of GRB 170817A (Granot et al. 2018; Lamb et al. 2018; Lazzati et al. 2018; Mooley et al. 2018b). Therefore, the blast wave in the envelope has an anisotropic power; however, it is less beamed than the jet.

When viewed at large angles from the polar axis, the explosion emission will be dominated by the blast wave in the envelope rather than the jet itself. By contrast, when viewed on-axis, the jet kinetic energy will strongly dominate over the energy of its forward shock in the envelope, and the observer will see a much brighter beamed GRB emitted by the jet plasma with  $\Gamma \gtrsim 10^2$ .

Below, we focus on the off-axis GRB expected from the blast wave in the envelope and compare it with GRB 170817A.

We will assume that the jet launches the blast wave into the envelope at a time comparable to 1 s after the merger.

### 5.1. Homologous Density Profile

Let us first evaluate the radial structure of the relativistic envelope before the blast wave from the jet. The envelope structure can be calculated in the ballistic approximation. The picture becomes particularly simple at long times/large radii: the envelope may be thought of as a sequence of shells ejected at  $r \approx 0$  and  $t \approx 0$  with different speeds, so that a shell with mass coordinate  $m$  and velocity  $v(m)$  has the radius

$$r(m, t) \approx v(m)t. \quad (37)$$

A shell  $dm$  has the thickness  $dr = t |dv/dm| dm$ , and hence, the envelope has the following density distribution (measured in the fixed lab frame):

$$\rho_{\text{lab}}(m, t) \approx \frac{dm}{4\pi r^2 dr} \approx \frac{1}{4\pi v^2 t^3 |dv/dm|}. \quad (38)$$

The proper density  $\rho = \rho_{\text{lab}}/\gamma$  is then given by

$$\rho(m, t) \approx \frac{\gamma^2}{4\pi\beta^2 c^3 t^3} \left| \frac{d(\gamma\beta)}{dm} \right|^{-1}, \quad (39)$$

where we have used the relation  $d(\gamma\beta) = \gamma^3 d\beta$ .

In particular, for the four-velocity distribution of the form  $\gamma\beta = (m/m_1)^{-\psi}$ , we find

$$\rho(m, t) \approx \frac{m_1}{4\pi\beta^4 c^3 t^3 \psi} \left( \frac{m}{m_1} \right)^{1-\psi}. \quad (40)$$

The mass stratification of the ultrarelativistic homologous envelope is such that the radial thickness of an outer shell  $m$  occupies a radial thickness of  $\Delta r/r \sim \gamma^{-2}(m)$ . The outer (smaller)  $m$  occupies a progressively smaller  $\Delta r \sim m(dm/dr)^{-1}$  because it moves faster and has a stronger relativistic compression factor  $\gamma^{-2}$ .

### 5.2. Blast-wave Emergence at the Photosphere

The edge of the envelope has a diverging Lorentz factor  $\gamma$ , and so, it is out of reach for the blast wave. Effectively, the blast wave propagates in an unbound medium, resembling the shocks in external media that produce GRB afterglows. Unlike the standard afterglow model, here, the medium is opaque, and the shock is mediated by radiation. Furthermore, the blast wave is *accelerating*, because the envelope density is steeply decreasing with radius and time, and its Lorentz factor  $\gamma$  increases with radius. The growth of the blast-wave Lorentz factor  $\Gamma$  gradually allows it to catch up with faster (and less massive) outer shells of the envelope. Eventually, the blast wave emerges at the photosphere of the envelope and produces a pulse of observed emission.

The photosphere of the homogeneously expanding envelope is located where the column number density of protons is  $\sim \sigma_T^{-1}$ . The photosphere mass coordinate  $m_*$  and radius  $R_*$  are related by

$$m_* = \frac{4\pi R_*^2 m_p}{Y_e \sigma_T}. \quad (41)$$

Radiation produced by the blast wave begins to escape to a distant observer when the shock mass coordinate  $m_s(t)$  approaches  $m_*(t)$ .

Relativistic shocks in a photon-rich medium are capable of creating copious  $e^\pm$  pairs (Beloborodov 2017; Ito et al. 2018; Lundman et al. 2018). This effect tends to prolong the photospheric emergence of the shock and delay its transition to complete transparency. However, the shock in the envelope described above will have a mildly relativistic jump, and pair creation will not be so efficient (we leave this for future study). In any case, the optical depth of pairs created by the shock is limited to  $\sim 1$  by  $e^\pm$  annihilation (Beloborodov 2017). Therefore, radiation will begin to escape the blast wave when  $m_s = m_*$  even if pair creation is strong.

