Quantitative modelling of type Ia supernovae spectral time series: Constraining the explosion physics

M. R. Magee[1]⋆, L. Siebenaler[2], K. Maguire[3], K. Ackley[1], T. Killestein[4]

[1]*Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[2]*Leiden Observatory, Leiden University, NL-2300 RA Leiden, Netherlands*
[3]*School of Physics, Trinity College Dublin, The University of Dublin, Dublin 2, Ireland*
[4]*Tuorla Observatory, Department of Physics and Astronomy, FI- 20014 University of Turku, Finland*

ABSTRACT

Multiple explosion mechanisms have been proposed to explain type Ia supernovae (SNe Ia). Empirical modelling tools have also been developed that allow for fast, customised modelling of individual SNe and direct comparisons between observations and explosion model predictions. Such tools have provided useful insights, but the subjective nature with which empirical modelling is performed makes it difficult to obtain robust constraints on the explosion physics or expand studies to large populations of objects. Machine learning accelerated tools have therefore begun to gain traction. In this paper, we present `riddler`, a framework for automated fitting of SNe Ia spectral sequences up to shortly after maximum light. We train a series of neural networks on realistic ejecta profiles predicted by the W7 and N100 explosion models to emulate full radiative transfer simulations and apply nested sampling to determine the best-fitting model parameters for multiple spectra of a given SN simultaneously. We show that `riddler` is able to accurately recover the parameters of input spectra and use it to fit observations of two well-studied SNe Ia. We also investigate the impact of different weighting schemes when performing quantitative spectral fitting and show that best-fitting models and parameters are highly dependent on the assumed weighting schemes and priors. As spectroscopic samples of SNe Ia continue to grow, automated spectral fitting tools such as `riddler` will become increasingly important to maximise the physical constraints that can be gained in a quantitative and consistent manner.

Key words: supernovae: general — radiative transfer

1 INTRODUCTION

Spectroscopic observations of type Ia supernovae (SNe Ia) are a key diagnostic probe of the explosion physics. Modelling and analysis of these observations give insight into the composition, density profile, and other physical properties of the SN that are directly linked to the explosion physics (see e.g. Jerkstrand 2017; Sim 2017 for overviews of some modelling techniques applied to SNe spectra). As spectroscopic datasets become increasingly large however, our ability to quickly and effectively model these data, so as to quantitatively extract the relevant physical parameters, has not been able to keep pace.

Multiple different modelling techniques are currently in use. Simple line identification codes (e.g. SYNAPPS, Thomas et al. 2011) provide useful information about which elements are present within the ejecta by allowing for individual tweaking of line strengths for different ions (e.g. Sullivan et al. 2011; Parrent et al. 2012; Hsiao et al. 2013). At the other extreme are highly sophisticated, self-consistent radiative transfer codes designed to produce synthetic observables from realistic explosion models (e.g. Höflich 1995, 2003; Blinnikov et al. 1998, 2006; Kasen et al. 2006; Kromer & Sim 2009; Hillier & Dessart 2012; van Rossum 2012; Ergon et al. 2018). These predic-

tions can then be directly compared against observations of a SN to determine whether it is consistent with a given explosion scenario (e.g. Baron et al. 2012; Dessart et al. 2014; Röpke & Sim 2018; Shen et al. 2021). Both of these options provide their own advantages and disadvantages. Line identification codes are often too parameterised and simplistic to provide robust predictions beyond the presence or absence of particular lines, while highly sophisticated codes come with huge computational expense that makes them ill-suited for exploration of large parameter spaces.

Radiative transfer codes intermediate to these two extremes have also been developed (e.g. Mazzali & Lucy 1993; Kerzendorf & Sim 2014). These codes are designed for significantly faster, empirical modelling of SNe spectra, and therefore make a few simplifying assumptions, but contain sufficient physics such that they may be used to produce a self-consistent model for interrogating the physical conditions of the ejecta. With these codes various input parameters, such as the temperature, density, and composition of the ejecta, can be used to generate synthetic spectra to directly confront the observations. The speed and flexibility of such codes means that they can be used to fit samples of SNe Ia (Mazzali et al. 2007). Investigating how changes to these parameters affect the quality of the fit can also provide useful insights (e.g. Stehle et al. 2005; Mazzali et al. 2008; Tanaka et al. 2011; Mazzali et al. 2014; Sasdelli et al. 2014; Ashall et al. 2016; Magee et al. 2016; Heringer et al. 2017; Magee et al.

⋆ E-mail: mrmagee.astro@gmail.com

2017, 2019; Vogl et al. 2019). Comparing these parameters against explosion model predictions can hence provide constraints on the explosion physics for individual objects or samples, however this empirical process is also challenging.

Determining which set of model parameters provide the best agreement is non-trivial. The most common approach is through visual inspection, in which input parameters are tweaked manually and the quality of the fit is determined purely through qualitative analysis. This makes it impossible to quantify the goodness of fit and experienced modellers may even disagree on which features are most important to reproduce, due to their own experience or biases. Furthermore, it is also not possible to determine whether the selected parameters represent a global or local minimum, as this would require computation of many thousands of spectra covering a huge parameter space that would each require visual inspection against the data.

Recent attempts have been made to provide more quantitative measures of the quality of fit and have shown promise (Ogawa et al. 2023). At least some visual inspection and fine-tuning are still required however and fits can be limited by the model spectra available. Incorporating machine learning techniques into spectroscopic analysis could overcome some of the primary limitations, including the computational cost associated with fitting large numbers of parameters. Chen et al. (2020) develop neural networks trained on TARDIS (Kerzendorf & Sim 2014) radiative transfer simulations and designed for predicting the chemical composition of SNe Ia around maximum light. Chen et al. (2024) apply this to estimate the ^{56}Ni abundance of SNe Ia and investigate how this correlates with observed light curve properties. Kerzendorf et al. (2021) present DALEK, which is designed to emulate TARDIS simulations and predict model spectra with a speed up of $\gtrsim 10\,000$. O'Brien et al. (2021) use DALEK to quantitatively fit a spectrum of SN 2002bo, a nomal SN Ia, approximately one week before maximum light and determine the chemical composition required to reproduce the observed spectral features. By comparing their inferred ejecta composition to explosion model predictions, they argue in favour of an explosion scenario containing a detonation for SN 2002bo. O'Brien et al. (2021) also find comparable results to previous studies without the need for manual and time-intensive tuning of model parameters, demonstrating the considerable power of this automated approach. O'Brien et al. (2023) extend this analysis to further apply DALEK to model spectra of a sample of normal and 91T-like SNe Ia (Li et al. 1999) using customised ejecta profiles and investigate differences in their physical properties.

Following from previous works, here we present riddler, a framework for automated, quantitative, and simultaneous fitting of full spectral time-series of SNe Ia up to shortly after maximum light. Similar to previous studies (Kerzendorf et al. 2021; O'Brien et al. 2021, 2023) we begin by training a series of neural networks to emulate TARDIS radiative transfer simulations. As in O'Brien et al. (2021, 2023), we then define a likelihood function and use nested sampling implemented in ULTRANEST (Buchner 2021) to perform quantitative comparisons between our emulated model spectra and observed spectra of SNe Ia. A key difference compared to previous works is that our training models cover a significantly larger time range and therefore rather than focus on indivdual spectra, our fitting approach has been extended to include multiple spectra at different epochs simultaneously. In addition, previous works have focused on inferring custom ejecta models for each SN. Here, we use realistic ejecta structures that are based on explosion simulations, allowing us to directly link the observed spectra back to predictions from explosion models and quantitatively determine the relative likelihood for the different ejecta structures predicted by the explosion models.

Using riddler, we are able to generate large numbers of model spectra with significantly reduced computational cost, enabling a detailed investigation of different metrics by which the goodness of fit may be quantified. In Sect. 2 we discuss our approach to generating a training dataset for our neural networks, which are described in Sect. 3. In Sect. 4 we discuss our approach to fitting spectra. Section 5 demonstrates riddler applied to spectra of SNe Ia, while Sect. 6 discusses the impact of different weighting schemes when calculating the likelihood. Finally, we discuss our results in Sect. 7 and present our conclusions in Sect. 8. riddler is publicly available on GitHub[1].

2 TRAINING MODELS

The purpose of this work is to investigate the early, optical spectral evolution of SNe Ia and investigate metrics through which to quantitatively determine the best-fitting ejecta structure predicted by different explosion models. For our spectral emulator neural networks, we therefore require a representative set of spectra that may be used for training. All model spectra used for training neural networks in this work were calculated with TARDIS (Kerzendorf & Sim 2014). TARDIS is an open-source, one-dimensional Monte Carlo radiative transfer code designed to rapidly produce model spectra based on a given set of input parameters defined by the user. The flexibility and speed of TARDIS, in combination with the appropriate level of physics included, means that it is ideally suited for producing the tens of thousands of spectra necessary for training neural networks. In Sect. 2.1 we discuss generating the TARDIS training set used for our neural networks, while in Sect. 2.2 we discuss the pre-processing applied before training the neural networks. The distributions of training parameters are shown in Fig. 1.

2.1 Generating the training set

Each TARDIS simulation for producing a SN Ia spectrum requires a number of input parameters: time since explosion, luminosity, inner boundary velocity, and the density and composition of the ejecta. With our neural networks we wish to cover a large parameter space, such that our models may be used to fit observations of SNe Ia at most epochs of interest. To determine a suitable range for each input parameter we use either samples of observed SNe Ia or predictions from theoretical explosion models.

We uniformly sample the time since explosion between 5 – 25 d. We select a lower time boundary of 5 d due to the increased computational cost in calculating models at very early times (as a result of the high density), while an upper boundary of 25 d is selected due to the photospheric approximation made by TARDIS. TARDIS assumes a sharp inner boundary separating optically thick and thin regions, which limits its validity to times up to shortly after maximum light (see Sect. 7.1). For ~10% of models, we find that our TARDIS simulations did not converge after 20 iterations and therefore exclude them from our training datasets. These models were typically at early times and with high densities at the inner boundary. This results in a somewhat skewed distribution for the time since explosion, rather than uniform.

Given that our sampled times since explosion cover a wide range of values, uniform sampling of the SN luminosity between fixed upper and lower boundaries would result in many models that are

---

[1] https://github.com/MarkMageeAstro/riddler
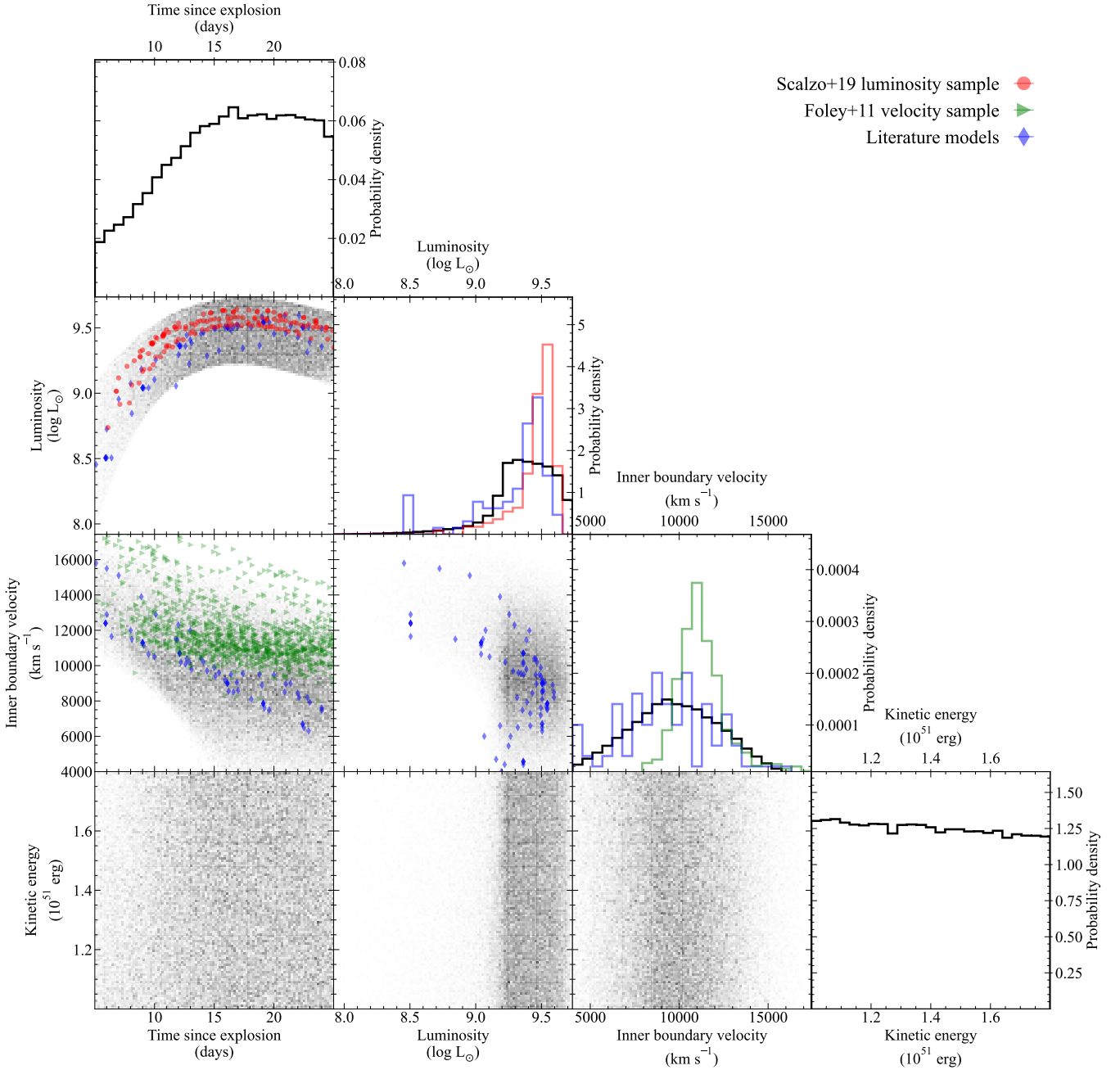
Figure 1. Distributions of the sampled input parameters used by our neural networks are shown in black. Bolometric luminosities for a sample of SNe Ia calculated by Scalzo et al. 2019 are shown in red. In green we show Si II $\lambda6\,355$ velocities calculated by Foley et al. 2011. Literature models discussed in the text are shown in blue.

over/under-luminous at early/late-times compared to observations of SNe Ia and hence are unphysical. Therefore, we adopt an empirical approach to determine the appropriate prior distribution for our input luminosity as a function of time since explosion. Scalzo et al. (2019) constructed bolometric light curves for a sample of low-redshift, well-observed SNe Ia observed by the Carnegie Supernova Project (CSP, Hamuy et al. 2006). Using these light curves we set an upper and lower limit on the bolometric luminosity as a function of time based on polynomial fits to the sample. Here we assume a typical rise time of 18.5 d (Ganeshalingam et al. 2011; Miller et al. 2020). We increase the upper limit on this luminosity range by a factor of 1.25, as TARDIS typically over-estimates the luminosity. Similarly, we

decrease the lower limit by a factor of 0.5 to allow our models to fit fainter SNe than those included by Scalzo et al. (2019). For each training model, the luminosity is then uniformly sampled within the appropriate boundaries for the previously selected time since explosion. The distribution of luminosities used for our training set is shown in Fig. 1, along with the observed SNe Ia luminosities calculated by Scalzo et al. (2019). Figure 1 also shows that our expanded sample region covers the range of luminosities used by a selection of literature models collected from the following sources: Stehle et al. (2005); Mazzali et al. (2008); Tanaka et al. (2011); Ashall et al. (2014); Mazzali et al. (2014); Heringer et al. (2017); Ashall et al. (2018); Heringer et al. (2019).
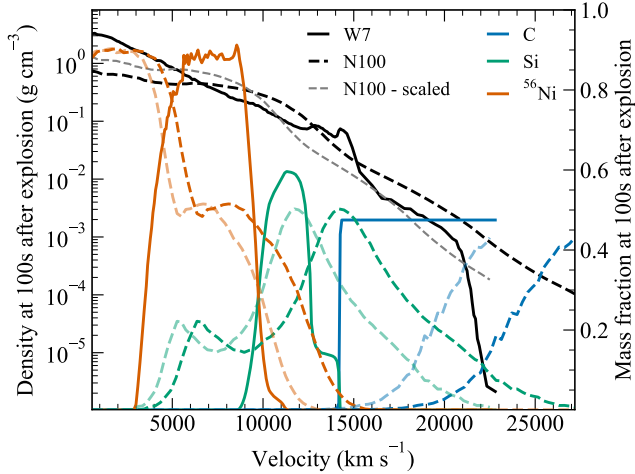
Figure 2. Density and representative mass fractions for the W7 and N100 models at 100s after explosion. A scaled N100 density and composition profile assuming $KE' = 1.0 \times 10^{51}$ erg (Eqns. 2 & 3) is shown as shaded lines.

For the inner boundary velocity we again adopt an empirical approach. We use the Si II velocities published by Foley et al. (2011) and Maguire et al. (2014) for samples of low-redshift SNe Ia to estimate an appropriate sampling range as a function of time. By analysing the velocity evolution of their sample, Foley et al. (2011) find that the Si II $\lambda 6355$ velocity v (in units of 10^3 km s^{-1}) at a time relative to maximum light t (in days) can be approximated with the following functional form:

$$v\,(v_0, t) = v_0\,(1 - 0.0322t) - 0.285t, \tag{1}$$

where v_0 represents the Si II $\lambda 6355$ velocity at maximum light. Using the previously selected time for each training model, and again assuming an 18.5 d rise time, we uniformly sample between the appropriate upper and lower boundaries at that time, which are defined assuming v_0 velocities of 15 000 km s^{-1} and 9 000 km s^{-1} respectively. Again, we increase the upper limit of our sample range by a factor of 1.25. The model photospheric velocity in radiative transfer simulations is highly correlated with, but not necessarily equal to, the Si II $\lambda 6355$ velocity (Fig. 1; c.f Parrent et al. 2012; Mazzali et al. 2014). To account for differences between the observed Si II $\lambda 6355$ velocity and the model photospheric velocity we subtract an additional offset uniformly sampled between $0 - 6\,000$ km s^{-1}. Figure 1 shows the velocities used for our training set along with photospheric velocities from a sample of literature models and Si II $\lambda 6355$ velocities for SNe Ia measured by Foley et al. (2011). As shown by Fig. 1, around maximum light the inner boundary velocity is typically lower than the observed Si II $\lambda 6355$ velocity.