The observed delay of the shock appearance at the photosphere  $R_*$  is given by

$$t_* = t_{\text{obs}}(R_*) \approx \frac{R_*}{2c\gamma_*}, \quad \gamma_* \equiv \gamma(m_*) \approx \left( \frac{m_*}{m_1} \right)^{-\psi}. \quad (42)$$

Hereafter, we assume that the relativistic envelope was inflated by the mechanism described in Section 4, and we use Equation (26). Combining Equations (41) and (42), we find

$$\gamma_*^{(1+4\psi)/\psi} = \frac{\sigma_T Y_e m_1}{16\pi m_p c^2 t_*^2}. \quad (43)$$

Note that  $\psi/(1+4\psi) = (1/8) \pm 0.01$  in the relevant range of  $0.2 < \psi < 0.3$ , and so,

$$\gamma_* \approx \left( \frac{\sigma_T Y_e m_1}{16\pi m_p c^2 t_*^2} \right)^{1/8} \approx 3 t_*^{-1/4} m_{1,27}^{1/8}, \quad (44)$$

where the observed time of the photospheric shock breakout  $t_*$  is expressed in seconds. Then, we also find  $R_*$  from Equation (42),

$$R_* \approx 5 \times 10^{11} t_*^{1/2} m_{1,27}^{1/4} \text{ cm}, \quad (45)$$

and the mass of the photospheric layers  $m_*$  (Equation 41),

$$m_* \approx 2 \times 10^{25} t_*^{1/2} m_{1,27}^{1/2} \text{ g}. \quad (46)$$

The shock power depends on the jump of the fluid Lorentz factor,

$$\chi \equiv \frac{\Gamma}{\gamma}. \quad (47)$$

The shock jump conditions give the relative velocity between the downstream ( $\Gamma$ ) and the upstream ( $\gamma$ ):  $\beta_{\text{rel}} = (\chi^2 - 1)/(\chi^2 + 1)$ , and the corresponding Lorentz factor is  $\gamma_{\text{rel}} = (1 - \beta_{\text{rel}}^2)^{-1/2} = (\chi^2 + 1)/2\chi$ . The energy dissipated by the shock as it crosses the photospheric layers of mass  $\sim m_*$  is given by

$$\begin{aligned} \mathcal{E}_* &\approx \frac{(\chi_* - 1)^2}{2} \gamma_* m_* c^2 \\ &\approx 3 \times 10^{46} (\chi_* - 1)^2 t_*^{1/2} m_{1,27}^{5/8} \text{ erg}. \end{aligned} \quad (48)$$

The photospheric shock breakout radiates a pulse of radiation with energy  $\sim \mathcal{E}_*$  and the characteristic peak duration  $\Delta t_{\text{obs}} \approx \Delta t_* \approx R_*/2\Gamma_* c$  determined by the Lorentz factor  $\Gamma_* = \Gamma(m_*)$  of the postshock plasma. The peak width can be

compared with its arrival time  $t_{\text{obs}} \approx t_*$ ,

$$\frac{\Delta t_{\text{obs}}}{t_{\text{obs}}} = \frac{\Delta t_*}{t_*} \approx \frac{\gamma_*^2}{\Gamma_*^2} \equiv \chi_*^{-2}. \quad (49)$$

Thus, the relative width of the observed peak,  $\Delta t_*/t_*$ , can be used as a proxy for the shock strength at the photosphere.

Elsewhere, we describe in more detail the blast-wave propagation through the homologous relativistic envelope and evaluate emission expected at  $t_{\text{obs}} \gg t_*$ , after the main peak of the off-axis GRB. It is produced by the deeper shells  $m_s \gg m_*$ , which were heated by the blast wave at smaller radii  $r_s \ll R_*$  and optical depths  $\tau_s \gg 1$ . The heated opaque shells behind  $m_*$  release their radiation with a delay  $\Delta t_{\text{obs}}(m_s)$ , when they expand to transparency. This delayed emission is partially thermalized and adiabatically cooled, and thus, it is much softer than the GRB peak.

### 5.3. Comparison with GRB 170817A

The above predictions can be compared with observations of GRB 170817A. Its arrival time was  $t_* \approx 1.7$  s and the main pulse had a width  $\sim 3$  shorter than  $t_*$ , which implies  $\chi_* \sim \sqrt{3}$  according to Equation (49). Substituting these values to Equations (48), we find

$$\mathcal{E}_* \sim 2 \times 10^{46} m_{1.27}^{5/8} \text{ erg}. \quad (50)$$

One can see that the envelope with  $m_1 \sim 3 \times 10^{27}$  g (which is in the expected range for the envelope model in Sections 4) is consistent with the observed energy of the main peak of GRB 170817A,  $\mathcal{E} \sim 4 \times 10^{46}$  erg (Goldstein et al. 2017). The corresponding mass of the photospheric layers is  $m_* \sim 6 \times 10^{25}$  g.

Furthermore, from Equations (44) and (45), we find that the preshock ejecta at the photosphere had Lorentz factor

$$\gamma_* \approx 3, \quad (51)$$

and the blast wave broke out at radius

$$R_* \approx 10^{12} \text{ cm}. \quad (52)$$

The Lorentz factor of the radiating plasma immediately behind the shock is

$$\Gamma_* = \chi_* \gamma_* \approx 5. \quad (53)$$

The observed spectrum of the initial pulse peaked at  $E_{\text{pk}} \sim 10^2$  keV, which roughly corresponds to the average photon energy  $\bar{E} \sim 10^2$  keV (the detailed shape of the spectrum of GRB 170817A is uncertain, because of poor photon statistics). This should be compared with the average energy of photons emitted by the blast wave at  $R_*$ ,

$$\begin{aligned} \bar{E}_* &\approx \frac{\Gamma_* \gamma_{\text{rel}} m_p c^2}{n_\gamma / n_b} = \frac{(\chi_* - 1) \gamma_* m_p c^2}{2 n_\gamma / n_b} \\ &\approx 100 \left( \frac{n_\gamma / n_b}{10^4} \right)^{-1} \text{ keV}. \end{aligned} \quad (54)$$

Thus, the observed  $E_{\text{pk}}$  is consistent with the photon-to-baryon ratio  $n_\gamma / n_b \sim 10^4$  expected in the envelope described in Section 4; see Equation (32).

GRB 170817A was also reported to have a soft tail of emission after the main peak. In an accompanying paper, we study the soft emission expected after the blast wave breaks out

of the opaque relativistic envelope. It has a decreasing luminosity  $L_{\text{obs}}(t_{\text{obs}})$  and a decreasing average photon energy  $\bar{E}(t_{\text{obs}})$ . However, quantitative tests of the tail prediction are difficult for GRB 170817A, because its tail is barely detected and its properties are poorly known and still debated (see Burgess et al. 2017).

## 6. Discussion

The observed timing of GRB 170817A and its luminosity implies ultrarelativistic expansion of the gamma-ray source,  $\Gamma \gtrsim 5$  (Figure 1). This constraint shows that neutron star mergers eject ultrarelativistic outflows at large angles from the rotation axis, at least up to  $\theta \sim 20\text{--}30^\circ$  (the viewing angle for GW170817).

We have argued that this broad ultrarelativistic outflow has the form of a self-similar envelope expanding from the center with a stratified Lorentz factor. This picture follows from our investigation of a possible mechanism for ultrarelativistic ejecta. Both mechanisms described in Sections 3 and 4 inflate a self-similar envelope with a Lorentz factor profile increasing outward. This envelope contains significant mass and is opaque. In particular, the ultrarelativistic envelope ejected by a magnetar shock (Section 4) can have a mass exceeding  $10^{-7} M_\odot$ . This is sufficient for inflating the GRB photosphere to radii  $r \sim 10^{12}$  cm.

A plausible scenario for producing gamma-rays invokes a delayed explosion from the merger remnant. The explosion launches a blast wave into the inflated envelope, which eventually emerges at its photosphere and emits a gamma-ray burst (Section 5, see Figure 8). The burst radius  $r \approx 10^{12}$  cm and Lorentz factor  $\Gamma \approx 5$  (the red circle in Figure 1) as well as the predicted burst luminosity  $L_\gamma \approx 10^{47} \text{ erg s}^{-1}$ , are in agreement with observations (Section 5.3). Furthermore, the expected average energy of the emitted photons  $\bar{E} \sim 10^2$  keV is consistent with observations.

### 6.1. Comparison with Previous Work

Our model for GRB 170817A shares some features with the shock breakout models of Gottlieb et al. (2018), Bromberg et al. (2018), and Nakar et al. (2018). However, there are important differences.

(i) The previous models require a small Lorentz factor  $\Gamma \sim 1$  in order to explain the observed photon energy  $\bar{E} \sim 10^2$  keV. These models adopted the plasma temperature behind the shock  $kT \approx 50$  keV, regulated by  $e^\pm$  creation, as discussed in earlier papers (e.g., Nakar & Sari 2012). Then, the observed average photon energy  $\bar{E} \approx 3kT \Gamma \approx 150 \Gamma$  keV implies  $\Gamma \sim 1$ . By contrast, we find that the observed light curve requires  $\Gamma \gtrsim 5$ . Therefore, we conclude that the shock model with  $kT \approx 50$  keV is in tension with observations of GRB 170817A.