Unlike the training models used in previous works (Chen et al. 2020; O'Brien et al. 2021), the density and composition of our training models does not follow an empirical approach and instead are more directly related to predictions from explosion models. For this initial, demonstrative study we limit our training datasets to compositions based on two well-studied explosion models: W7 (Nomoto et al. 1984) and N100 (Seitenzahl et al. 2013; Kromer et al. 2017). The W7 model is a one-dimensional, parameterised pure deflagration explosion with a deflagration speed such that the resulting composition and spectra closely match observations of SNe Ia around maximum light. The N100 model however is a deflagration-to-detonation transition explosion calculated via a fully realistic multi-dimensional simulation. Here we use the angle-averaged N100 model provided

by HESMA[2] (Kromer et al. 2017). In Fig. 2 we show density profiles for both the W7 and N100 models, along with mass fractions of C, Si, and ^{56}Ni. For each training set we assume the composition is fixed and taken directly from the respective explosion model. While both models do show good agreement with spectra of SNe Ia around maximum light, neither provides perfect agreement (Branch et al. 1985; Stehle et al. 2005; Sim et al. 2013; Mazzali et al. 2014). These limitations will also be reflected in our training datasets and we therefore do not expect that either model will provide perfect agreement with observations. Nevertheless, by taking compositions directly from existing explosion models we avoid scenarios where the best-fitting model determined via fitting contains an unrealistic structure, which may arise from fitting the abundance of each element independently.

Given that families of explosion models can generally produce similar structures with different kinetic energies (e.g. Seitenzahl et al. 2013), we also allow for some variation in the ejecta structure of our training models by scaling to higher or lower kinetic energies (Hachinger et al. 2009; Ashall et al. 2016). The velocity and density of each cell in the model are scaled to a new kinetic energy KE' and ejecta mass M'. The updated velocities v' and densities ρ' are given by:

$$v' = v_{exp} \left(\frac{KE'}{KE_{exp}} \right)^{\frac{1}{2}} \left(\frac{M'}{M_{exp}} \right)^{-\frac{1}{2}}, \tag{2}$$

$$\rho' = \rho_{exp} \left(\frac{KE'}{KE_{exp}} \right)^{-\frac{3}{2}} \left(\frac{M'}{M_{exp}} \right)^{\frac{5}{2}}, \tag{3}$$

where KE_{exp} and M_{exp} are the kinetic energies and ejecta masses of the input explosion models, in this case W7 and N100. For the current work, as we are focused only on Chandrasekhar mass explosions, we assume the ejecta mass is fixed and only allow for variation in the kinetic energy. We uniformly sample the kinetic energy between $10^{51} - 10^{51.26}$ erg, which covers the range of predictions from Chandrasekhar mass explosion models[3] (Nomoto et al. 1984; Sim et al. 2013).

In summary, we generate two complete training datasets based on the W7 and N100 explosion models. For both datasets we use the same input parameters, which are defined as time since explosion, luminosity, inner boundary velocity, and kinetic energy, and the same randomly selected values for each input parameter. Compared to the training sets presented by Chen et al. (2020) and O'Brien et al. (2021), our models cover a wider range of times and luminosities, making it possible to fit a sequence of spectra for a single SN Ia and determine the changes in the best fitting parameters. In addition to the ranges previously specified, we apply further criteria to ensure that our randomly selected parameter values represent physical and useful models. We set a minimum inner boundary velocity of 4 000 km s^{-1} and reject models where the inner boundary is <2 500 km s^{-1} below the maximum velocity of the ejecta. We randomly generate 120 000 models that meet these requirements for both our W7 and N100 training datasets. Of these, we use 100 000 to train the neural networks

and 20 000 test models to determine the typical accuracy of the neural networks (see Sect. A). The choice of appropriate prior space for our input parameters does require some manual selection and will significantly impact the determination of the overall best-fitting explosion model. We stress that we are only able to comment on which model provides a better match to the data within the boundaries set by our priors. In Sect. 7.1 we discuss the limitations of our training datasets and future improvements.

2.2 Pre-processing

The models developed for training as part of this work cover a wide range of luminosities within each wavelength bin and input parameters. Therefore, they require pre-processing before they can be used effectively by the neural network. Our first step is to rebin all of the model spectra. We chose 1 000 log-spaced wavelength bins in the range 3 000 – 9 000 Å. To allow our neural network to be used for fitting a wide range of SN Ia spectra, with different signal-to-noise ratios, we follow Chen et al. (2020) and apply smoothing to each spectrum using a Savitzky-Golay filter (order 2, window 25).

We also experimented with data augmentation to boost the size of our training dataset, as was done by Chen et al. (2020). For this augmented training data, each spectrum was included ten times with different levels of smoothing applied, resulting in ten slightly different spectra for a given set of input parameters. This was designed to approximately mimic different signal-to-noise ratios. Despite the increase in the size of the available training dataset, we found the overall performance of the neural networks decreased following data augmentation with this method. This likely arose from the different levels of smoothing resulting in varying degrees of blending for spectral features and noise within the training data inhibiting the ability of the neural networks to learn effectively. We therefore do not include additional data augmentation during training, but future work will explore alternative augmentation methods.

After rebinning and smoothing, the luminosity within each wavelength bin is processed by first taking $\log_{10}$ and transforming using the STANDARD scaler implemented in SKLEARN. Each of the input parameters used for generating our TARDIS spectra are processed in the same manner.

3 NEURAL NETWORKS

All neural networks were trained using TENSORFLOW (Abadi et al. 2015) and KERAS (Chollet et al. 2015). Extensive hyperparameter testing and tuning was performed both manually and with OPTUNA (Akiba et al. 2019). For both the W7 and N100 models, we train a final set of 20 neural networks each that are discussed throughout this work. Figure 3 shows an example of the neural network architecture used. Table A1 in the Appendix gives the hyperparameters of each neural network, along with the median mean fractional error and median maximum fractional error across the 20 000 W7 and N100 test models.

The input layer of each neural network is defined by four neurons, corresponding to the four (processed) parameters used to generate each TARDIS spectrum (time since explosion, luminosity, inner boundary velocity, and kinetic energy). All neural networks are trained on the same set of 100 000 input parameters with the same validation split of 20%. The number of subsequent layers and neurons per layer are given in Table A1. We do not include batch normalisation after any layer, as this was found to decrease the overall performance of the neural networks. The output layer is defined by 1 000 neurons and

corresponds to the (processed) flux in each of the 1 000 wavelength bins. In general, we found that the choice of optimiser had little impact on the performance and therefore chose the Nadam optimiser for all neural networks. In addition, we found the leaky-ReLU (Maas et al. 2013) activation function and mean squared error loss function provided the best performance and therefore were used for all neural networks during training.

Each neural network was trained for a total of 50 000 epochs on three Nvidia Quadro RTX 6 000 GPUs. To prevent over-fitting, we monitor the performance of the neural networks on the validation data and save the epoch with the best validation performance. In Sect. A in the Appendix, we discuss the accuracy of our neural networks in more detail, but we find typical accuracies of up to a few percent (Table A1).

4 SPECTRAL SEQUENCE FITTING METHOD

In the following section, we discuss our approach to fitting multiple spectra of SNe Ia simultaneously. Using the neural networks described in Sect. 3, we aim to determine the best-fitting model input parameters for each observed spectrum. In other words, the set of input parameters θ with the highest likelihood given the observations o, $\mathcal{L}(\theta|o)$. We also aim to quantitatively determine the relative likelihood of the W7 and N100 models for a given SN Ia by comparing the evidence $\mathcal{Z}$ for each model (Thrane & Talbot 2020) assuming our goodness-of-fit metric and priors. In Bayesian inference, the evidence is given by the likelihood function marginalised over the prior distribution $\pi(\theta)$:

$$\mathcal{Z} = \int \mathcal{L}(\theta|o)\pi(\theta)d\theta, \tag{4}$$

which is calculated for both the W7 and N100 models based on the total likelihood across all spectra included in the fit.

The prior distributions, $\pi(\theta)$, for our input parameters used during fitting are the same as those in Sect. 2.1. As discussed in Sect. 2.1, the input parameters used to generate our TARDIS spectra are given by the time since explosion, luminosity (L), inner boundary velocity (v), and the density and composition of the ejecta. For the current work, the density and composition are taken from either the W7 or N100 model, with some variation given by the kinetic energy (KE). We note that as we do allow for variation in the kinetic energy, the emulated models are not strictly the exact predictions from either W7 or N100. We therefore refer to the resulting best-fit models as either W7- or N100-like. The explosion epoch, t_{exp}, and KE are both fixed for all spectra in the sequence. This results in a set of $2 + 2N$ parameters defining the spectral sequence, $\theta = \{t_{exp}, KE, L_1, v_1, ..., L_N, v_N\}$, where N is the number of spectra included in the fit. To reduce degeneracy between the free parameters, we add an additional constraint to ensure that the inner boundary velocity decreases for later spectra, $v_i \geqslant v_{i+1}$, which is typically observed in SNe Ia (Foley et al. 2011; Maguire et al. 2014) and used in spectral modelling (Stehle et al. 2005). In Sect. 7.1, we discuss the impact of this assumption.

Fits are performed with ULTRANEST (Buchner 2021). The benefit of nested sampling routines such as ULTRANEST is that they can be used to simultaneously infer posterior distributions and calculate the model evidence. For each set of input parameters θ, we use a subset of the neural networks described in Sect. 3 to generate synthetic spectra covering the spectral sequence. We assume a Gaussian likelihood
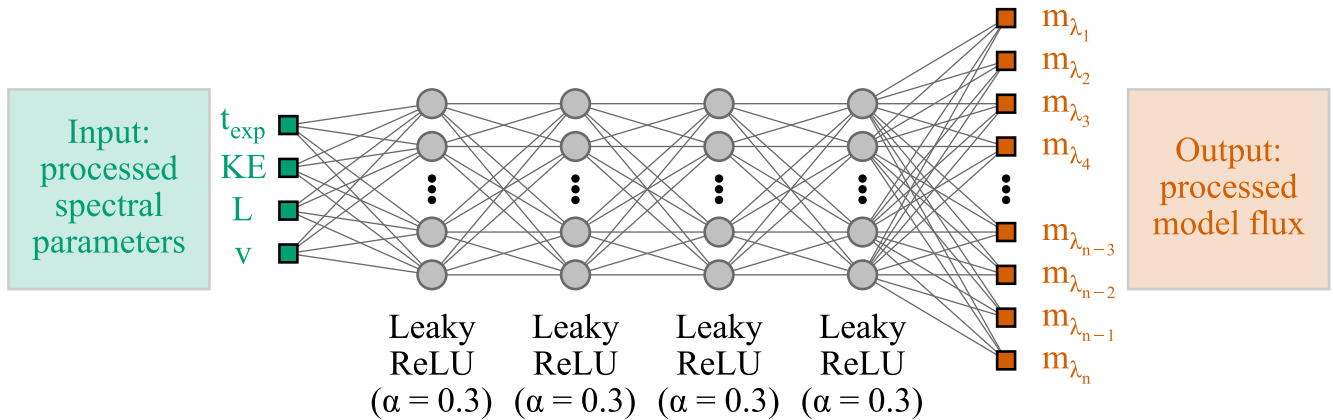
Figure 3. Schematic diagram showing an example neural network architecture with four hidden layers. As input the neural network takes the processed spectral parameters defining our TARDIS models, which are the time since explosion (t_{exp}), the kinetic energy of the ejecta (KE), the luminosity of the spectrum (L), and the inner boundary velocity (v). Each neuron uses a leaky-ReLU activation function in which the gradient for negative values is set to $\alpha = 0.3$. The outputs of the neural network are the processed flux values within each wavelength bin, m_{λ_n}.

function for each input spectrum i given by,

$$\ln \mathcal{L}_i(\theta|o) \propto -\frac{1}{2} \sum_\lambda w_\lambda \left[\left(\frac{m_\lambda(\theta) - o_\lambda}{s_\lambda(\theta)} \right)^2 + \ln(2\pi s_\lambda(\theta)^2) \right], \quad (5)$$

where

$$s_\lambda(\theta)^2 = \sigma_{o,\lambda}^2 + f^2 m_\lambda(\theta)^2 + \Delta_\lambda(t,v)^2 m_\lambda(\theta)^2. \quad (6)$$

Here, o_λ and $\sigma_{o,\lambda}$ give the observed flux and uncertainty, $m_\lambda(\theta)$ gives the model flux for a given set of input parameters θ, $\Delta_\lambda(t,v)$ gives the fractional uncertainty of the prediction for the input time since explosion t and inner boundary velocity v, f is a nuisance parameter included to account for underestimated uncertainties, and w is a weighting parameter. The integration is performed over all wavelengths λ. In cases where the observed flux uncertainty is unavailable we assume an uncertainty of 2%. Sect. A discusses the uncertainty of our neural network predictions ($\Delta_\lambda(t,v)$) in more detail, however we note that this is typically of order a few percent. The parameter f is included as an additional source of uncertainty to account for systematic differences between off-the-shelf explosion model predictions and individual SNe Ia. We note that these systematic differences between explosion model spectra and observed SNe Ia are typically $\gtrsim 20\%$ (Kerzendorf et al. 2021). Therefore, without this additional parameter, the posteriors give unreasonably well-constrained parameters. This arises from the fact that none of the models constitute a 'good' fit based on the χ^2 value and therefore, small deviations away from the best-fitting model result in large changes in the χ^2 value and only a narrow range of acceptable parameters. The parameter w_λ allows for different relative weighting of individual features or wavelengths. By default, we make the simplest assumption of no weighting applied during fitting. For the initial work presented here, we also explore the impact of excluding different wavelengths during fitting in Sect. 6. We note that riddler allows for the inclusion of any arbitrary weighting scheme implemented by the user.

Fits are performed independently with each of the top six neural networks as determined by their mean and maximum fractional errors (see Table A1 in the Appendix). We combine their posteriors

determined by ULTRANEST to estimate the overall best-fitting parameters. This approach helps to overcome limitations associated with the accuracy of any given neural network and provides a more reliable estimate for the total uncertainty of the model parameters, including systematic uncertainties associated with the different neural network architectures. For the W7 model, the best performing neural networks used throughout the rest of this work are NNs 2, 3, 4, 16, 17, & 18. For the N100 model, these are NNs 8, 13, 14, 16, 17, & 18. To compare the relative likelihoods of two models (in this case either W7- or N100-like ejecta structures) we use the evidences calculated by ULTRANEST and the Bayes factor given by

$$\ln \text{BF} = \ln \mathcal{Z}_1 - \ln \mathcal{Z}_2, \quad (7)$$

where $\mathcal{Z}_{1,2}$ refer to the evidence for either W7- or N100-like models. In this case, a Bayes factor >1 indicates a preference for model 1 relative to model 2.

5 APPLICATION

Having described the neural networks designed to emulate TARDIS simulations and our assumed likelihood function, we now apply riddler to fitting spectra. We begin by applying riddler to a model spectrum, for which the true input parameters are known, and then to observed SNe Ia.

5.1 Test models

Here we apply riddler to fitting model spectra generated using the same approach discussed in Sect. 2.1 (but not seen during training) and demonstrate that it is able to accurately recover the parameters of the input model. During fitting, we assume a 2% flux uncertainty to mimic real data and account for noise in the radiative transfer simulation. In Fig. 4 we show an example fit applied to one of our W7 model spectra. We find comparable results when fitting our N100 spectra.
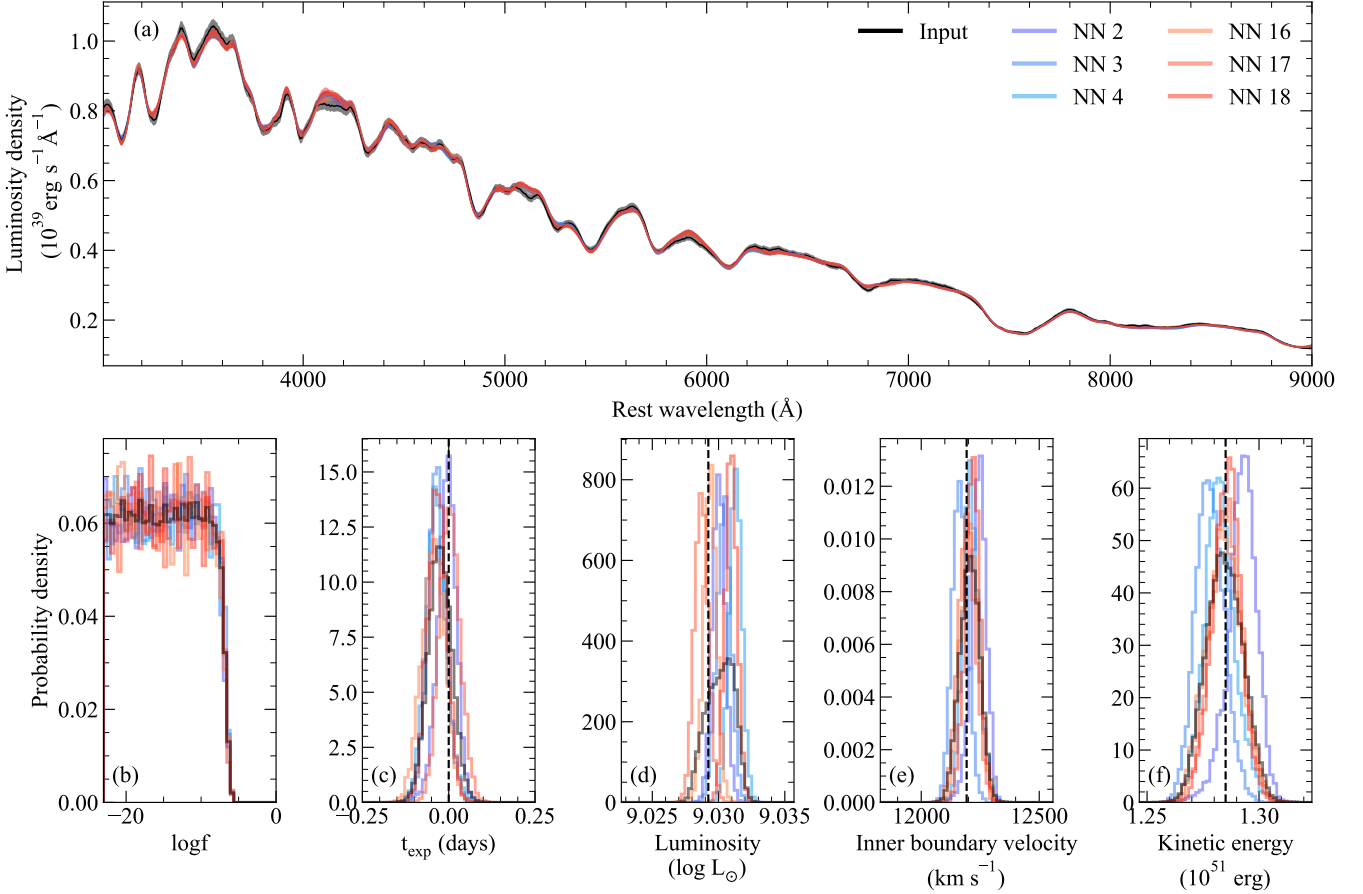
Figure 4. *Panel a:* Comparison between our input spectrum (black) and the neural network reconstructions of the best-fitting parameters based on our ULTRANEST fits. The grey shaded region shows the assumed 2% uncertainty on the model flux, while coloured shaded regions show the model flux error, Δ_λ. *Panels b − f:* Posterior distributions of fitting parameters determined by ULTRANEST. The posterior distribution for each neural network is shown as a coloured line, while the total posterior across all neural networks is shown as a grey line. In *Panels c − f* we also denote the true input value as a vertical dahsed line.