(ii) The previous models assumed that shock breakout occurs in a photon-poor cloud,  $n_\gamma / n_b \ll 10^3$ . By contrast, the envelope inflated by the mechanism described in Section 4 is photon rich,  $n_\gamma / n_b > 10^4$ . Then, the delayed explosion in the envelope emits a spectrum with a reduced energy per photon in the jet rest frame, which is consistent with the high- $\Gamma$  Doppler boost giving the observed  $\bar{E} \sim 10^2$  keV. We have not calculated yet the detailed GRB spectrum expected from shock breakout in the photon-rich envelope. A similar problem was studied by Levinson (2012), and we are currently working on

complete, first-principle simulations that will give the emitted spectra with various  $n_\gamma/n_b \gg 10^3$ . The results will be reported in a future paper. Note that our photon-rich model neglected the photon number generated by the shock itself via downstream bremsstrahlung emission while the previous models relied on this emission. The self-generation of photons could reach the required  $n_\gamma/n_b \sim 10^4$  if the shock is slower and/or the outer layers of the envelope manage to synthesize heavy nuclei (Nakar 2019).

(iii) In the previous models, the shock acceleration and the production of gamma-rays occurred at a well defined characteristic radius—the cloud “edge” where density suddenly and steeply drops by many orders of magnitude. In this respect, the models were similar to the canonical shock breakout in a stellar explosion. By contrast, the envelope described in this paper is equivalent to an infinite medium. The fact that at any given time the envelope extends to a finite radius becomes irrelevant, since its leading edge has a diverging Lorentz factor and is out of reach for a blast wave.

The expanding envelope may be idealized as a flow ejected impulsively from the center with a self-similar (power-law) distribution of Lorentz factor  $\gamma(m)$ . Its density profile is determined by  $\gamma(m)$  and is also self-similar (Section 5.1). The acceleration of a blast wave launched in such an envelope occurs over a few decades in radius rather than at an edge of a cloud. This qualitatively changes the dynamics and radiation of the blast wave. It has to “chase” each layer of the envelope and catches up with layers of higher  $\gamma$  at progressively larger radii  $r \propto \gamma^2$  until finally reaching the photosphere.

## 6.2. Future Observational Tests

Our results suggest a few observational implications that may be tested in the future.

(1) Our model for off-axis short GRBs predicts that the relative width of the gamma-ray pulse  $\Delta t_{\text{obs}}/t_{\text{obs}}$  reflects the shock strength at the photosphere (Equation 49). The blast-wave power is expected to decrease with the viewing polar angle  $\theta$ . This suggests that the luminosity of the gamma-ray counterpart should decrease with  $\theta$  while  $\Delta t_{\text{obs}}/t_{\text{obs}}$  should increase. Our model also predicts an anticorrelation between the pulse hardness  $\bar{E}$  and relative duration  $\Delta t_{\text{obs}}/t_{\text{obs}}$ , as both are controlled by the blast-wave strength at the photosphere,  $\Gamma_*/\gamma_*$ . These correlations may be tested by future observations. The blast wave is fastest when viewed on-axis, directly in front of the powerful collimated jet with  $\Gamma \gtrsim 10^2$ . Its gamma-ray emission is also shortest when viewed on-axis,  $\Delta t_{\text{obs}}/t_{\text{obs}} \ll 1$ . However, the on-axis luminosity from the blast wave should be outshined by the extremely bright, beamed emission from the jet itself, and therefore, the above correlations should break at small  $\theta$  where the jet comes into view. The jet is expected to emit a canonical short GRB many orders of magnitude brighter than GRB 170817A.

(2) Relativistic ablation of the neutron star surface at the onset of the merger (Sections 3) suggests an immediate gamma-ray burst, overlapping with the gravitational wave signal. Relativistic ablation creates an ultrarelativistic fireball in a short time  $t \sim 3 \times 10^{-5}$  s after the two stars touch. Its energy  $\mathcal{E}_f$  has a flat distribution over  $\gamma$ , up to enormous  $\gamma \sim 10^3$  (Figure 5). Its outermost, fastest layers become transparent while still being radiation dominated,  $w \gtrsim 1$ , and, hence, radiate away most of their energy, similar to the fireball models of Paczynski (1986) and Goodman (1986). Thus, a significant

fraction of the ablation fireball energy  $\mathcal{E}_f$  is radiated away. The fireball energy is quite uncertain though. It is sensitive to both the neutrino luminosity from the collision sandwich and the geometric parameter  $a$  that describes the effective area of the neutrino-sphere near the stellar surface. In the most optimistic simulation with  $a = 4$  km, this energy is  $\mathcal{E}_f \sim 3 \times 10^{45}$  erg; then, the ablation fireball emits a quasi-thermal burst with luminosity up to  $10^{51}$  erg s $^{-1}$  and duration  $\sim 10^{-5}$  s. Its observed temperature is close to the temperature at the base of the outflow,  $kT \sim 3\text{--}6$  MeV. Even in the optimistic model, this initial, hard “ablation burst” is weak and difficult to detect; however, it might become detectable with future, more sensitive detectors.

(3) The ejecta acceleration by the internal shock simulated in Section 4 has observational consequences for the kilonova emission. Optical/IR emission from GW170817 can be explained as a superposition of a “blue” and a “red” component that came from material with different opacities, with and without synthesized lanthanides; see Kasen et al. (2017). Waxman et al. (2018) pointed out that the data could also be fitted by a model with a simple (fixed, gray) opacity, if the emitting material was ejected with a power-law velocity distribution  $v \propto m^{-\psi}$  with  $\psi \sim 0.6$ . For the cloud simulated in Section 4, both opacity and velocity may be expected to vary. The early internal shock creates a monotonic four-velocity distribution with a changing slope  $\psi = -d \ln(\gamma\beta)/d \ln m$  (Figure 7). In the present paper, we focused on the outer layers ( $m \ll M_{\text{ej}}$ ), which are relevant for the GRB production. The kilonova is emitted by much deeper layers (the massive part of the cloud,  $m$  comparable to  $M_{\text{ej}}$ ). Qualitatively, one may expect that the variation of speed and density across the postshock cloud will lead to the emission of blue and red kilonova components. The faster parts of the cloud will have a lower density, fail to produce lanthanides, and emit the blue component. The slower and denser parts may synthesize the lanthanide material of high opacity and emit the red component. However, it is unclear if our simple spherically symmetric simulations are capable of giving enough mass at low velocities needed for the observed red kilonova. The presence of heavier and slower outflow at large polar angles may need to be invoked to increase red emission and explain the GW170817 observations. We leave the detailed analysis of this topic for future work.

(4) The ejected envelope is eventually decelerated by an external medium and produces afterglow emission for a broad range of viewing angles. Such deceleration afterglow is in general expected for dynamical ejecta from mergers (Hotokezaka et al. 2018; Nakar & Piran 2018), regardless of the presence or absence of a collimated jet. Furthermore, Nakar & Piran (2018) argue that the initial slow rise of the afterglow of GW170817 comes from ejecta moving toward us, along the line of sight, rather than the collimated jet viewed from the side. At present, it is unclear if/when the envelope described in this paper can dominate the observed afterglow. Its calculated stratification  $\gamma\beta(m)$  (Figure 7) may be used to develop a detailed afterglow model and check if the decelerating envelope could overshadow the off-axis emission from the decelerating jet at late stages of the explosion.

We thank Ehud Nakar, the referee of this paper, for his comments. A.M.B. is supported by NSF grant AST-1816484, NASA grant NNX15AE26G, a grant from the Simons

Foundation #446228, and the Humboldt Foundation. Y.L. is supported by NSF grant AST-1816484. C.L. was supported by the Swedish National Space Board under grant No. Dnr. 107/16.

### Appendix Compressional Heating at the Collision Interface

The pressure growth in the sandwich between the colliding stars implies a strong compressional heating of their surface layers, as seen from the following consideration. Let  $\rho_0$  be the initial, preshock density of an old layer in the sandwich, and  $P_0$  is its pressure when it was just crossed by the shock. Pressure in low-density layers is strongly dominated by radiation (and  $e^\pm$  pairs), so their adiabatic index is  $\hat{\gamma} \approx 4/3$ . As long as neutrino cooling is negligible, the layer compression by increasing  $P$  to a higher density,  $\rho'_0$ , occurs adiabatically,

$$\frac{\rho'_0}{\rho_0} = \left( \frac{P'_0}{P_0} \right)^{1/\hat{\gamma}} \approx \left( \frac{P}{P_0} \right)^{1/\hat{\gamma}}, \quad (\text{A1})$$

where we used pressure balance across the sandwich,  $P'_0 \approx P$ . Energy per baryon in the compressed layer grows proportionally to  $P'_0/\rho'_0 \propto P^{1-1/\hat{\gamma}}$ .