Figure 4(a) shows our input model spectrum compared against neural network reconstructions of the best-fitting parameters. The mean fractional errors of these spectra are ∼1.3%. As shown by Fig. 4 our best-fitting spectra reproduce the strongest spectral features in the input spectrum with the correct velocity, strength, and width. For weaker spectral features however, such as those at λ ∼4 500 Å and ∼5 150 Å, our best-fitting spectra struggle to reproduce the input model. We note that the inability to reproduce weak features is a systematic issue with our neural networks (Sect. A) and could result in absent or blended features in the best-fitting spectra. Our assumed 2% flux uncertainty likely exacerbates the inability to fit weak features further as they can become lost in the spectrum uncertainty, but is nevertheless a better representation of fitting real data. Overall, our best-fitting models also reproduce the flux level and colour of the input spectrum, but there are some systematic differences. Around ∼3 400 − 3 600 Å the neural networks predict systematically lower flux than in the input model, while around ∼4 000 − 4 100 Å they predict systematically higher flux. In both cases however the neural network predictions are within our assumed 2% uncertainty for the input model spectrum.

Figure 4(b) shows that the posterior distribution of log f is consistent with essentially no additional systematic uncertainty (≲ 0.5%), which is unsurprising given that the model was constructed in the same manner as the training dataset and with the same underlying ejecta structure. In Figs. 4(c) − (f) we present posteriors for the best-fitting parameters of the input spectrum compared to the true value. While individual neural networks do show some variation in their posterior distributions, they all generally produce results that are consistent with each other and with the true input value. We note that in some cases the posterior means may be systematically offset from the input value. We therefore take the upper and lower limits of the full posterior across all neural networks as a conservative estimate of the total uncertainty. These limits are typically ≲ ±5% of the input value and include the input value itself.

5.2 Observations

Having shown that `riddler` is able to accurately recover the input parameters of our TARDIS model spectra, we now apply it to observations and estimate the best-fitting parameters for observed SNe Ia. For this purpose, we use the well-observed SN 2011fe (Nugent et al. 2011), which was the subject of previous, detailed spectroscopic modelling by Mazzali et al. (2014) and Heringer et al. (2017), and SN 2013dy (Zheng et al. 2013). All spectra were obtained from WISeREP (Yaron & Gal-Yam 2012). Both SNe are fit independently using our W7 and N100 models. No wavelength-dependent weighting (see Sect. 6) or additional flux scaling has been applied to these fits. Details of the spectra included during fitting are given in Table 1,

Table 1. Best-fitting parameters, likelihoods, and evidences for SN 2011fe and SN 2013dy with either W7 or N100 models assuming uniform wavelength weighting

Date	MJD	Phase (days)	Luminosity (log $L_\odot$)	Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$	Luminosity (log $L_\odot$)	Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$
			W7			N100		
			SN 2011fe					
2011 Aug. 28	5 5801.17	−13.1	$8.56^{8.58}_{8.54}$	12221^{12598}_{11623}	-88349^{-88330}_{-88357}	$8.56^{8.58}_{8.55}$	13987^{14348}_{13583}	-88263^{-88252}_{-88273}
2011 Aug. 31	5 5804.25	−10.1	$9.11^{9.12}_{9.09}$	12181^{12589}_{11591}	-89153^{-89148}_{-89158}	$9.10^{9.11}_{9.08}$	12355^{13013}_{11828}	-89197^{-89194}_{-89200}
2011 Sep. 03	5 5807.38	−6.9	$9.40^{9.42}_{9.38}$	6740^{6956}_{6671}	-89730^{-89727}_{-89735}	$9.39^{9.41}_{9.38}$	11267^{11995}_{10829}	-89785^{-89783}_{-89788}
2011 Sep. 07	5 5811.37	−2.9	$9.53^{9.54}_{9.51}$	5730^{6185}_{5330}	-90022^{-90020}_{-90024}	$9.52^{9.54}_{9.51}$	7225^{7738}_{6837}	-90248^{-90239}_{-90253}
2011 Sep. 10	5 5814.39	+0.1	$9.54^{9.56}_{9.52}$	4840^{5320}_{4350}	-90225^{-90222}_{-90227}	$9.52^{9.54}_{9.50}$	4035^{4340}_{4000}	-90288^{-90280}_{-90295}
2011 Sep. 13	5 5817.67	+3.4	$9.50^{9.53}_{9.48}$	4033^{4338}_{4000}	-90520^{-90514}_{-90525}	$9.49^{9.51}_{9.47}$	4011^{4249}_{4000}	-90526^{-90517}_{-90538}
			$\ln \mathcal{Z}$		$-538059^{-538032}_{-538082}$	$\ln \mathcal{Z}$		$-538366^{-538351}_{-538377}$
			SN 2013dy					
2013 Jul. 21	5 6494.48	−6.6	$9.42^{9.44}_{9.40}$	8956^{9490}_{8404}	-89846^{-89835}_{-89853}	$9.41^{9.45}_{9.39}$	11167^{13222}_{10325}	-89850^{-89833}_{-89868}
2013 Jul. 25	5 6498.60	−2.5	$9.50^{9.53}_{9.48}$	5307^{5519}_{5009}	-89943^{-89936}_{-89948}	$9.48^{9.51}_{9.46}$	10422^{12155}_{9206}	-90287^{-90254}_{-90330}
2013 Jul. 27	5 6500.32	−0.8	$9.51^{9.54}_{9.48}$	4465^{4613}_{4257}	-89945^{-89939}_{-89954}	$9.48^{9.51}_{9.47}$	9961^{11743}_{8510}	-90451^{-90432}_{-90474}
2013 Jul. 29	5 6502.31	+1.2	$9.48^{9.50}_{9.46}$	4003^{4066}_{4000}	-90012^{-90001}_{-90019}	$9.47^{9.50}_{9.45}$	9293^{11346}_{7572}	-90559^{-90530}_{-90569}
2013 Aug. 01	5 6505.57	+4.5	$9.41^{9.43}_{9.39}$	4001^{4026}_{4000}	-90241^{-90232}_{-90250}	$9.42^{9.44}_{9.40}$	5363^{6417}_{4000}	-90564^{-90462}_{-90596}
			$\ln \mathcal{Z}$		$-450039^{-450024}_{-450050}$	$\ln \mathcal{Z}$		$-451751^{-451739}_{-451762}$

Note. The best-fitting model parameters are determined based on the median of the total posterior across all of the neural networks included in the fits. Upper and lower limits are also based on the total posterior across all neural networks. We stress that these parameters assume uniform wavelength weighting, which has important limitations (see Sect. 6). Likelihoods and evidences are given by the mean with upper and lower limits determined from the minimum and maximum values, across all neural networks,.

along with best-fitting values, and upper and lower limits for each input parameter. In addition, Table 1 also gives the mean, minimum, and maximum evidences (Eqn. 4) for a given model based on the total likelihood across all spectra included in the fit and likelihoods (Eqn. 5) for a given model and individual spectra, as determined from our ULTRANEST fits.

We again note that as we are using predictions from explosion models, and do not allow the structure or composition of our model ejecta to vary freely, we do not expect to find perfect agreement with the observations. Our approach with riddler however is able to find the best-fitting set of input parameters, including explosion epoch, for a given explosion model template assuming our likelihood function and prior distributions. We focus on quantifying the relative agreement between different models, based on the likelihoods and evidences.
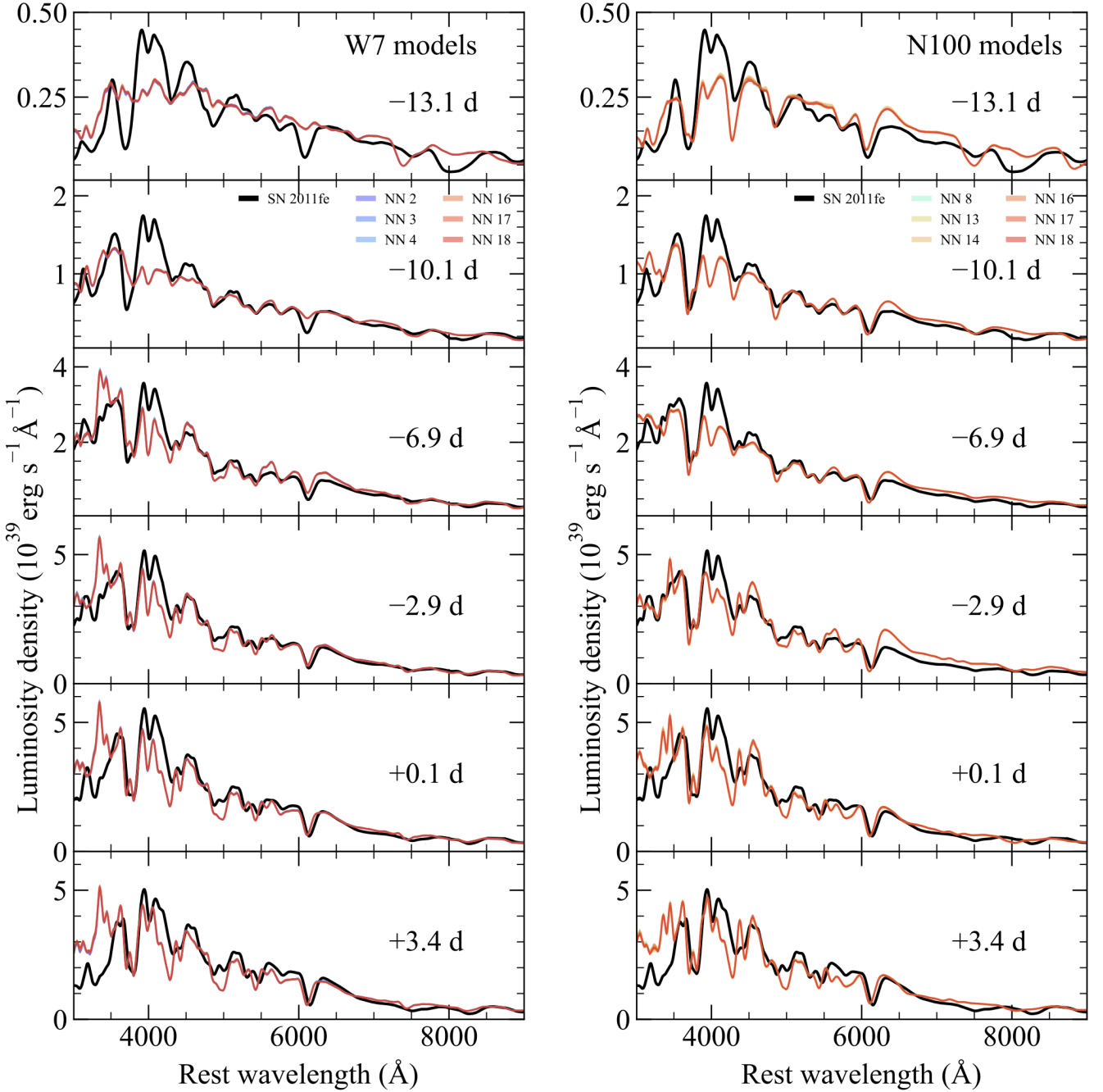
5.2.1 SN 2011fe

For SN 2011fe, we use the HST spectra presented by Mazzali et al. (2014) and assume a distance modulus of $\mu = 29.04$ mag and total extinction of $E(B - V) = 0.02$ mag (Mazzali et al. 2014). Figure 5 shows the spectra reconstructed by our neural networks for the W7 and N100 models and their respective best-fitting parameters. The posterior distributions from each fit are shown in Fig. 6.

From Fig. 5 it is clear that the earliest SN 2011fe spectrum is not reproduced by a W7-like ejecta as the models predict features that

are significantly weaker than those observed. At this phase models with an N100-like ejecta produce features that more closely resemble SN 2011fe, although the overall spectral shapes show some differences and are redder than SN 2011fe. Fits with our N100 models favour kinetic energies towards the lower boundary of our input parameters ($\sim 1.0 \times 10^{51}$ erg), which represents a $\sim$30% decrease in the kinetic energy and results in scaled density profiles similar to that of W7 (Fig. 2). Given that the best-fitting models approach the lower boundary, this would indicate that our input parameter space is not sampling the parameter range necessary to fully reproduce SN 2011fe, however further decreasing the kinetic energy may be unphysical. At this phase the ejecta above the photosphere in the W7-like models is dominated by unburned carbon and oxygen, while the N100-like models have more extended distributions of iron-group and intermediate-mass elements and hence show stronger spectral features. At −10.1 d we again find that the N100-like models produce better agreement with SN 2011fe for some spectral features compared to the W7-like models. In particular, the N100-like models show stronger Si II $\lambda 6\,355$ and Ca II H&K that are more consistent with SN 2011fe, but over-predict the strength of the Si II and Fe III blend around $\sim$4 800 Å and near-ultraviolet (NUV) flux. Despite showing visually worse agreement with SN 2011fe spectral features, the W7-like models are favoured at this phase based on the likelihood, which is primarily driven by improved agreement with the continuum at longer wavelengths and the need for a higher systematic uncertainty (f) for the N100-like models. In addition, while the

Figure 5. Comparison between SN 2011fe and neural network reconstructions of the best-fitting W7 (*left*) and N100 (*right*) model spectra for each neural network assuming uniform wavelength weighting. Spectra are shown on an absolute luminosity scale with no additional scaling or offsets applied. Phases are given relative to *B*-band maximum.

pseudo-equivalent widths of features for the N100-like models show better agreement with SN 2011fe, the lower overall flux around these wavelengths (relative to the observations) can lead to lower likelihoods. Conversely, the weaker features in the W7-like models can lead to higher likelihoods. This would indicate that additional weighting, potentially based on the overall strengths of the features, could be considered to avoid situations where spectra that do not produce features are favoured quantitatively (Ogawa et al. 2023). By −6.9 d W7-like models begin to show improved agreement with SN 2011fe and are generally able to match the Si II $\lambda6\,355$, Si II $\lambda5\,972$, and

S II-W features, and the complex silicon/iron blend around ~5 000 Å. Again however we find that the N100-like models show better fits to these features, but the W7-like models have higher likelihoods due to their better agreement with the continuum at $\lambda \gtrsim 6\,500$ Å.

Around maximum light, the differences between our best-fitting W7-like and N100-like models become less pronounced. Both ejecta structures generally reproduce many of the spectral features, but systematically over-estimate the NUV flux and do not reproduce NUV spectral features. Our W7-like models also produce reasonable agreement with the Ca II NIR triplet, while our N100-like models are
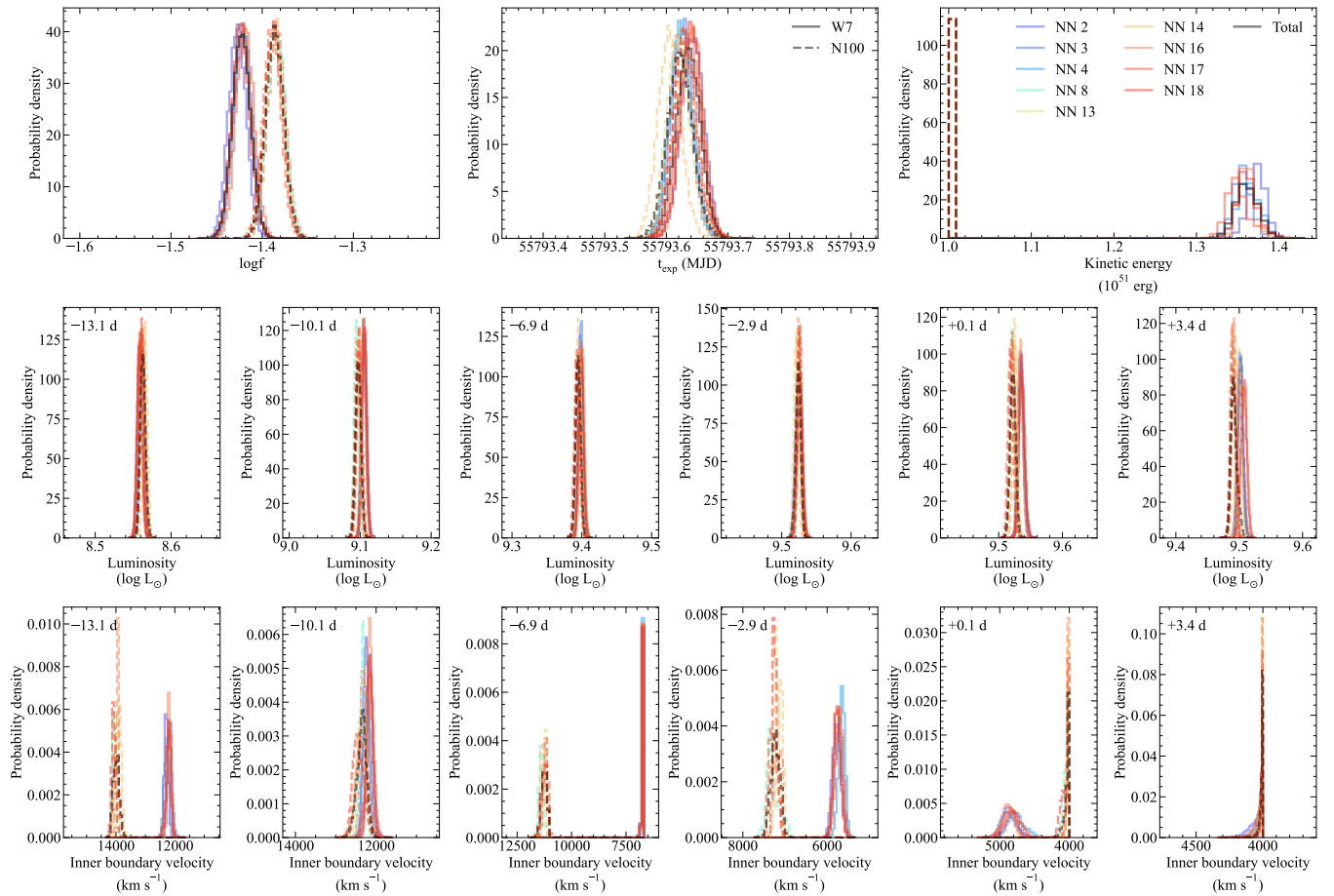
Figure 6. Posterior distributions for best-fitting parameters to SN 2011fe using W7 and N100 models for each neural network assuming uniform wavelength weighting. Phases are given relative to *B*-band maximum.

not able to match this feature. As with earlier epochs, the W7-like models are better able to match the flux at longer wavelengths and therefore produce higher likelihoods. We note however that for the spectrum at +3.4 d, both models show overlapping likelihood ranges across all of the neural networks used here, indicating no strong preference for either model at this epoch.

Figure 6 shows that the W7-like and N100-like models have similar levels of systematic differences (parameterised by $\log f$, ~23%) between the models and observed spectra, although N100-like models are typically higher. Despite differences in the overall level of agreement, both sets of models predict a consistent explosion epoch of MJD = 5 5793.6±0.1 and consistent luminosities for each spectrum. As previously mentioned, our N100-like model fits favour a density profile scaled to a lower kinetic energy, which results in similar profiles to that of W7 (Fig. 2). Therefore while the composition of the ejecta differs between the W7- and N100-like models, our fits for SN 2011fe argue for a consistent density profile.

Unlike most other model parameters, we find significant differences in the location of the inner boundary velocity between our W7- and N100-like models. The location of the inner boundary velocity is highly model-dependent as this sets the amount and composition of material above the photosphere, which determines the overall spectrum. In addition, the inner boundary and composition, in combination with the time since explosion and luminosity, will also strongly impact the overall temperature of the model. It is therefore unsurprising that we generally find different inner boundary veloci-

ties for the W7- and N100-like models. Around maximum light our fits using both W7- and N100-like models find inner boundary velocities that are likely too low and indeed lie towards the lower limit of our training dataset, 4 000 km s^{-1}. Such velocities are significantly lower (by up to ~3 500 km s^{-1}) than previous custom fitting of SN 2011fe (see Sect. 7.2; Mazzali et al. 2014). While this could point to a limitation in the construction of our training dataset, lower inner boundary velocities are instead likely a consequence of fitting the entire spectrum, with no weighting for features, and the photospheric approximation made by TARDIS. The result is that the fits will systematically favour low and potentially unrealistic inner boundary velocities, which produce spectra with higher temperatures that better match the overall shape of the observed spectra. This point is discussed further in Sect. 6.

Overall, we find that riddler is able to produce reasonable fits to SN 2011fe using both W7- and N100-like ejecta structures. From ULTRANEST, we find the evidence for our W7-like models is $\ln \mathcal{Z} \sim -538059$, while for our N100-like models $\ln \mathcal{Z} \sim -538366$. Calculating the Bayes factor (Eqn. 7), we find $\ln \mathrm{BF} = 307$. Kaas & Raftery (1995) argue that a Bayes factor of $2 \ln \mathrm{BF} > 10$ indicates a very strong preference for a given model. Therefore our fitting indicates that a W7-like ejecta structure is heavily favoured, however we stress that this is dependent on our likelihood function, which should be treated cautiously when fitting full spectra, and our assumed priors, which do not cover the full parameter space required for N100-like models (Fig. 6). Considering individual spectra and

their likelihoods however, the somewhat more mixed structure of N100 (compared to the highly stratified W7) is favoured at very early phases shortly after explosion.

5.2.2 *SN 2013dy*

We fit the HST spectra of SN 2013dy presented by Pan et al. (2015). We assume a distance modulus of $\mu = 30.68$ mag and total extinction of $E(B-V) = 0.35$ mag (Pan et al. 2015). The best-fitting W7 and N100 spectra are shown in Fig. D1 & D3, with posterior distributions shown in Fig. D2 & D4, in the Appendix.

In general, we find that SN 2013dy shows visually somewhat worse agreement with both W7- and N100-like models compared to SN 2011fe. At -6.6 d both sets of models have similar likelihood ranges and show stronger spectral features and bluer spectra than SN 2013dy. As with SN 2011fe, we find our N100-like models favour lower kinetic energies, but we also find similarly low kinetic energies for our W7-like models. Given that N100 scaled to 1.0×10^{51} erg produces a similar density structure to W7 (Fig. 2), this would again imply that our N100 training dataset does not extend to sufficiently low energy to reproduce the overall density structure observed. Towards maximum light, both models show improved agreement with most spectral features, including Si II $\lambda 6355$ and Ca II H&K, but over-predict the NUV flux, as in the case of SN 2011fe.

With the exception of the first epoch, we find that W7-like models produce lower likelihoods compared to N100-like models. Again this is due to the better agreement between the model spectra and the continuum at longer wavelengths. At -6.6 d, our neural networks show overlapping likelihoods for both models, indicating no strong preference for either. We also note that each of our N100 neural networks show considerably larger variations in their predictions than our W7 or SN 2011fe fits. This likely arises from the overall poor quality of the fits.

From Fig. D4 we find that N100-like models have significantly larger systematic differences relative to SN 2013dy than W7-like models (log $f \sim 30\%$ compared to $\sim23\%$), which is consistent with the generally worse quality of fits. As evident from the range of best-fitting spectra in Fig. D3, our N100-like models can show large variations in the best-fitting parameters. Across all neural networks we find much broader ranges of best-fitting values than for SN 2011fe, with NN 18 also showing a bi-modal distribution for the explosion epoch, kinetic energy, and inner boundary velocities. As was the case for SN 2011fe, some fits find inner boundary velocities that are too low and close to the lower limit of our training data, $4\,000$ km s^{-1}. This indicates that either our training dataset does not cover the full required range, or more likely the uniform weighting scheme is not appropriate at these phases.

Overall, we find that `riddler` produces reasonable fits to SN 2013dy using W7-like models, but generally fails to find reasonable fits with an N100-like structure. Comparing the evidences calculated by ULTRANEST, we find ln BF = 1 712 – indicating overwhelming evidence in favour of the W7 model across the full spectral sequence. Again we stress that the Bayes factor is useful only to determine the relative preference for a given model, based on our assumed priors and likelihood. It does not indicate that W7 is the correct model for a given SN and indeed our fitting indicates a wider parameter space is necessary for the N100 models but it is not clear if such a prior distribution would be physically realistic.

5.3 Summary

In summary, we applied `riddler` to fitting model spectra that were not seen by the neural networks during training. We showed that we are able to recover the input parameters of the models to within a few percent ($\lesssim 5\%$). Applying `riddler` to the well-observed SNe Ia, SN 2011fe and SN 2013dy, we find models containing a W7-like ejecta structure are generally preferred and able to reproduce observations within approximately one week of maximum light. For SN 2011fe, we find the earliest spectra (more than approximately 10 d before maximum light) are better fit by models with an N100-like ejecta structure, while for SN 2013dy we find that N100-like ejecta structures produce significantly worse fits at all epochs. Based on the Bayes factor, we find strong evidence in favour of a W7-like model compared to an N100-like model when considering the full spectral sequence for both SN 2011fe and SN 2013dy. Again we stress that these conclusions are based on our assumed likelihood function and priors, and different models may be affected by the prior distributions in different ways.

6 IMPACT OF ALTERNATIVE WEIGHTING SCHEMES

In Sect. 5, we applied `riddler` to fit spectra of SNe Ia from 3 000 – 9 000 Å. During our fitting process, we made the simplest assumption of treating all wavelengths uniformly with no additional weighting or flux scaling applied. The photospheric approximation used by TARDIS however means that certain wavelengths are expected to have larger systematic offsets than others and therefore should not necessarily be treated equally.

TARDIS assumes a sharp boundary separating optically thick and thin regions. This allows TARDIS to model SNe spectra quickly by avoiding simulating regions of the ejecta with high optical depths. Due to this approximation, all Monte Carlo packets are injected into the simulation at the same physical position in the ejecta, regardless of their wavelengths. The optical depth at longer wavelengths however is generally lower than at shorter wavelengths, therefore the photosphere at longer wavelengths generally should be deeper inside the ejecta. By injecting all packets at the same location, those with longer wavelengths are typically able to escape the ejecta more easily than if they were injected deeper inside the ejecta. This usually manifests as an increased flux at longer wavelengths and is commonly observed closer to and beyond maximum light as the photospheric approximation becomes increasingly questionable (e.g. Stehle et al. 2005; Mazzali et al. 2014). Therefore, it is reasonable to allow for a systematically higher flux at longer wavelengths during our fitting process.

To demonstrate the impact of this, we perform additional fits in which wavelengths longer than 6 500 Å were excluded, which we call 'blue only' weighting. We note that the choice of $\lambda > 6\,500$ Å is somewhat arbitrary. The flux excess induced by the photospheric approximation becomes more pronounced at longer wavelengths, but does not necessarily begin at 6 500 Å. This value was chosen simply to demonstrate its impact and due to the presence of few spectral features. We also test a third weighting scheme, 'features only' weighting, in which only specific spectral features are included. This is designed to broadly mimic the approach taken by Ogawa et al. (2023), which was partially motivated by presence of weak-spectral features that are typically not well-fit by overall likelihood estimates (as is the case for `riddler`). We note however that we do not include additional comparisons between the model and data calculated by Ogawa et al. (2023), such as equivalent widths and velocity minima,
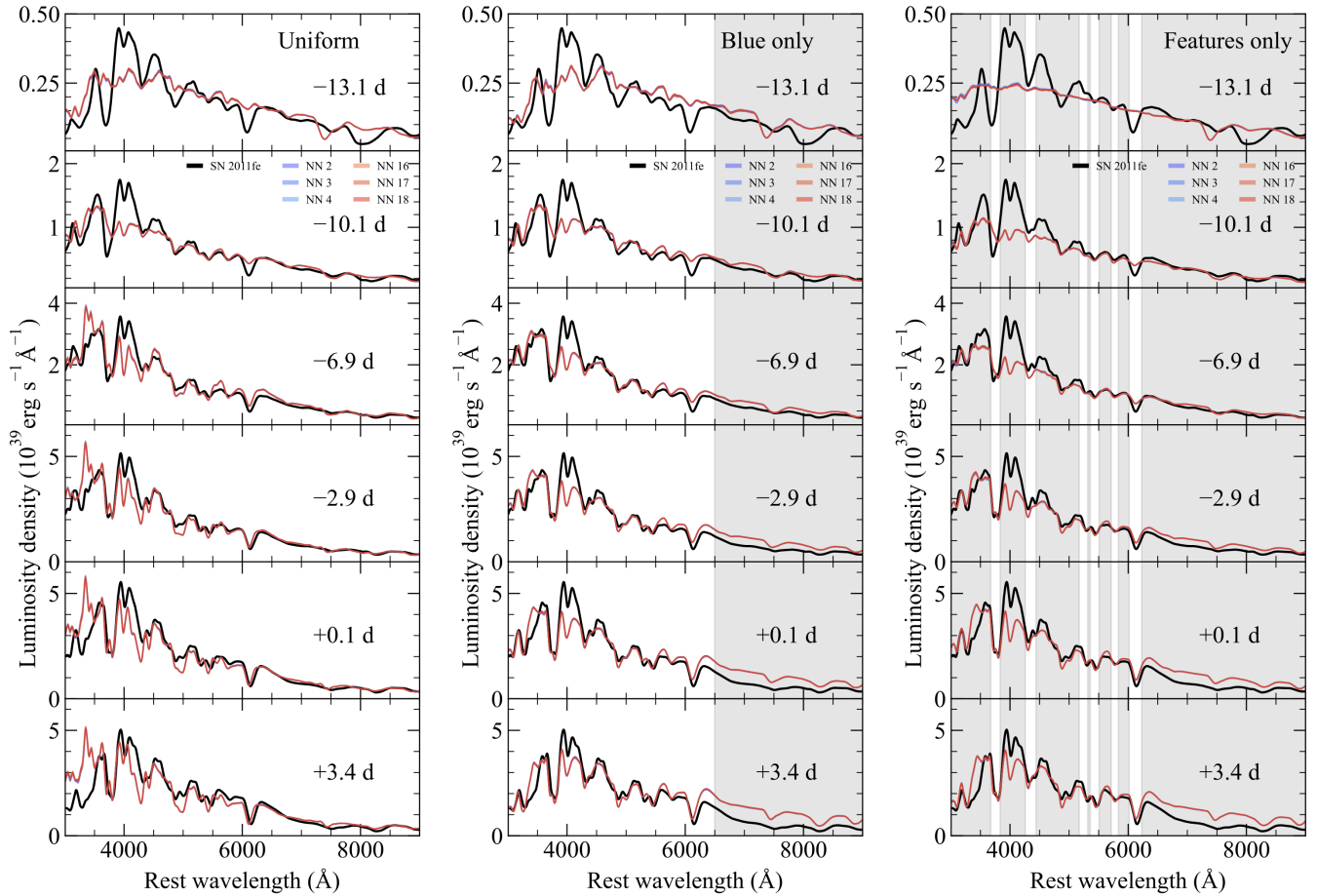
Figure 7. Comparison between SN 2011fe and neural network reconstructions of the best-fitting W7 model spectra for each neural network and assuming different weighting schemes. Spectra are shown on an absolute luminosity scale with no additional scaling or offsets applied. Phases are given relative to *B*-band maximum. Grey shaded regions are not included during fitting.

which could yield better agreement in some cases (see Sect. 5). For these tests and demonstrative purposes, the different schemes are applied at all epochs, but we again stress that `riddler` allows for any arbitrary weighting scheme, including those that vary with both wavelength and phase.

Figure 7 shows our best-fitting W7-like spectra compared to SN 2011fe assuming each of our three weighting schemes. Shaded regions are not included during the fits. The resulting posterior distributions are given in Fig. 8, while best-fitting parameter values are given in Table B1 in the Appendix. Comparisons for each weighting scheme against N100-like models are shown in the Appendix in Fig. C1, with posterior distributions shown in Fig. C2 and best-fitting parameters given in Table B2. Similar figures for SN 2013dy are also given in Appendix D, again with best-fitting values in Tables B1 & B2.

From Fig. 7, it is clear that alternative weighting schemes, including only fitting specific spectral features, do not find improved agreement between SN 2011fe and W7-like models for the earliest spectra – the outer ejecta of W7 simply does not contain the required structure to reproduce the spectral features observed. Beginning at −6.9 d however we find model fits that do not include wavelengths $\lambda > 6\,500$ Å provide significantly better fits to spectral features, including the iron blends around ∼4 500 Å and ∼5 000 Å, and the NUV. Improved agreement with the NUV is a direct result of exclud-

ing longer wavelengths from the fit and therefore reducing the impact of the photospheric approximation. As longer wavelengths typically show an excess of flux, the best-fitting models will preferentially be those with higher temperatures, which naturally shifts more flux from longer to shorter wavelengths. This will better match the former at the expense of the latter, resulting in an excess of flux in the NUV. This is further demonstrated by the fact that Fig. 8 shows these models also have systematically higher inner boundary velocities and therefore lower temperatures. For SN 2011fe, we find similar results for our feature weighting scheme (Fig. 7). Fits with N100-like models also show similar levels of improved agreement compared to SN 2011fe when using our alternative weighting schemes (Fig. C1).

Considering the likelihood values, we find some changes in the best-fitting models depending on the weighting scheme. With uniform weighting, all spectra after −10.1 d show higher likelihoods for W7-like models. Our blue only and features only weighting schemes however show slightly higher likelihoods for N100-like models between −10.1 − −2.9 d. Despite the majority of spectra showing higher likelihoods for N100-like models, ULTRANEST estimates an overall higher evidence in favour of W7-like models, which is primarily driven by the much higher likelihood for the +3.4 d spectrum. We note however that while the Bayes factor does still indicate a strong preference for the W7 model relative to N100, this is significantly reduced compared to uniform weighting. Across all neural networks,
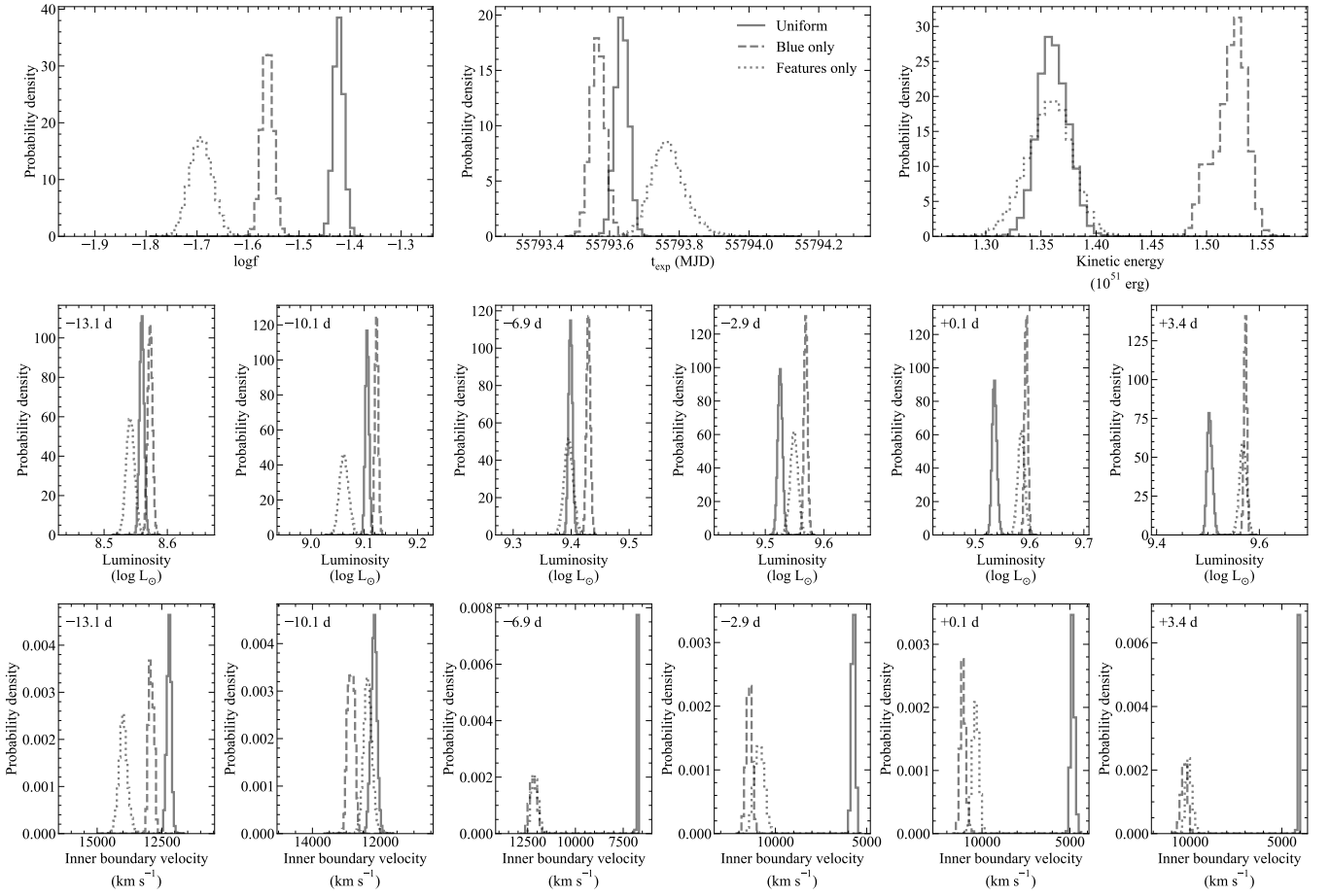
Figure 8. Total posterior distributions for best-fitting parameters to SN 2011fe using W7 models across all neural networks fit and assuming different weighting schemes. Phases are given relative to *B*-band maximum.

we find ln BF ~12 – 60 for blue only weighting compared to ~300 for uniform weighting.