Instead of energy per baryon, it is more convenient to consider dimensionless enthalpy per unit mass,  $w$ . The initial  $w_0 = (P_0 + U_0)/\rho_0 c^2$  is related to  $\hat{E}$  defined in Equation (5) by

$$w_0 = \frac{4}{3} \frac{\hat{E}}{c^2}. \quad (\text{A2})$$

As the sandwich pressure grows to  $P \gg P_0$ , the dimensionless enthalpy of the layer is amplified as

$$\frac{w'_0}{w_0} = \left( \frac{P}{P_0} \right)^{1-1/\hat{\gamma}} \approx \left( \frac{P}{P_0} \right)^{1/4} \approx \left( \frac{\rho}{\rho_0} \right)^{1/4}, \quad (\text{A3})$$

where  $\rho \gg \rho_0$  is the present density of matter just upstream of the shocks. Note that  $\rho$  and  $\rho_0$  are the preshock densities of different layers;  $\rho/\rho_0$  should not be confused with the compression factor of the old layer due to the increasing pressure,  $\rho'_0/\rho_0$ . The last equality in Equation (A3) may be slightly changed (by a numerical factor close to unity) when the shock pressure becomes dominated by nucleons, which leads to the postshock adiabatic index  $\hat{\gamma} \approx 5/3$  instead of  $4/3$ . However, the scaling  $w'_0/w_0 \propto \rho^{1/4}$  is weakly affected by this transition, because the scaling applies to the old, low-density layers, which remain radiation dominated with  $\hat{\gamma} = 4/3$ .

The moderate initial  $w_0 < 0.1$  increases with time to  $w'_0 > 1$  when the shocks propagate into layers of density  $\rho > w_0^{-4} \rho_0$ . The layers compressed to relativistic enthalpy  $w'_0 \gg 1$  have the potential of being ejected with highly relativistic speeds, if their internal energy has a chance to convert to bulk kinetic energy via adiabatic expansion without losing it to the neighboring heavy and nonrelativistic layers.

The maximum compression factor may be reached close to the moment when the squeezed matter begins to leak out from the sandwich. At the beginning of the collision, the size of the collision interface in the  $y$ - $z$  plane expands with a superluminal speed  $dr_A/dt$ . Here,  $r_A = (y_A^2 + z_A^2)^{1/2}$  represents the curve where the surfaces of the two stars intersect, which defines the edge of the sandwich (this curve is not a circle, as the tangential

motion of the colliding stars breaks the axial symmetry of the interface). The interface area grows with time because of the converging motion of the two stars, which brings it into contact with more material. The two shocks bounding the collision sandwich intersect at  $r_A$ , which initially grows with rate  $dr_A/dt > c$ . Later,  $dr_A/dt$  is reduced (because of the curvature of the stellar surfaces) and eventually becomes subluminal and subsonic. Then, the hot interface material pushes the shocks aside (so that they no longer intersect at the edge) and leaks out with the sound speed through the edges of the sandwich. This mass loss will buffer the growth of pressure in the sandwich.

The time at which the sandwich matter begins to be ejected from the edges may be roughly estimated as  $t \sim R/c \approx 3 \times 10^{-5}$  s. During this time, the two shocks propagate a significant distance of  $x \sim t v_{\text{sh}} \gtrsim 1$  km into the deep stellar interior, where the upstream density can be quite large,  $\rho \sim 10^{13}$ – $10^{14}$  g cm $^{-3}$ . Therefore, we estimate the maximum possible compression in the sandwich using the ram pressure that corresponds to the upstream  $\rho \sim 10^{14}$  g cm $^{-3}$ . The corresponding  $\rho_0$  of layers reaching  $w'_0 > 1$  is found from Equation (A3),

$$\rho_0 \approx 10^{10} \left( \frac{w_0}{0.1} \right)^{-4} \rho_{14} \text{ g cm}^{-3}. \quad (\text{A4})$$

Recalling that  $\rho_0$  is the preshock density of subsurface layers with a hydrostatic scale-height  $h \sim 0.3$ – $1$  km, one can roughly estimate their mass as  $m \sim r^2 h \rho_0$ , where  $r \lesssim R \sim 10$  km is the characteristic extension of the old layers along the interface. The initial volume occupied by these layers is  $V_0 \sim r^2 h \lesssim 10^{16}$  cm $^3$ , and their nucleon mass is

$$m \sim V_0 \rho_0 \sim 10^{26} \left( \frac{w_0}{0.1} \right)^4 \rho_{14} \text{ g}. \quad (\text{A5})$$

This mass has the potential of being ejected with a highly relativistic speed. However, a more detailed analysis suggests that this mechanism is hindered by two processes in the sandwich between the colliding stars. (1) The idealized picture where mass  $m$  estimated in Equation (A5) resides in a very thin (strongly compressed) layer is likely incorrect, because the sandwich is also the site of Kelvin–Helmholtz instability. The high- $w$  material is likely to be dispersed into small bubbles or filaments and mixed into dense, massive, low- $w$  material before escaping the sandwich. Only a small fraction of the bubbles might be able to escape with a highly relativistic momentum and avoid sharing it with the nonrelativistic matter. (2) Neutrino emission and transport tends to steal energy from the layers with high  $w$  and reduce their pressure. This amplifies their compression and limits enthalpy per unit rest mass,  $w$ .