For SN 2013dy, our uniform weighting scheme favours W7-like models for all epochs. With our blue weighting scheme, N100-like models are favoured for the first epoch only, however the difference in likelihoods is sufficiently large that the evidence is strongly in favour of an N100-like model overall (ln BF ~9 – 30). Indeed, our blue weighting scheme significantly reduces the scatter in best-fitting parameters for N100-like models and also produces visually better agreement, although we note that these models do somewhat over-predict the flux even at wavelengths <6 500 Å. This is again likely due to the photospheric approximation and the arbitrary wavelength cut-off used here. For our features weighting scheme, we again find W7-like models are favoured at most epochs, with N100-like being favoured only for the −2.5 d spectrum.

In summary, we have shown how the best-fitting model parameters, and even ejecta structure, can be significantly impacted by the choice of weighting schemes used during fitting. Given the known limitations of the photospheric approximation we argue that, of the three simple schemes presented here, our blue weighting scheme provides the best compromise between fitting spectral features and continuum shape. More complicated weighting schemes are beyond the scope of this paper, but should be explored further. Such schemes include those that do not exclude longer wavelengths entirely, but instead allow for a systematically higher flux, and those that vary the weights as a function of phase, in addition to wavelength. Our results show

that the best-fitting model parameters, and even the overall preferred explosion model, are sensitive to somewhat arbitrary choices of the likelihood function, demonstrated here by the omission of certain wavelengths. We stress that this quantitative fitting is nevertheless reproducible, due to the explicitly defined likelihood function and prior distributions. Manual fitting based on visual inspection is similarly arbitrary, but not reproducible due to the subjective nature with which model comparisons are made. We therefore strongly encourage future automated fitting routines and their applications to make explicitly clear what assumptions have been made during the fitting process and their expected impact.

7 DISCUSSION

Here we discuss the results of our `riddler` fitting. In Sect. 7.1, we discuss the limitations and assumptions of this work and where caution should be applied when using `riddler` to fit observations of SNe Ia. In Sect. 7.2 we compare the results of our fitting to previous studies and alternative fitting methods. Finally, the quantitative method of spectral fitting outlined here allows for inferences about the SN explosion physics. In Sect. 7.3 we discuss constraints on the explosion physics derived from our fitting.

7.1 Limitations and assumptions

As discussed in Sect. 6, the assumed wavelength-dependent weighting used during fitting can have a significant impact on the best-fitting model parameters and therefore care must be taken when applying automated fitters to observations. Conclusions can only be drawn within the boundaries of the assumed goodness-of-fit metric and priors. Here we discuss other assumptions and limitations of this work.

As with any machine learning-based technique, this work is limited by the datasets used to train our neural networks. We present neural networks trained on two Chandrasekhar mass explosion models, W7 and N100. Both models have been extensively studied within the literature and in general show good agreement with spectra of SNe Ia, but some differences are clearly apparent (see e.g. Branch et al. 1985; Sim et al. 2013). We therefore expect that in general our neural networks will not be able to reproduce all features observed in SNe Ia. Nevertheless, the relative level of agreement between different models can still provide useful insight into the explosion physics. While similar fractional offsets are typically not calculated for customised manual fitting, given the similar levels of agreement with observations we expect they have similar values.

In addition to these limitations, neither model is capable of capturing the full diversity among SNe Ia and therefore our fitting is currently limited to 'normal' SNe Ia (Benetti et al. 2005; Branch et al. 2006). By confining our training data to explosion model predictions however, we ensure that all of our models are physically consistent. Allowing individual elemental abundances to vary independently could result in statistically well-fitting, but physically unrealistic models. Such a scenario would make it difficult to provide meaningful constraints on the explosion mechanism for observed SNe Ia. Future work will include an expanded set of explosion models and mechanisms as the basis of our training datasets, thereby enabling further robust and quantitative constraints to be placed on the explosion physics for a larger and more diverse sample of SNe Ia.

Aside from the underlying structure of the ejecta, when constructing our training datasets we also adopted a few additional assumptions. To determine a suitable range for the input luminosity and photospheric velocity, we used observations and measurements of SNe Ia from Scalzo et al. (2019) and Foley et al. (2011) that were calculated relative to maximum light. Our TARDIS simulations however require a spectral phase relative to the time of explosion. We therefore assumed a rise time of 18.5 d for all SNe Ia. While some SNe Ia do show rise times of ~18.5 d, significant diversity also exists, with rise times extending from as low as ~15 up to ~25 d (Ganeshalingam et al. 2011; Firth et al. 2015; Miller et al. 2020). The exact measured value of the rise time is dependent on both the model assumptions and the data quality used during fitting (Miller et al. 2020). Future work will include a wider range of rise times within the training datasets, which will likely be necessary as other explosion models with different ejecta masses are included (Scalzo et al. 2014). We note however that given the wide range of luminosities and velocities used in our training datasets, our fitting technique does not assume a rise time of 18.5 d when applied to observations. This is demonstrated from our fits to both SN 2011fe and SN 2013dy, which imply rises times of ~20.6 d and ~16 – 20 d, respectively (depending on the model and assumed weighting). We also set an arbitrary lower limit on the inner boundary velocity of our models of $4\,000$ km s^{-1}. As discussed in Sect. 5, some of our best-fitting models favour inner boundaries at this lower limit, indicating our training dataset does not cover the full parameter space required to match the observations. Again however this only affects fits with uniform weighting

of all wavelengths and arises from the photospheric approximation. Alternative schemes that account for this approximation do not show similarly skewed velocities.

The spectra used to train our neural networks were generated from $3\,000 - 9\,000$ Å and therefore, our neural networks cannot currently be used to fit UV spectra. These wavelengths are highly sensitive to metallicity effects and can provide important constraints on the progenitor metallicity (Foley & Kirshner 2013; Brown et al. 2015). Future work will include an expanded wavelength range, but we again stress the importance of appropriate weighting schemes, which can significantly impact the results.

Finally, during our fitting process we also set a prior constraint that the inner boundary velocity is always decreasing with time, which helps to improve the convergence of the ULTRANEST fits. This has also generally been found by manual spectral fitting (e.g. Stehle et al. 2005; Mazzali et al. 2014). To test the impact of this explicit assumption, we run additional fits for both SN 2011fe and SN 2013dy with no prior constraint. For SN 2011fe, we find no significant differences for the best-fitting parameters, which do naturally show a decreasing inner boundary velocity with time. One exception however is the +3.4 d spectrum for which we find a best-fitting inner boundary velocity closer to the previous +0.1 d spectrum (~4 840 km s^{-1}) rather than the lower limit of our training dataset (4 000 km s^{-1}). Both fits however are also consistent with each other within the full uncertainty range. In the case of SN 2013dy, we find noticeable changes in the best-fitting parameters and that the inner boundary velocity does not monotonically decrease with time. Again we note that the N100-like models show overall poor agreement with SN 2013dy when considering a uniform wavelength weighting. Assuming our blue weighting scheme however, which shows significantly better agreement, our best-fitting models recover a monotonically decreasing inner boundary velocity and predict comparable best-fitting parameters to those with a prior velocity constraint. We therefore argue that the assumed prior constraint on the velocity evolution is unlikely to have a significant impact on models that provide a good match to the observations. For well-sampled spectral sequences, with much smaller time intervals between subsequent observations, a somewhat relaxed velocity constraint may be necessary.

7.2 Comparisons with existing models and methods

The spectra of SN 2011fe fit with `riddler` in Sect. 5 were the subject of a previous analysis by Mazzali et al. (2014). Heringer et al. (2017) also present spectral comparisons between these spectra and TARDIS models calculated using similar input parameters. Here we compare the results of our fits with these previous results. We note that model spectra presented by Mazzali et al. (2014) were calculated with a different radiative transfer code and therefore we expect some discrepancies between best-fitting parameters and spectra to arise due to differences in the radiative transfer treatments and atomic data (Heringer et al. 2017). Mazzali et al. (2014) present spectra calculated for SN 2011fe using both the W7 and WS15DD1 models (Iwamoto et al. 1999). The WS15DD1 model is a delayed detonation explosion similar to the N100 model used here however we note that the WS15DD1 model was calculated in one dimension, has a slightly smaller ^{56}Ni mass, and shows a different abundance distribution.

Based on their spectral fits with the W7 model, Mazzali et al. (2014) argue for an explosion epoch of MJD = 55 795.2±0.5, whereas our `riddler` fits predict an earlier explosion epoch (by ~1.6±0.7 d) of MJD = 55 793.63±0.5 (across all of the weighting schemes included here). Mazzali et al. (2014) show how differences in the rise time can have a significant impact on the earliest spectrum at −15.3 d, with

longer rise times (i.e. earlier explosion epochs) producing worse agreement. Spectra even ~2 d later however are less sensitive to the rise time and show only minor variations. The −15.3 d spectrum modelled by Mazzali et al. (2014) was not included in our fits as the estimated phase relative to explosion was outside the time range covered by our input parameters (5 – 25 d). This likely explains the difference in rise time estimates and further highlights the need for spectra as early as possible to place tight constraints on the explosion epoch through spectral modelling. We note that both explosion epoch estimates are still earlier than those predicted from the light curve (Nugent et al. 2011; Magee et al. 2020).

Overall, we find qualitatively similar levels of agreement between our best-fitting spectra and those presented by Mazzali et al. (2014) (and Heringer et al. (2017)) for our blue only weighting scheme, which we consider a more appropriate comparison. Both sets of models generally predict similar velocities and widths for most spectral features. We find that our spectra typically provide better agreement with the relative strengths of features, such as the S II W feature and the iron blends around ~4 500 and 5 000 Å, however the Mazzali et al. (2014) spectra provide better matches to the NUV. Comparing the model input parameters, we find that our luminosities are comparable to those estimated by Mazzali et al. (2014) and differ by ≲0.1 dex. The inner boundary velocities however show significant differences, with our values being systematically higher (by up to ~3 900 km s^{-1}) than those of Mazzali et al. (2014) when assuming a blue only weighting scheme. For our uniform weighting scheme, we find our inner boundary velocities are instead systematically lower (by up to ~3 500 km s^{-1}). We note that the inner boundary velocities presented by Heringer et al. (2017) also show some differences (50 – 850 km s^{-1}) relative to Mazzali et al. (2014), despite the same explosion epoch and luminosity. Given that many of our spectral line ratios provide better matches to SN 2011fe when assuming a blue only weighting scheme, this would indicate that the temperature and ionisation state of these models is closer to that of SN 2011fe. This further demonstrates the importance of an appropriate weighting scheme to account for excess flux due to the photospheric approximation.

Aside from spectra calculated with existing explosion models, by modelling progressively later spectra Mazzali et al. (2014) also develop a custom ejecta profile. This custom profile is motivated primarily by differences in the UV spectra and provides better qualitative agreement with SN 2011fe spectra. This profile contains a tail of material extending to higher velocities than in the W7 model, but with densities lower than the WS15DD1 delayed detonation model. Similarly, Mazzali et al. (2014) find a more mixed composition for the ejecta produces better agreement than the standard W7 model. Although we do not construct custom ejecta models and do not consider the UV here, we find qualitatively similar results. Both our best-fitting W7- and N100-like models predict similar density profiles. With our approach however, neither best-fitting model has the freedom to extend the density profile to higher velocities. In addition, we find that the less stratified structure of the N100-like models produces better agreement with many of the spectral features in SN 2011fe, particularly at early times. Therefore, even though our models do not have complete freedom when fitting individual SNe, comparative analysis between multiple explosion models can provide useful insights without the need for customisation of models for each SN.

Finally, we also compare our results to a DALEK fit of the −10.1 d SN 2011fe spectrum using the method outlined by O'Brien et al. (2023). This spectrum was selected by O'Brien et al. (2023) as it lies approximately 8 – 12 d post-explosion. We note that this DALEK fit is performed using continuum normalised spectra and therefore

we cannot directly compare on an absolute flux scale, as in our `riddler` fits. Based on our DALEK fit, we find an explosion epoch of 55 794.3±0.2, which lies approximately between the values predicted by our fits and Mazzali et al. (2014). We also find an inner boundary velocity of ~ 10 800±160 km s^{-1}, which is lower than the best-fitting value for our blue only weighted N100-like model (the best match at this epoch, ~11 900 – 13 300 km s^{-1}) and marginally lower than the Mazzali et al. (2014) custom model (11 300 km s^{-1}).

Using the masses determined by our DALEK fit for the −10.1 d SN 2011fe spectrum, we use equations 1 – 7 in O'Brien et al. (2023) to reconstruct the best-fitting ejecta structure. We note however that the simulation only includes material above the inner boundary (10 800± 160 km s^{-1} at this epoch). We find that the structure of the DALEK model shows a generally shallower density slope and lower overall density than the W7 or N100 models. This lower overall density is the likely cause of the lower inner boundary velocity, as the photosphere needs to be placed deeper inside the model to produce the required temperature and features. Indeed, the density at the inner boundary is comparable across all of the models. The abundance distributions also show some differences. From our DALEK fit, the distribution of intermediate mass elements peaks at ~14 000 km s^{-1}. This is comparable to the N100 model, but slightly higher than our best-fitting W7- and N100-like models (~11 000 – 12 000 km s^{-1}). The velocity range of intermediate mass elements is also comparable to the W7 model, although more narrow than N100. At the inner regions of the ejecta model, our DALEK fit is dominated by ^{56}Ni, while both our W7- and N100-like models show more mixed structures with some ^{56}Ni, other iron-group elements, and a significant fraction of intermediate mass elements. Despite the different training data and methods used, we find qualitatively consistent results between our `riddler` and DALEK fits, although again we note that we are unable to directly compare the absolute luminosities. Both methods indicate a larger fraction of intermediate mass elements at high velocities relative to the standard W7 model is required to reproduce the early spectra of SN 2011fe.

In summary, we find that our `riddler` fits are able to produce comparable best-fitting spectra to existing methods using either customised or automated fitting. Current techniques rely on modelling of progressively later spectra, as the photosphere recedes deeper inside the ejecta, to determine customised abundance profiles. As new layers of the ejecta are exposed, the composition in the outer regions is held fixed. This ensures a self-consistent model. A key benefit of `riddler` however is that all spectra are fit simultaneously, naturally ensuring a consistent ejecta profile, and reproducing the luminosity and temperature evolution of the observed SN.

7.3 Explosion physics constraints

Our fits to SN 2011fe and SN 2013dy show that neither the W7 nor N100 explosion model is able to reproduce all features of either SN, however different regions of the ejecta do show better agreement with individual models. At early times, we find that the N100 model produces better agreement with SN 2011fe than W7, while at later times this is reversed. Such changes in the best-fitting models at different phases can provide further constraints on the explosion physics.

The outer regions of the W7 model are dominated by unburned carbon and oxygen, following quenching of the deflagration front (Nomoto 1984). Consistent with previous studies, our results show that these outer layers in the W7 model cannot reproduce early observations of SNe without some additional mixing (Branch et al. 1985; Stehle et al. 2005). Relative to SN 2011fe, our W7-like models show

stronger carbon and oxygen features, in particular C II $\lambda6\,578$ and O I $\lambda7\,774$, and significantly weaker intermediate mass element features at these times. While features indicative of unburned material have been detected in SN 2011fe and other SNe Ia at early times (Nugent et al. 2011; Parrent et al. 2012; Folatelli et al. 2012), they are not as strong as those shown by our models. This would indicate that enhanced burning is required or increased mixing to reduce the abundance in the outer layers and extend this distribution down to lower velocities. For the N100 model, unburned fuel is also found in the outer ejecta, but these regions are instead mostly dominated by a larger fraction of burned material in the form of intermediate mass elements (Seitenzahl et al. 2013). Unburned material does however extend to much lower velocities than in the W7 model. The O I $\lambda7\,774$ feature predicted by our N100 models shows good agreement with the shape of the feature around $\sim7\,500$ Å in SN 2011fe, but is slightly stronger than observed for the earliest spectrum at -13.1 d. This could indicate that while the overall velocity-distribution is similar to SN 2011fe, a somewhat reduced mass fraction is preferred for the outermost ejecta at least. Such differences in the distribution of burned and unburned material are a natural consequence of buoyancy (Khokhlov 1995; Townsley et al. 2007). In multi-dimensional simulations of deflagrations, buoyancy of the deflagration ash drives plumes towards the outer ejecta, while in one-dimensional models, such as W7, no such buoyancy can occur (Pakmor et al. 2024). This leads to a highly stratified ejecta with burned material from the high density regions of the white dwarf confined to the inner ejecta.

None of our best fitting models are able to reproduce the high-velocity Ca II NIR features observed in SN 2011fe. High-velocity features are commonly observed in spectra of SNe Ia up to maximum light at velocities $\sim5\,000 - 10\,000$ km s^{-1} above the photosphere (Mazzali et al. 2005; Maguire et al. 2014; Silverman et al. 2015). Although the cause of these features is currently unknown, it has been suggested that they could arise from properties of the progenitor system, through interaction with circumstellar material, or from properties of the explosion, as certain explosion scenarios predict high-velocity shells of material from incomplete silicon burning (Wang et al. 2003; Gerardy et al. 2004; Mazzali et al. 2005). Magee et al. (2021) present models of double detonation explosions (Bildsten et al. 2007; Kromer et al. 2010; Shen & Bildsten 2014; Polin et al. 2019) and demonstrate the impact of a high-velocity shell of material on the Si II $\lambda6\,355$ feature at early times. In addition, Clark et al. (2021) recently argued against circumstellar interaction based on multiple probes believed to be linked to circumstellar interaction and the lack of any correlation. The W7 and N100 models implemented into the current version of `riddler` do not contain high-velocity shells of material or circumstellar interaction, therefore the lack of agreement with the high-velocity features is unsurprising. Nevertheless our models do cover a wide range of ionisation states for ejecta. The fact that these are unable to simultaneously match the photospheric and high-velocity features indicates additional complexity is required. Future versions of `riddler` will incorporate models that do contain high-velocity shells and therefore may be able to provide direct evidence in favour of or refute specific interpretations.

8 CONCLUSIONS

In this work, we presented `riddler`, a method for automated and quantitative fitting of SNe Ia spectral time series, beginning a few days after explosion up to shortly after maximum light. Using predictions from the well-studied W7 (Nomoto 1984) and N100 (Seitenzahl et al. 2013) explosion models, we developed datasets consisting of

100 000 spectra per model with the TARDIS radiative transfer code (Kerzendorf & Sim 2014). These datasets were then used to train a series of neural networks that act as emulators of full TARDIS simulations. Using our emulators, we fit spectra of SNe Ia with ULTRANEST (Buchner 2021) to quantify the relative likelihoods of explosion models and determine the best-fitting parameters and posterior distributions. Compared to previous studies using similar emulators, our models incorporate densities and compositions from realistic explosion models. They can therefore be used to directly quantify the relative agreement between multiple theoretically predicted ejecta structures. Our fits are also performed to multiple spectra of a given SN at different epochs simultaneously, thereby naturally producing a self-consistent model and matching the luminosity and temperature evolution of the observed SN.

To demonstrate the viability of `riddler`, we fit model spectra generated in the same manner as our training datasets, but not seen during training. We showed that `riddler` is able to recover the input parameters of the model with reasonable accuracy. We then used `riddler` to fit observations of two well-studied SNe Ia, SN 2011fe (Nugent et al. 2011) and SN 2013dy (Zheng et al. 2013). Based on our fits, we showed that in general the W7 model is strongly favoured overall for our assumed likelihood and priors. Through qualitative comparisons and by considering the likelihoods of the models at each phase however, we showed that the earliest spectra are not well reproduced by the W7 model. Instead a more mixed ejecta (similar to the N100 model) is favoured, consistent with previous studies. Comparing our results to existing models and methods, we find comparable agreement between spectra.

Using `riddler`, we also demonstrated the importance of different weighting schemes when quantifying the goodness-of-fit between model spectra and observations. Depending on the relative weighting of wavelengths or specific features, the best-fitting model parameters and spectra can show significant variation. This also applies in general to the likelihood function and prior distributions assumed when performing such fits, both of which can have a significant impact on the overall best-fitting model and parameters. Manual fitting of spectra however is typically performed in a qualitative and subjective manner, and is therefore not reproducible. Although the exact goodness-of-fit metric and priors do require somewhat arbitrary choices to be made when using automated fitters, these choices can at least be made explicit and quantified, enabling reproducible and consistent studies of SNe spectra.

In the coming years, spectral modelling of SNe Ia will necessarily become increasingly automated. Manual fitting of individual spectra is both time and computationally intensive, and cannot provide a quantitative measurement of the relative agreement between models nor the uncertainties for their best-fitting parameters. The number of SNe Ia spectra currently archived already dwarfs our ability to perform manual fitting of them. With large scale spectroscopic surveys, such as the SED Machine as part of Zwicky Transient Facility (Blagorodnova et al. 2018; Bellm et al. 2019) and TiDES on 4MOST (Swann et al. 2019), this will continue to grow even larger. Automated fitters are therefore required to make continued progress in constraining the explosion physics of SNe Ia. With `riddler`, we have developed a tool that enables entire spectral sequences to be fit simultaneously and allows for robust comparisons across explosion models. Future work will continue to improve the training dataset and neural networks used in `riddler`, including expansion to a wider variety of explosion models.

ACKNOWLEDGEMENTS

DATA AVAILABILITY

`riddler` is publicly available on GitHub[5].

REFERENCES

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/
Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, arXiv e-prints, p. arXiv:1907.10902
Ashall C., Mazzali P., Bersier D., Hachinger S., Phillips M., Percival S., James P., Maguire K., 2014, MNRAS, 445, 4427
Ashall C., Mazzali P. A., Pian E., James P. A., 2016, MNRAS, 463, 1891
Ashall C., et al., 2018, MNRAS, 477, 153
Baron E., Höflich P., Krisciunas K., Dominguez I., Khokhlov A. M., Phillips M. M., Suntzeff N., Wang L., 2012, ApJ, 753, 105
Bellm E. C., et al., 2019, PASP, 131, 018002
Benetti S., et al., 2005, ApJ, 623, 1011
Bildsten L., Shen K. J., Weinberg N. N., Nelemans G., 2007, ApJ, 662, L95
Blagorodnova N., et al., 2018, PASP, 130, 035003
Blinnikov S. I., Eastman R., Bartunov O. S., Popolitov V. A., Woosley S. E., 1998, ApJ, 496, 454
Blinnikov S. I., Röpke F. K., Sorokina E. I., Gieseler M., Reinecke M., Travaglio C., Hillebrandt W., Stritzinger M., 2006, A&A, 453, 229
Branch D., Doggett J. B., Nomoto K., Thielemann F. K., 1985, ApJ, 294, 619
Branch D., et al., 2006, PASP, 118, 560
Brown P. J., Baron E., Milne P., Roming P. W. A., Wang L., 2015, ApJ, 809, 37
Buchner J., 2021, The Journal of Open Source Software, 6, 3001
Chen X., Hu L., Wang L., 2020, ApJS, 250, 12
Chen X., Wang L., Hu L., Brown P. J., 2024, ApJ, 962, 125
Chollet F., et al., 2015, Keras, https://keras.io
Clark P., Maguire K., Bulla M., Galbany L., Sullivan M., Anderson J. P., Smartt S. J., 2021, MNRAS, 507, 4367
Dessart L., Blondin S., Hillier D. J., Khokhlov A., 2014, MNRAS, 441, 532

Ergon M., Fransson C., Jerkstrand A., Kozma C., Kromer M., Spricer K., 2018, A&A, 620, A156
Firth R. E., et al., 2015, MNRAS, 446, 3895
Folatelli G., et al., 2012, ApJ, 745, 74
Foley R. J., Kirshner R. P., 2013, ApJ, 769, L1
Foley R. J., Sanders N. E., Kirshner R. P., 2011, ApJ, 742, 89
Ganeshalingam M., Li W., Filippenko A. V., 2011, MNRAS, 416, 2607
Gerardy C. L., et al., 2004, ApJ, 607, 391
Hachinger S., Mazzali P. A., Taubenberger S., Pakmor R., Hillebrandt W., 2009, MNRAS, 399, 1238
Hamuy M., et al., 2006, PASP, 118, 2
Heringer E., van Kerkwijk M. H., Sim S. A., Kerzendorf W. E., 2017, ApJ, 846, 15
Heringer E., van Kerkwijk M. H., Sim S. A., Kerzendorf W. E., Graham M. L., 2019, ApJ, 871, 250
Hillier D. J., Dessart L., 2012, MNRAS, 424, 252
Höflich P., 1995, ApJ, 443, 89
Höflich P., 2003, in Hubeny I., Mihalas D., Werner K., eds, Astronomical Society of the Pacific Conference Series Vol. 288, Stellar Atmosphere Modeling. p. 185
Hsiao E. Y., et al., 2013, ApJ, 766, 72
Iwamoto K., Brachwitz F., Nomoto K., Kishimoto N., Umeda H., Hix W. R., Thielemann F.-K., 1999, ApJS, 125, 439
Jerkstrand A., 2017, Spectra of Supernovae in the Nebular Phase. Springer International Publishing, Cham, pp 795–842, doi:10.1007/978-3-319-21846-5_29, https://doi.org/10.1007/978-3-319-21846-5_29
Kaas R. E., Raftery A. E., 1995, Journal of the American Statistical Association, 90, 773
Kasen D., Thomas R. C., Nugent P., 2006, ApJ, 651, 366
Kerzendorf W. E., Sim S. A., 2014, MNRAS, 440, 387
Kerzendorf W. E., Vogl C., Buchner J., Contardo G., Williamson M., van der Smagt P., 2021, ApJ, 910, L23
Khokhlov A. M., 1995, ApJ, 449, 695
Kromer M., Sim S. A., 2009, MNRAS, 398, 1809
Kromer M., Sim S. A., Fink M., Röpke F. K., Seitenzahl I. R., Hillebrandt W., 2010, ApJ, 719, 1067
Kromer M., Ohlmann S., Röpke F. K., 2017, Mem. Soc. Astron. Italiana, 88, 312
Li W. D., et al., 1999, AJ, 117, 2709
Maas A. L., Hannun A. Y., Ng A. Y., 2013, Proc. ICML, 30
Magee M. R., et al., 2016, A&A, 589, A89
Magee M. R., et al., 2017, A&A, 601, A62
Magee M. R., Sim S. A., Kotak R., Maguire K., Boyle A., 2019, A&A, 622, A102
Magee M. R., Maguire K., Kotak R., Sim S. A., Gillanders J. H., Prentice S. J., Skillen K., 2020, A&A, 634, A37
Magee M. R., Maguire K., Kotak R., Sim S. A., 2021, MNRAS, 502, 3533
Maguire K., et al., 2014, MNRAS, 444, 3258
Mazzali P. A., Lucy L. B., 1993, A&A, 279, 447
Mazzali P. A., et al., 2005, ApJ, 623, L37
Mazzali P. A., Röpke F. K., Benetti S., Hillebrandt W., 2007, Science, 315, 825
Mazzali P. A., Sauer D. N., Pastorello A., Benetti S., Hillebrandt W., 2008, MNRAS, 386, 1897
Mazzali P. A., et al., 2014, MNRAS, 439, 1959
Miller A. A., et al., 2020, ApJ, 902, 47
Nomoto K., 1984, ApJ, 277, 791
Nomoto K., Thielemann F.-K., Yokoi K., 1984, ApJ, 286, 644
Nugent P. E., et al., 2011, Nature, 480, 344
O'Brien J. T., et al., 2021, ApJ, 916, L14
O'Brien J. T., et al., 2023, arXiv e-prints, p. arXiv:2306.08137
Ogawa M., Maeda K., Kawabata M., 2023, arXiv e-prints, p. arXiv:2308.03834
Pakmor R., Seitenzahl I. R., Ruiter A. J., Sim S. A., Roepke F. K., Taubenberger S., Bieri R., Blondin S., 2024, arXiv e-prints, p. arXiv:2402.11010
Pan Y. C., et al., 2015, MNRAS, 452, 4307
Parrent J. T., et al., 2012, ApJ, 752, L26
Polin A., Nugent P., Kasen D., 2019, ApJ, 873, 84

---

[4] https://johannesbuchner.github.io/UltraNest/
[5] https://github.com/MarkMageeAstro/riddler

Röpke F. K., Sim S. A., 2018, Space Sci. Rev., 214, 72

Sasdelli M., Mazzali P. A., Pian E., Nomoto K., Hachinger S., Cappellaro E., Benetti S., 2014, MNRAS, 445, 711

Scalzo R. A., Ruiter A. J., Sim S. A., 2014, MNRAS, 445, 2535

Scalzo R. A., et al., 2019, MNRAS, 483, 628

Seitenzahl I. R., et al., 2013, MNRAS, 429, 1156

Shen K. J., Bildsten L., 2014, ApJ, 785, 61

Shen K. J., Boos S. J., Townsley D. M., Kasen D., 2021, ApJ, 922, 68

Silverman J. M., Vinkó J., Marion G. H., Wheeler J. C., Barna B., Szalai T., Mulligan B. W., Filippenko A. V., 2015, MNRAS, 451, 1973

Sim S. A., 2017, Spectra of Supernovae During the Photospheric Phase. Springer International Publishing, Cham, pp 769–793, doi:10.1007/978-3-319-21846-5_28, https://doi.org/10.1007/978-3-319-21846-5_28

Sim S. A., et al., 2013, MNRAS, 436, 333

Stehle M., Mazzali P. A., Benetti S., Hillebrandt W., 2005, MNRAS, 360, 1231

Sullivan M., et al., 2011, ApJ, 732, 118

Swann E., et al., 2019, The Messenger, 175, 58

Tanaka M., Mazzali P. A., Stanishev V., Maurer I., Kerzendorf W. E., Nomoto K., 2011, Monthly Notices of the Royal Astronomical Society, 410, 1725

Thomas R. C., Nugent P. E., Meza J. C., 2011, PASP, 123, 237

Thrane E., Talbot C., 2020, Publ. Astron. Soc. Australia, 37, e036

Townsley D. M., Calder A. C., Asida S. M., Seitenzahl I. R., Peng F., Vladimirova N., Lamb D. Q., Truran J. W., 2007, ApJ, 668, 1118

Vogl C., Sim S. A., Noebauer U. M., Kerzendorf W. E., Hillebrandt W., 2019, A&A, 621, A29

Wang L., et al., 2003, ApJ, 591, 1110

Yaron O., Gal-Yam A., 2012, PASP, 124, 668

Zheng W., et al., 2013, ApJ, 778, L15

van Rossum D. R., 2012, ApJ, 756, 31

APPENDIX A: NEURAL NETWORK TRAINING

Here we provide a more detailed discussion of the accuracies for the neural networks used in this work. As discussed in Sect. 3, to determine the accuracy of each neural network, we use our final set of 20 000 testing models. These models were not previously seen during training of any of the neural networks. In Fig. A1 we show predicted W7 spectra with the highest mean fractional errors from our worst performing neural networks, while in Fig. A2 we show predicted spectra with the lowest mean fractional errors from our best performing neural networks. In general, the neural networks are able to reproduce most of the prominent features in the spectra, with the correct velocity and width, but struggle to reproduce weaker and more narrow features – likely due to their overall lower contribution to the neural network loss during training. For our worst performing neural networks, Fig. A1 shows that they struggle to reproduce the flux level for some spectral features and could be discrepant by $\gtrsim$10%. For our W7 neural networks, the largest errors occur at longer wavelengths where the spectra do not show any strong features. Conversely, our N100 neural networks show the largest errors, potentially up to ~20% in extreme cases, at wavelengths $\lesssim$4 000 Å.

In Fig. A3, we show the distribution of mean fractional errors across our 20 000 W7 and N100 testing models. For both the W7 and N100 models, we find that most spectra show typical mean fractional errors of ~1.6 – 1.8%, with the largest mean fractional errors being ~5 – 20% across all of the neural networks. Unsurprisingly, we find the mean fractional errors of our predicted spectra are highly correlated with sparsely-populated regions of the training dataset parameter space. In particular, those models with velocities towards the lower limit for a given time typically show higher mean fractional errors. Figure A4 shows the mean fractional error as a function of wavelength across the 20 000 W7 and N100 testing models, which is
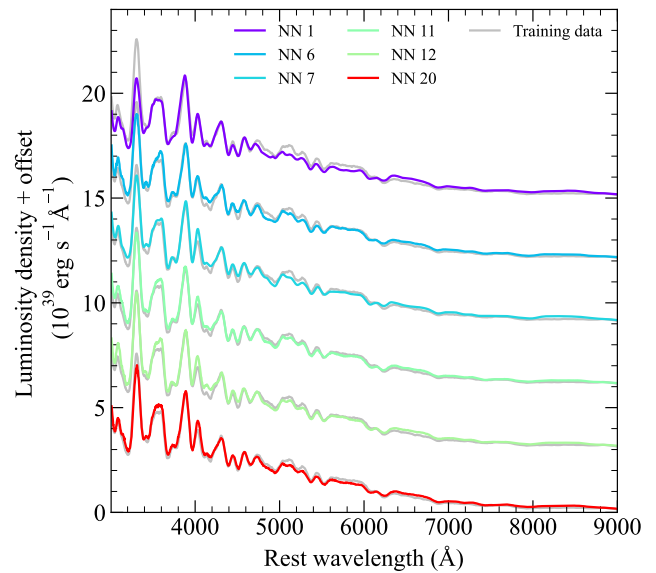
Figure A1. Predicted spectra for our worst performing neural networks compared against the true model spectrum with the largest mean fractional error (grey). Predicted spectra show typical mean fractional errors of ~9 – 14%. Maximum fractional errors for each predicted spectrum range from ~40 – 70% and always occur between ~8 100 – 8 300 Å.
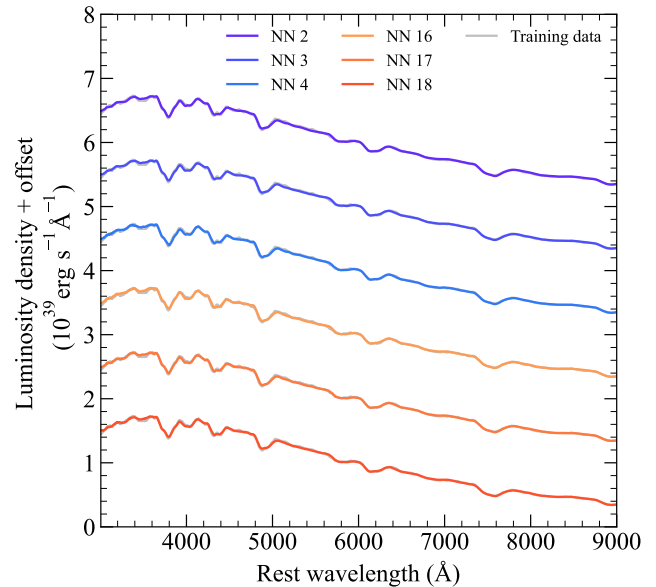