### References

- Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017a, *ApJL*, **848**, L13  
 Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017b, *PhRvL*, **119**, 161101  
 Abramowicz, M. A., Novikov, I. D., & Paczynski, B. 1991, *ApJ*, **369**, 175  
 Bauswein, A., Goriely, S., & Janka, H.-T. 2013, *ApJ*, **773**, 78  
 Beloborodov, A. M. 2003, *ApJ*, **588**, 931  
 Beloborodov, A. M. 2010, *MNRAS*, **407**, 1033  
 Beloborodov, A. M. 2011, *ApJ*, **737**, 68  
 Beloborodov, A. M. 2013, *ApJ*, **764**, 157  
 Beloborodov, A. M. 2017, *ApJ*, **838**, 125  
 Belyanin, A. A., Derishev, E. V., Kocharovskiy, V. V., & Kocharovskiy, V. V. 2001, arXiv:astro-ph/0111558  
 Birkel, R., Aloy, M. A., Janka, H.-T., & Müller, E. 2007, *A&A*, **463**, 51

- Bromberg, O., Tchekhovskoy, A., Gottlieb, O., Nakar, E., & Piran, T. 2018, *MNRAS*, **475**, 2971
- Burgess, J. M., Greiner, J., Begue, D., et al. 2017, arXiv:1710.05823
- Chen, W.-X., & Beloborodov, A. M. 2007, *ApJ*, **657**, 383
- Coulter, D. A., Foley, R. J., Kilpatrick, C. D., et al. 2017, *Sci*, **358**, 1556
- Dessart, L., Ott, C. D., Burrows, A., Rosswog, S., & Livne, E. 2009, *ApJ*, **690**, 1681
- Drout, M. R., Piro, A. L., Shappee, B. J., et al. 2017, *Sci*, **358**, 1570
- Duffell, P. C., Quataert, E., Kasen, D., & Klion, H. 2018, *ApJ*, **866**, 3
- Evans, P. A., Cenko, S. B., Kennea, J. A., et al. 2017, *Sci*, **358**, 1565
- Goldstein, A., Veres, P., Burns, E., et al. 2017, *ApJL*, **848**, L14
- Goodman, J. 1986, *ApJL*, **308**, L47
- Goodman, J., Dar, A., & Nussinov, S. 1987, *ApJL*, **314**, L7
- Gottlieb, O., Nakar, E., Piran, T., & Hotokezaka, K. 2018, *MNRAS*, **479**, 588
- Granot, J., Gill, R., Guetta, D., & De Colle, F. 2018, *MNRAS*, **481**, 1597
- Hallinan, G., Corsi, A., Mooley, K. P., et al. 2017, *Sci*, **358**, 1579
- Hotokezaka, K., Kiuchi, K., Kyutoku, K., et al. 2013, *PhRvD*, **87**, 024001
- Hotokezaka, K., Kiuchi, K., Shibata, M., Nakar, E., & Piran, T. 2018, *ApJ*, **867**, 95
- Ito, H., Levinson, A., Stern, B. E., & Nagataki, S. 2018, *MNRAS*, **474**, 2828
- Johnson, M. H., & McKee, C. F. 1971, *PhRvD*, **3**, 858
- Kasen, D., Metzger, B., Barnes, J., Quataert, E., & Ramirez-Ruiz, E. 2017, *Natur*, **551**, 80
- Kasliwal, M. M., Nakar, E., Singer, L. P., et al. 2017, *Sci*, **358**, 1559
- Kiuchi, K., Kyutoku, K., Sekiguchi, Y., & Shibata, M. 2018, *PhRvD*, **97**, 124039
- Kluźniak, W., & Ruderman, M. 1998, *ApJL*, **505**, L113
- Kyutoku, K., Ioka, K., & Shibata, M. 2014, *MNRAS*, **437**, L6
- Lamb, G. P., Mandel, I., & Resmi, L. 2018, *MNRAS*, **481**, 2581
- Lazzati, D., Deich, A., Morsony, B. J., & Workman, J. C. 2017a, *MNRAS*, **471**, 1652
- Lazzati, D., López-Cámara, D., Cantiello, M., et al. 2017b, *ApJL*, **848**, L6
- Lazzati, D., Perna, R., Morsony, B. J., et al. 2018, *PhRvL*, **120**, 241103
- Levinson, A. 2012, *ApJ*, **756**, 174