Figure A2. Predicted spectra for our best performing neural networks compared against the true model spectrum with the smallest mean fractional error (grey). Predicted spectra show typical mean fractional errors of ~0.8 – 0.9%. Maximum fractional errors for each predicted spectrum range from ~3 – 4% and occur in either the first or last wavelength bin, with the exception of NN 16 for which the maximum error occurs at ~8 600 Å.

generally highest in NUV and NIR spectral regions. To account for differences in the level of accuracy as a function of the model input parameters, we calculate the mean fractional error as a function of wavelength in a series of time and velocity bins. We then define an interpolator to calculate the mean fractional error for any values of time and velocity. When fitting spectra, we include this systematic

Table A1. Hyperparameters and performance for neural networks trained on W7 and N100 models

Neural network	Batch size	Learning rate	Neurons	Layers	W7 Median Mean FE	W7 Median Max FE	N100 Median Mean FE	N100 Median Max FE
					Varying layers			
1	10 000	1×10^{-3}	400	1	0.0166	0.0712	0.0192	0.0852
2	10 000	1×10^{-3}	400	2	0.0156	0.0684	0.0174	0.0802
3	10 000	1×10^{-3}	400	3	0.0155	0.0679	0.0174	0.0802
4	**10 000**	1×10^{-3}	**400**	**4**	**0.0156**	**0.0683**	**0.0174**	**0.0801**
5	10 000	1×10^{-3}	400	5	0.0157	0.0688	0.0174	0.0806
6	10 000	1×10^{-3}	400	6	0.0158	0.0693	0.0178	0.0811
					Varying neurons			
7	10 000	1×10^{-3}	100	4	0.0158	0.0693	0.0177	0.0809
8	10 000	1×10^{-3}	200	4	0.0156	0.0681	0.0174	0.0799
9	10 000	1×10^{-3}	300	4	0.0156	0.0682	0.0174	0.0802
4	**10 000**	1×10^{-3}	**400**	**4**	**0.0156**	**0.0683**	**0.0174**	**0.0801**
10	10 000	1×10^{-3}	500	4	0.0156	0.0680	0.0175	0.0800
11	10 000	1×10^{-3}	600	4	0.0157	0.0685	0.0176	0.0805
					Varying batch size			
12	20 000	1×10^{-3}	400	4	0.0157	0.0688	0.0175	0.0799
4	**10 000**	1×10^{-3}	**400**	**4**	**0.0156**	**0.0683**	**0.0174**	**0.0801**
13	5 000	1×10^{-3}	400	4	0.0156	0.0687	0.0174	0.0797
14	2 500	1×10^{-3}	400	4	0.0156	0.0683	0.0174	0.0800
15	1 000	1×10^{-3}	400	4	0.0156	0.0680	0.0175	0.0799
					Varying learning rate			
4	**10 000**	1×10^{-3}	**400**	**4**	**0.0156**	**0.0683**	**0.0174**	**0.0801**
16	10 000	5×10^{-4}	400	4	0.0155	0.0682	0.0173	0.0797
17	10 000	1×10^{-4}	400	4	0.0155	0.0681	0.0173	0.0799
18	10 000	5×10^{-5}	400	4	0.0155	0.0681	0.0173	0.0799
19	10 000	1×10^{-5}	400	4	0.0156	0.0688	0.0174	0.0806
20	10 000	5×10^{-6}	400	4	0.0157	0.0692	0.0175	0.0809
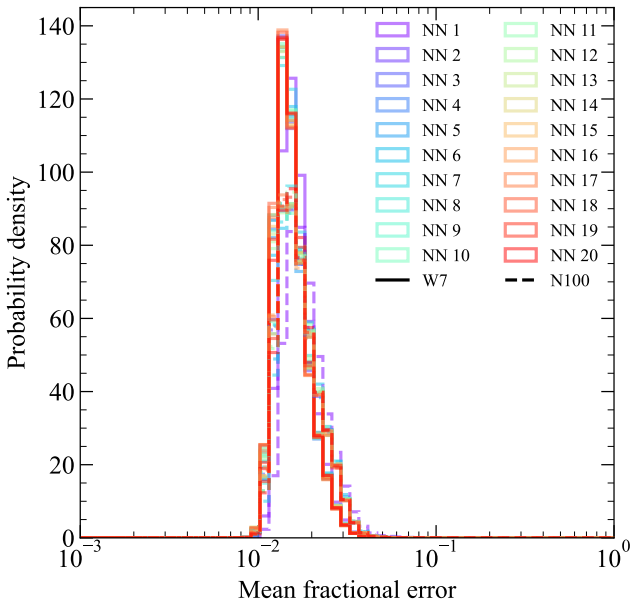
Figure A3. Distribution of mean fractional errors across predicted spectra for the 20 000 models in both our W7 and N100 testing datasets.

offset, based on the input parameters of the each model spectrum, as an estimate of the neural network prediction error ($\Delta_\lambda(t, v)$; Eqn. 6).

In summary, we find our neural networks reach typical accuracies of a few percent and are likely limited by sparse regions of the input parameter space for our training dataset. The inclusion of active learning could aid in training the neural networks in these regions and will be explored in future work (O'Brien et al. 2023). In addition, future work will also explore larger training datasets and different methods of data augmentation, in addition to more complicated neural network structures such those that can self-consistently estimate the uncertainty of the predictions (e.g. O'Brien et al. 2023).
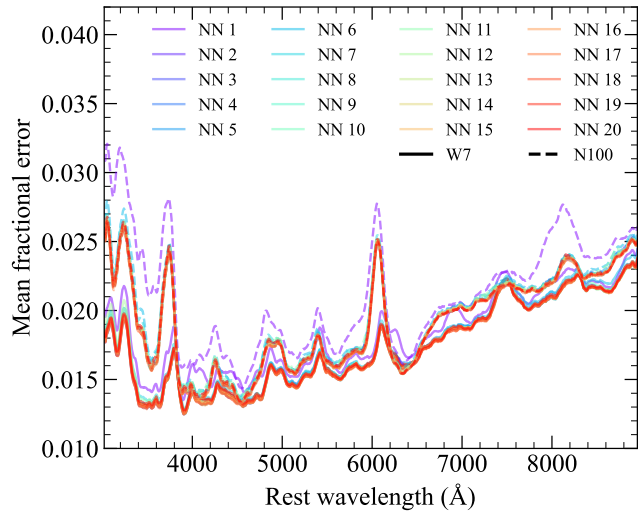
Figure A4. Mean fractional error as a function of wavelength for the 20 000 models in both our W7 and N100 testing datasets.

Table B1. Best-fitting parameters for SN 2011fe and SN 2013dy with W7 models and different weighting schemes

Phase (days)	Uniform Luminosity (log $L_\odot$)	Uniform Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta)$	Blue only Luminosity (log $L_\odot$)	Blue only Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$	Features only Luminosity (log $L_\odot$)	Features only Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$
				SN 2011fe					
-13.1	$8.56^{8.58}_{8.54}$	12221^{12598}_{11623}	-88349^{-88330}_{-88357}	$8.57^{8.59}_{8.56}$	12909^{13315}_{12579}	-62153^{-62150}_{-62156}	$8.54^{8.57}_{8.51}$	13992^{14996}_{12167}	-15628^{-15627}_{-15630}
-10.1	$9.11^{9.12}_{9.09}$	12181^{12589}_{11591}	-89153^{-89148}_{-89158}	$9.12^{9.14}_{9.11}$	12858^{13228}_{12415}	-62833^{-62827}_{-62835}	$9.06^{9.13}_{9.03}$	12366^{13678}_{11720}	-15734^{-15731}_{-15736}
-6.9	$9.40^{9.42}_{9.38}$	6740^{6956}_{6671}	-89730^{-89727}_{-89735}	$9.43^{9.44}_{9.42}$	12210^{12886}_{11492}	-63199^{-63198}_{-63201}	$9.40^{9.44}_{9.36}$	12115^{12757}_{10368}	-15803^{-15802}_{-15804}
-2.9	$9.53^{9.54}_{9.51}$	5730^{6185}_{5330}	-90022^{-90020}_{-90024}	$9.57^{9.58}_{9.56}$	11445^{12140}_{10775}	-63426^{-63425}_{-63427}	$9.55^{9.58}_{9.52}$	10888^{11924}_{9847}	-15857^{-15856}_{-15858}
$+0.1$	$9.54^{9.56}_{9.52}$	4840^{5320}_{4350}	-90225^{-90222}_{-90227}	$9.59^{9.61}_{9.58}$	11080^{11635}_{10239}	-63485^{-63480}_{-63487}	$9.58^{9.61}_{9.55}$	10369^{11097}_{9647}	-15871^{-15871}_{-15872}
$+3.4$	$9.50^{9.53}_{9.48}$	4033^{4338}_{4000}	-90520^{-90514}_{-90525}	$9.57^{9.59}_{9.56}$	10311^{11001}_{9666}	-63431^{-63425}_{-63435}	$9.57^{9.60}_{9.54}$	10058^{10719}_{9178}	-15873^{-15872}_{-15873}
	$\ln \mathcal{Z}$		$-538059^{-538032}_{-538082}$	$\ln \mathcal{Z}$		$-378589^{-378578}_{-378597}$	$\ln \mathcal{Z}$		-94816^{-94814}_{-94819}
				SN 2013dy					
-6.6	$9.42^{9.44}_{9.40}$	8956^{9490}_{8404}	-89846^{-89835}_{-89853}	$9.45^{9.51}_{9.43}$	11340^{12577}_{10226}	-63309^{-63300}_{-63315}	$9.52^{9.54}_{9.46}$	10778^{14242}_{10097}	-15755^{-15754}_{-15756}
-2.5	$9.50^{9.53}_{9.48}$	5307^{5519}_{5009}	-89943^{-89936}_{-89948}	$9.51^{9.52}_{9.49}$	5337^{5885}_{4884}	-63428^{-63420}_{-63433}	$9.55^{9.57}_{9.52}$	10569^{11466}_{9733}	-15811^{-15809}_{-15814}
-0.8	$9.51^{9.54}_{9.48}$	4465^{4613}_{4257}	-89945^{-89939}_{-89954}	$9.50^{9.52}_{9.49}$	5306^{5790}_{4853}	-63455^{-63452}_{-63457}	$9.54^{9.56}_{9.52}$	10438^{11255}_{9633}	-15807^{-15805}_{-15808}
$+1.2$	$9.48^{9.50}_{9.46}$	4003^{4066}_{4000}	-90012^{-90001}_{-90019}	$9.48^{9.50}_{9.47}$	5282^{5733}_{4834}	-63474^{-63472}_{-63476}	$9.52^{9.55}_{9.51}$	10188^{10858}_{9473}	-15803^{-15802}_{-15804}
$+4.5$	$9.41^{9.43}_{9.39}$	4001^{4026}_{4000}	-90241^{-90232}_{-90250}	$9.44^{9.46}_{9.43}$	5247^{5698}_{4818}	-63454^{-63451}_{-63459}	$9.48^{9.50}_{9.45}$	9912^{10506}_{9060}	-15810^{-15809}_{-15811}
	$\ln \mathcal{Z}$		$-450039^{-450024}_{-450050}$	$\ln \mathcal{Z}$		$-317169^{-317157}_{-317175}$	$\ln \mathcal{Z}$		-79031^{-79028}_{-79034}

Note. The best-fitting model parameters are determined based on the median of the total posterior across all of the neural networks included in the fits. Similarly, upper and lower limits are also based on the total posterior across all neural networks. We stress that these parameters assume uniform wavelength weighting, which has important limitations (see Sect. 6).

APPENDIX B: ADDITIONAL TABLES

APPENDIX C: ADDITIONAL FIGURES FOR SN 2011FE

APPENDIX D: ADDITIONAL FIGURES FOR SN 2013DY

This paper has been typeset from a TEX/LATEX file prepared by the author.

Table B2. Best-fitting parameters for SN 2011fe and SN 2013dy with N100 models and different weighting schemes

Phase (days)	Uniform Luminosity (log $L_\odot$)	Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$	Blue only Luminosity (log $L_\odot$)	Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$	Features only Luminosity (log $L_\odot$)	Inner boundary (km s^{-1})	$\ln \mathcal{L}(\theta\|o)$
					SN 2011fe				
−13.1	$8.56^{8.58}_{8.55}$	13987^{14348}_{13583}	-88263^{-88252}_{-88273}	$8.58^{8.60}_{8.56}$	13554^{14001}_{13093}	-62107^{-62098}_{-62115}	$8.57^{8.61}_{8.54}$	16084^{16819}_{15095}	-15556^{-15551}_{-15563}
−10.1	$9.10^{9.11}_{9.08}$	12359^{13013}_{11828}	-89197^{-89194}_{-89200}	$9.12^{9.14}_{9.10}$	12550^{13265}_{11902}	-62831^{-62827}_{-62833}	$9.11^{9.14}_{9.07}$	13389^{15594}_{12463}	-15687^{-15684}_{-15690}
−6.9	$9.39^{9.41}_{9.38}$	11267^{11995}_{10829}	-89785^{-89783}_{-89788}	$9.41^{9.43}_{9.40}$	11022^{11679}_{10541}	-63180^{-63179}_{-63182}	$9.41^{9.43}_{9.38}$	12546^{13611}_{11634}	-15777^{-15775}_{-15778}
−2.9	$9.52^{9.54}_{9.51}$	7225^{7738}_{6837}	-90248^{-90239}_{-90253}	$9.55^{9.57}_{9.54}$	10456^{10821}_{10121}	-63411^{-63409}_{-63415}	$9.55^{9.58}_{9.53}$	11602^{12453}_{10865}	-15853^{-15851}_{-15855}
+0.1	$9.52^{9.54}_{9.50}$	4035^{4340}_{4000}	-90288^{-90280}_{-90295}	$9.58^{9.59}_{9.57}$	10130^{10467}_{9840}	-63504^{-63500}_{-63508}	$9.60^{9.62}_{9.58}$	10955^{11832}_{10423}	-15905^{-15903}_{-15908}
+3.4	$9.49^{9.51}_{9.47}$	4011^{4249}_{4000}	-90526^{-90517}_{-90538}	$9.56^{9.57}_{9.55}$	9931^{10294}_{9405}	-63525^{-63520}_{-63528}	$9.62^{9.64}_{9.59}$	10471^{11679}_{9719}	-15979^{-15975}_{-15981}
	$\ln \mathcal{Z}$	$-538366^{-538351}_{-538377}$		$\ln \mathcal{Z}$	$-378622^{-378609}_{-378636}$		$\ln \mathcal{Z}$	-94811^{-94801}_{-94821}	
					SN 2013dy				
−6.6	$9.41^{9.45}_{9.39}$	11167^{13222}_{10325}	-89850^{-89833}_{-89868}	$9.45^{9.48}_{9.43}$	11987^{12989}_{11306}	-63246^{-63241}_{-63251}	$9.45^{9.49}_{9.41}$	14243^{16277}_{12084}	-15765^{-15756}_{-15775}
−2.5	$9.48^{9.51}_{9.46}$	10422^{12155}_{9206}	-90287^{-90254}_{-90330}	$9.52^{9.54}_{9.51}$	10888^{11306}_{10563}	-63438^{-63437}_{-63441}	$9.53^{9.56}_{9.51}$	12379^{13534}_{11808}	-15809^{-15808}_{-15810}
−0.8	$9.48^{9.51}_{9.47}$	9961^{11743}_{8510}	-90451^{-90432}_{-90474}	$9.53^{9.54}_{9.51}$	10622^{10983}_{10258}	-63462^{-63460}_{-63466}	$9.54^{9.56}_{9.51}$	12141^{12866}_{11382}	-15830^{-15826}_{-15835}
+1.2	$9.47^{9.50}_{9.45}$	9293^{11346}_{7572}	-90559^{-90530}_{-90596}	$9.52^{9.53}_{9.51}$	10309^{10712}_{9930}	-63478^{-63474}_{-63482}	$9.54^{9.56}_{9.51}$	11846^{12808}_{10798}	-15875^{-15869}_{-15883}
+4.5	$9.42^{9.44}_{9.40}$	5363^{6417}_{4000}	-90564^{-90462}_{-90596}	$9.49^{9.51}_{9.48}$	10075^{10566}_{9585}	-63467^{-63463}_{-63471}	$9.52^{9.56}_{9.48}$	8005^{12788}_{4000}	-16020^{-15987}_{-16073}
	$\ln \mathcal{Z}$	$-451751^{-451739}_{-451762}$		$\ln \mathcal{Z}$	$-317145^{-317141}_{-317148}$		$\ln \mathcal{Z}$	-79316^{-79312}_{-79322}	