- Lipunova, G. V., & Lipunov, V. M. 1998, *A&A*, **329**, L29
- Lithwick, Y., & Sari, R. 2001, *ApJ*, **555**, 540
- Lundman, C., Beloborodov, A. M., & Vurm, I. 2018, *ApJ*, **858**, 7
- Margutti, R., Berger, E., Fong, W., et al. 2017, *ApJL*, **848**, L20
- Matsumoto, T., Nakar, E., & Piran, T. 2019, *MNRAS*, **483**, 1247
- Matzner, C. D., Levin, Y., & Ro, S. 2013, *ApJ*, **779**, 60
- Mestel, L., & Spruit, H. C. 1987, *MNRAS*, **226**, 57
- Metzger, B. D., Martínez-Pinedo, G., Darbha, S., et al. 2010, *MNRAS*, **406**, 2650
- Metzger, B. D., Thompson, T. A., & Quataert, E. 2007, *ApJ*, **659**, 561
- Metzger, B. D., Thompson, T. A., & Quataert, E. 2018, *ApJ*, **856**, 101
- Mooley, K. P., Deller, A. T., Gottlieb, O., et al. 2018a, *Natur*, **561**, 355
- Mooley, K. P., Frail, D. A., Dobie, D., et al. 2018b, *ApJL*, **868**, L11
- Nakar, E. 2007, *PhR*, **442**, 166
- Nakar, E. 2019, arXiv:1912.05659
- Nakar, E., Gottlieb, O., Piran, T., Kasliwal, M. M., & Hallinan, G. 2018, *ApJ*, **867**, 18
- Nakar, E., & Piran, T. 2018, *MNRAS*, **478**, 407
- Nakar, E., & Sari, R. 2012, *ApJ*, **747**, 88
- Paczynski, B. 1986, *ApJL*, **308**, L43
- Paczynski, B. 1998, in AIP Conf. Ser. 428, Gamma-Ray Bursts, 4th Hunstville Symp., ed. C. A. Meegan, R. D. Preece, & T. M. Koshut (Melville, NY: AIP), 783
- Pan, M., & Sari, R. 2006, *ApJ*, **643**, 416
- Pozanenko, A. S., Barkov, M. V., Minaev, P. Y., et al. 2018, *ApJL*, **852**, L30
- Price, D. J., & Rosswog, S. 2006, *Sci*, **312**, 719
- Qian, Y.-Z., & Woosley, S. E. 1996, *ApJ*, **471**, 331
- Radice, D., Galeazzi, F., Lippuner, J., et al. 2016, *MNRAS*, **460**, 3255
- Radice, D., Perego, A., Hotokezaka, K., et al. 2018, arXiv:1809.11161
- Savchenko, V., Ferrigno, C., Kuulkers, E., et al. 2017, *ApJL*, **848**, L15
- Siegel, D. M., & Metzger, B. D. 2017, *PhRvL*, **119**, 231102
- Soares-Santos, M., Holz, D. E., Annis, J., et al. 2017, *ApJL*, **848**, L16
- Tan, J. C., Matzner, C. D., & McKee, C. F. 2001, *ApJ*, **551**, 946
- Tanaka, M., Utsumi, Y., Mazzali, P. A., et al. 2017, *PASJ*, **69**, 102
- Tanvir, N. R., Levan, A. J., González-Fernández, C., et al. 2017, *ApJL*, **848**, L27
- Thompson, T. A., Burrows, A., & Meyer, B. S. 2001, *ApJ*, **562**, 887
- Thompson, T. A., Chang, P., & Quataert, E. 2004, *ApJ*, **611**, 380
- Troja, E., Piro, L., van Eerten, H., et al. 2017, *Natur*, **551**, 71
- Veres, P., Mészáros, P., Goldstein, A., et al. 2018, arXiv:1802.07328
- Vurm, I., & Beloborodov, A. M. 2016, *ApJ*, **831**, 175
- Waxman, E., Ofek, E. O., Kushnir, D., & Gal-Yam, A. 2018, *MNRAS*, **481**, 3423
- Xie, X., Zrake, J., & MacFadyen, A. 2018, *ApJ*, **863**, 58
- Zalamea, I., & Beloborodov, A. M. 2011, *MNRAS*, **410**, 2302
- Zel'dovich, Y. B., & Raizer, Y. P. 1967, Physics of Shock waves and High-temperature Hydrodynamic Phenomena (New York: Academic)
- Zrake, J., & MacFadyen, A. I. 2013, *ApJL*, **769**, L29