Note. The best-fitting model parameters are determined based on the median of the total posterior across all of the neural networks included in the fits. Similarly, upper and lower limits are also based on the total posterior across all neural networks. We stress that these parameters assume uniform wavelength weighting, which has important limitations (see Sect. 6).
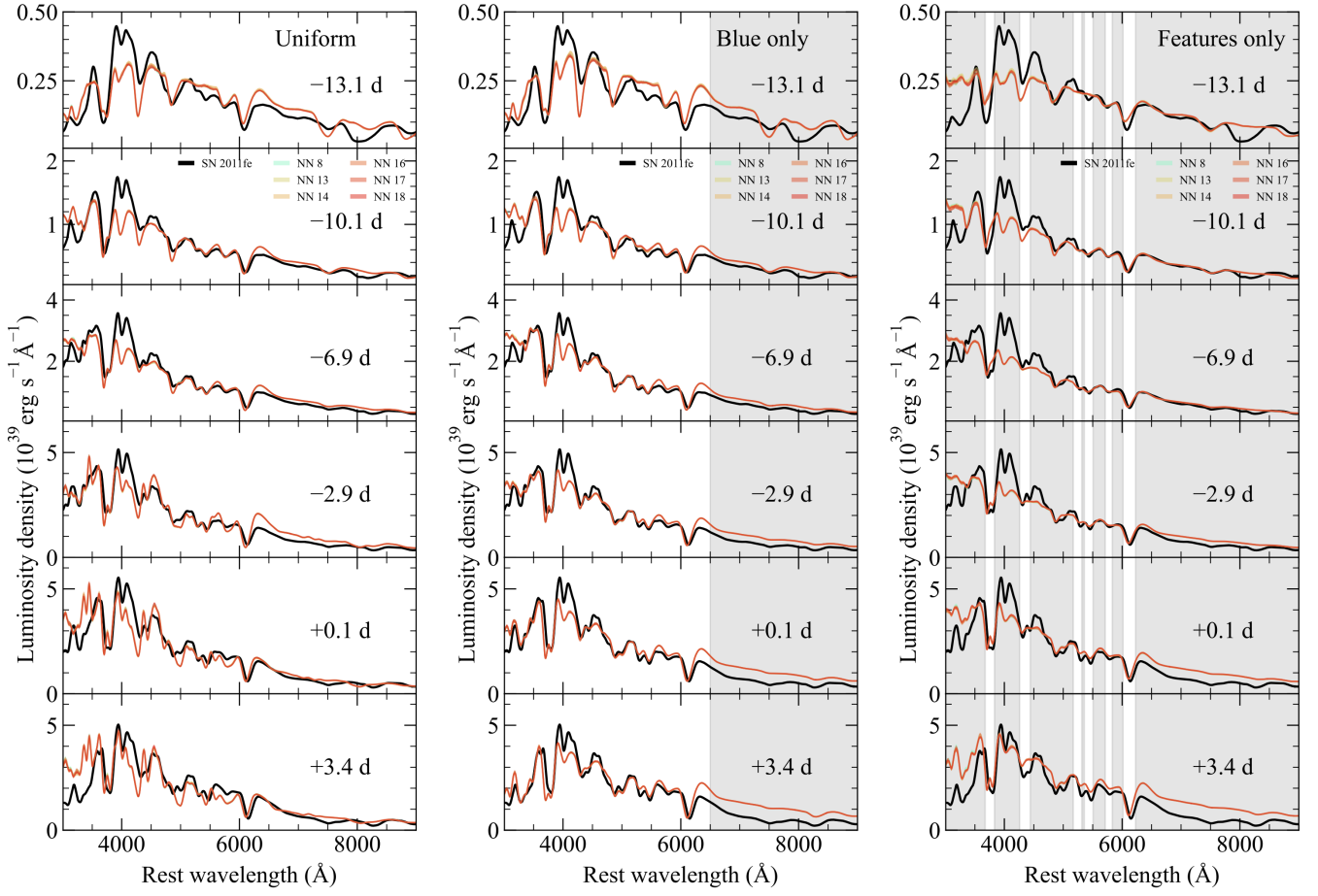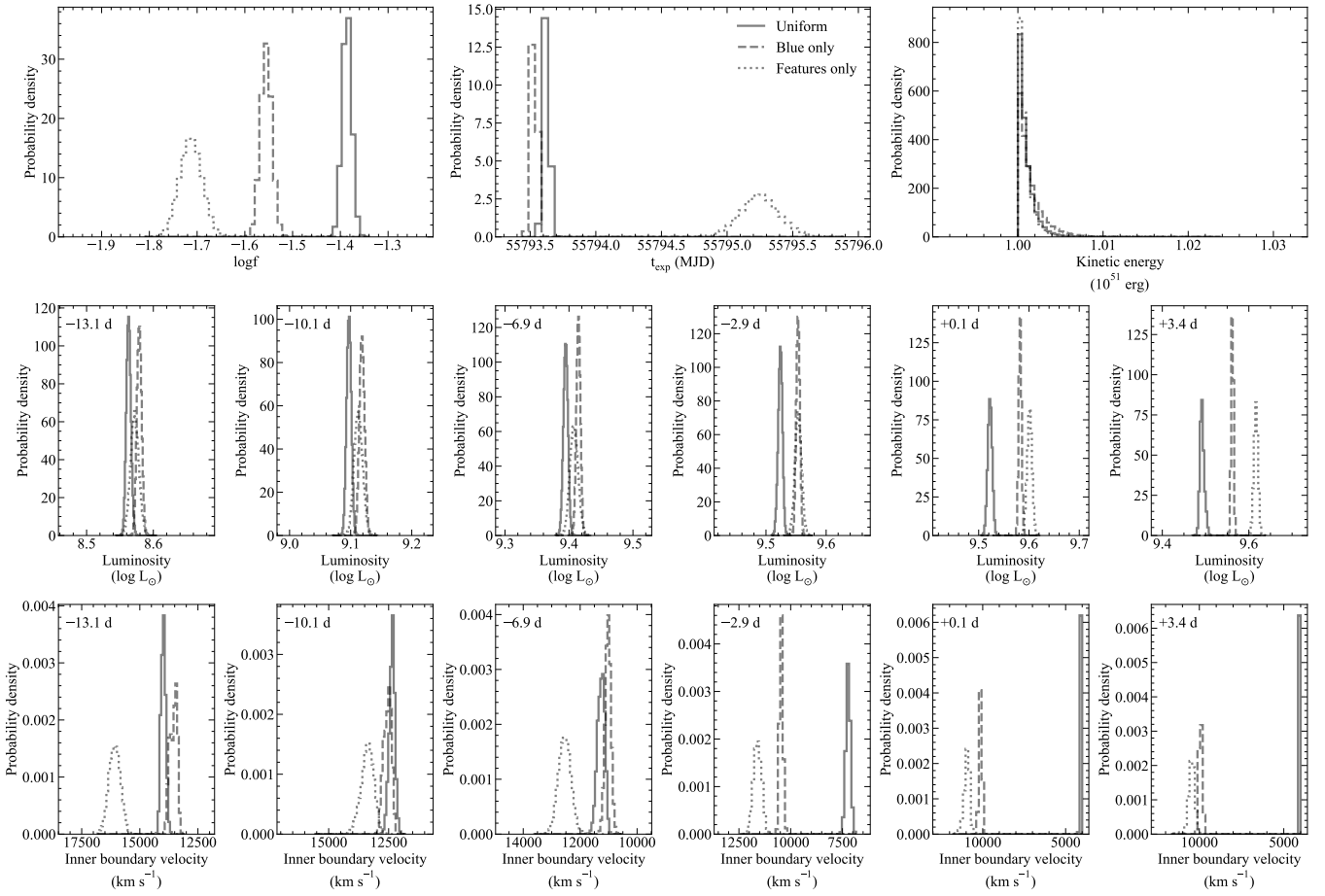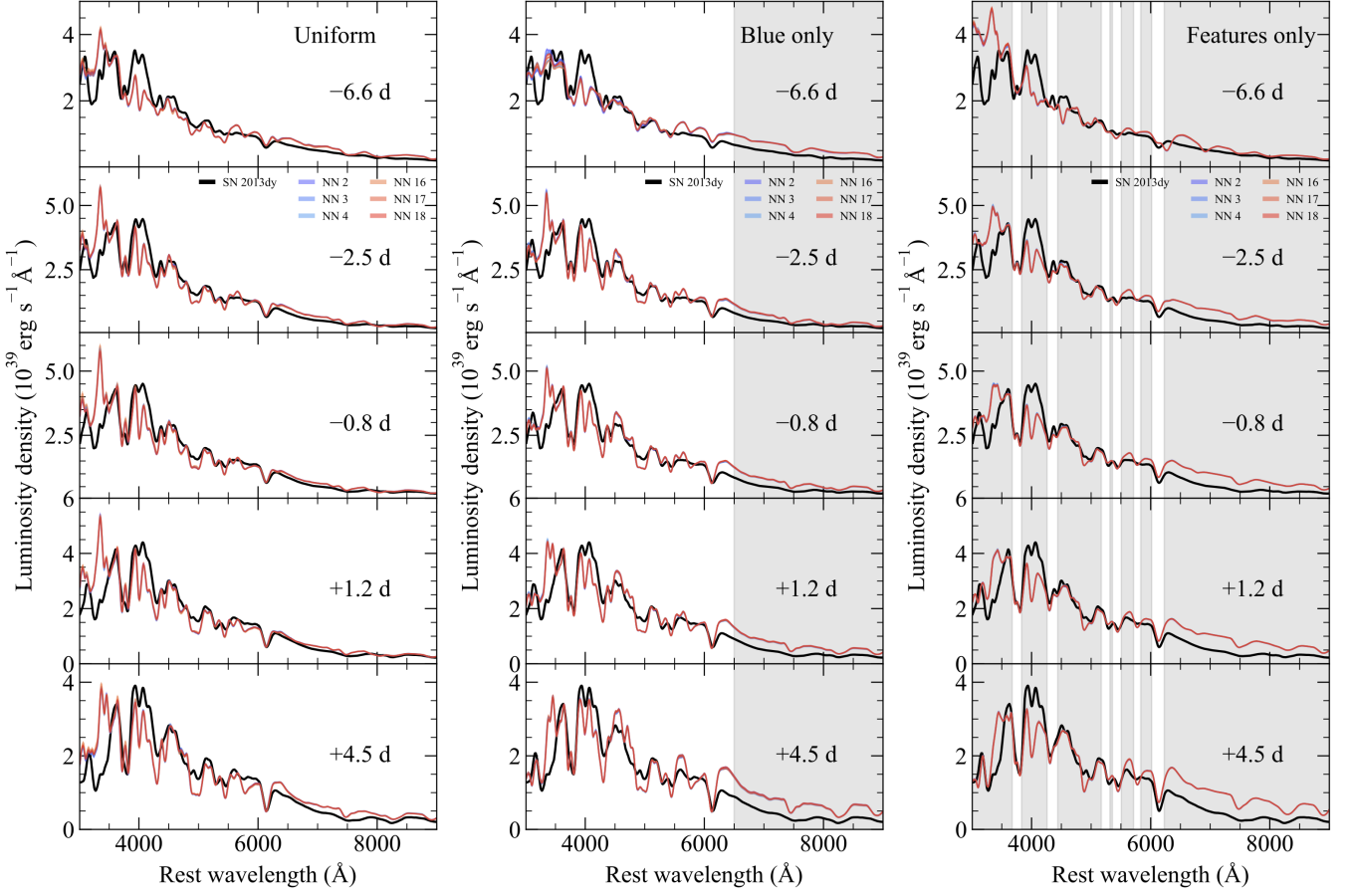
**Figure C1.** As in Fig. 7 for SN 2011fe and N100 models.

**Figure C2.** As in Fig. 8 for SN 2011fe and N100 models.

Figure D1. Comparison between SN 2013dy and neural network reconstructions of the best-fitting W7 model spectra for each neural network and assuming different weighting schemes. Spectra are shown on an absolute luminosity scale with no additional scaling or offsets applied. Phases are given relative to *B*-band maximum. Grey shaded regions are not included during fitting.
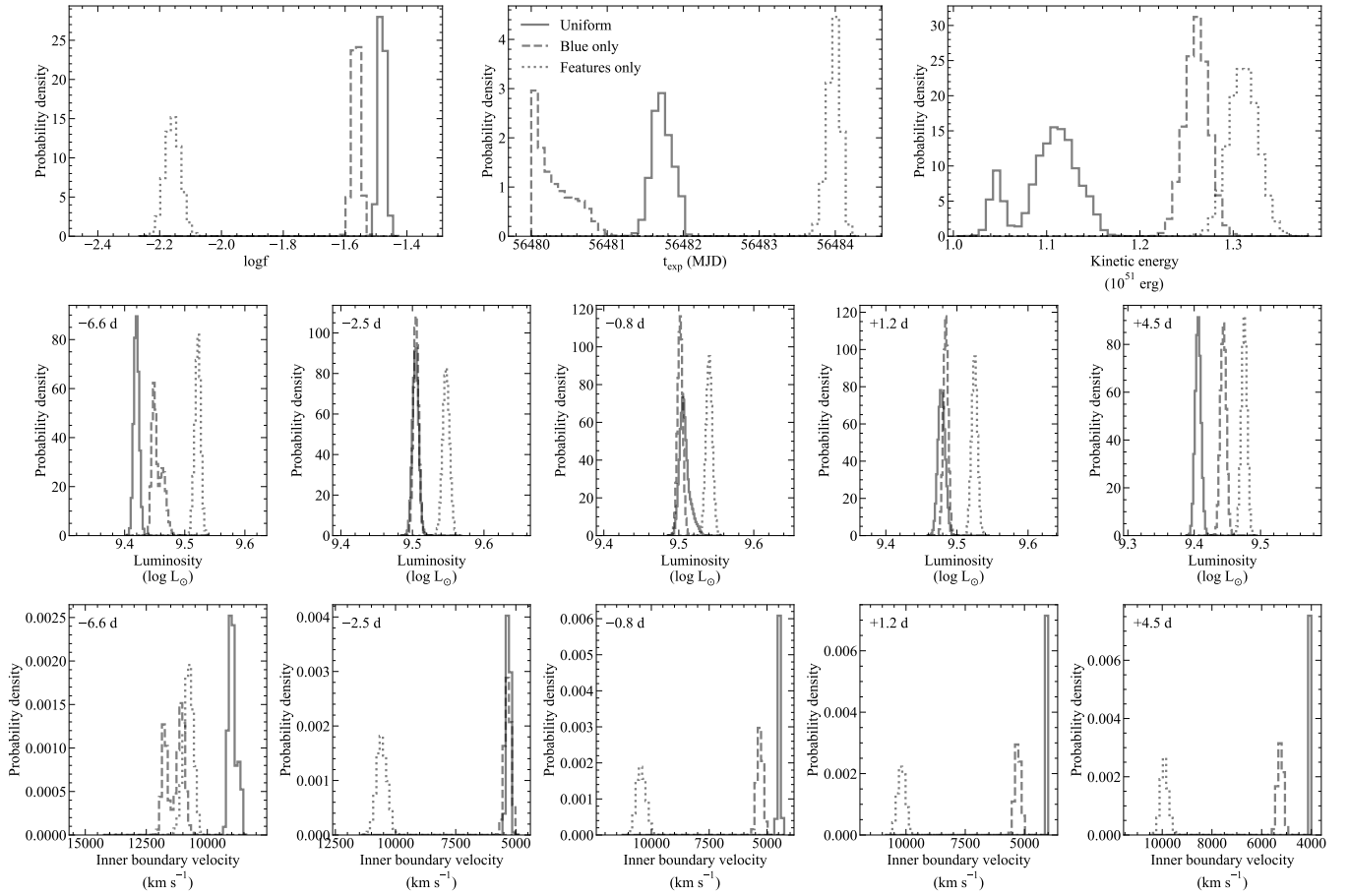
Figure D2. Total posterior distributions for best-fitting parameters to SN 2013dy using W7 models across all neural networks fit and assuming different weighting schemes. Phases are given relative to *B*-band maximum.
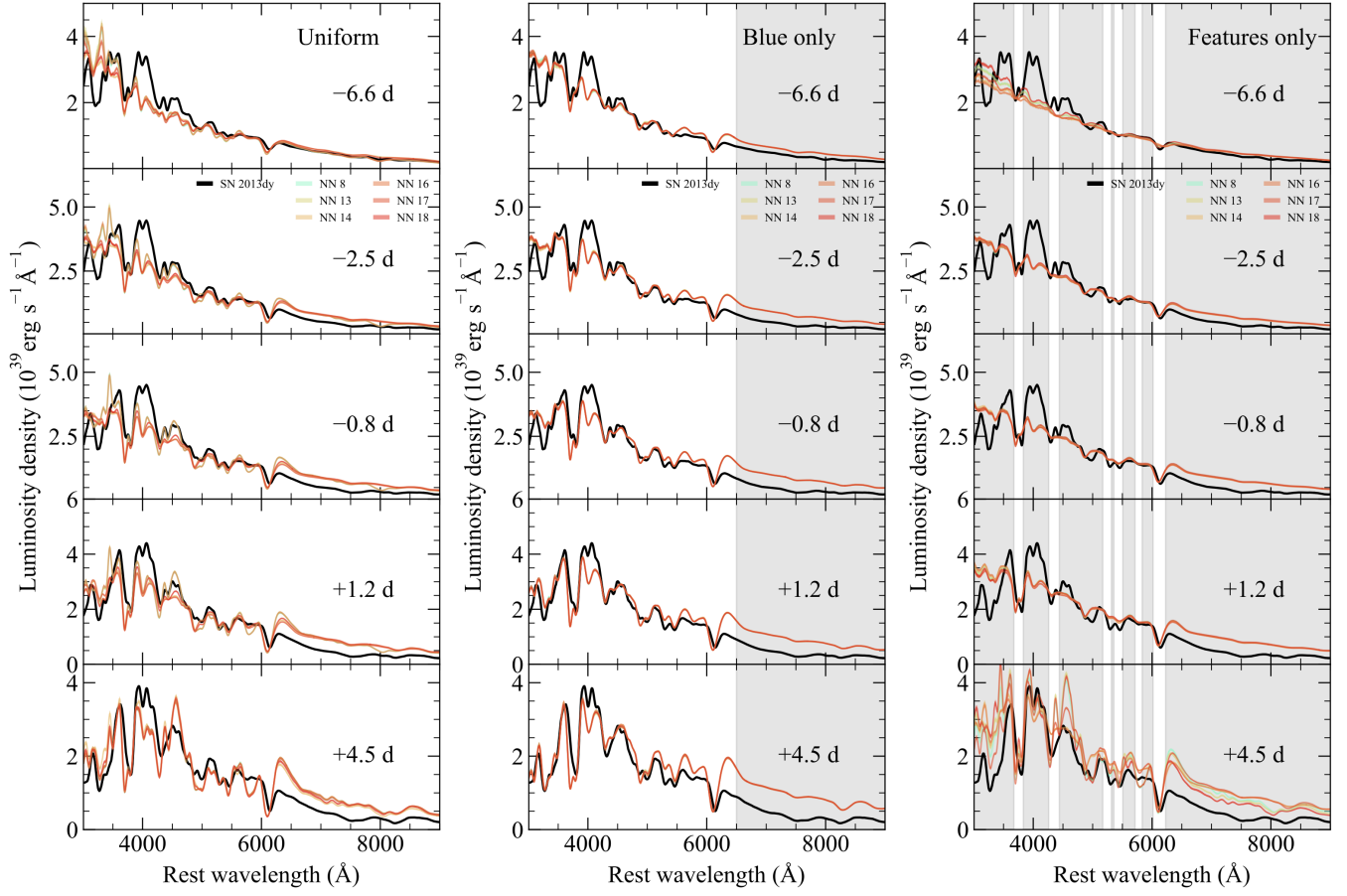
Figure D3. As in Fig. D1 for SN 2013dy and N100 models.
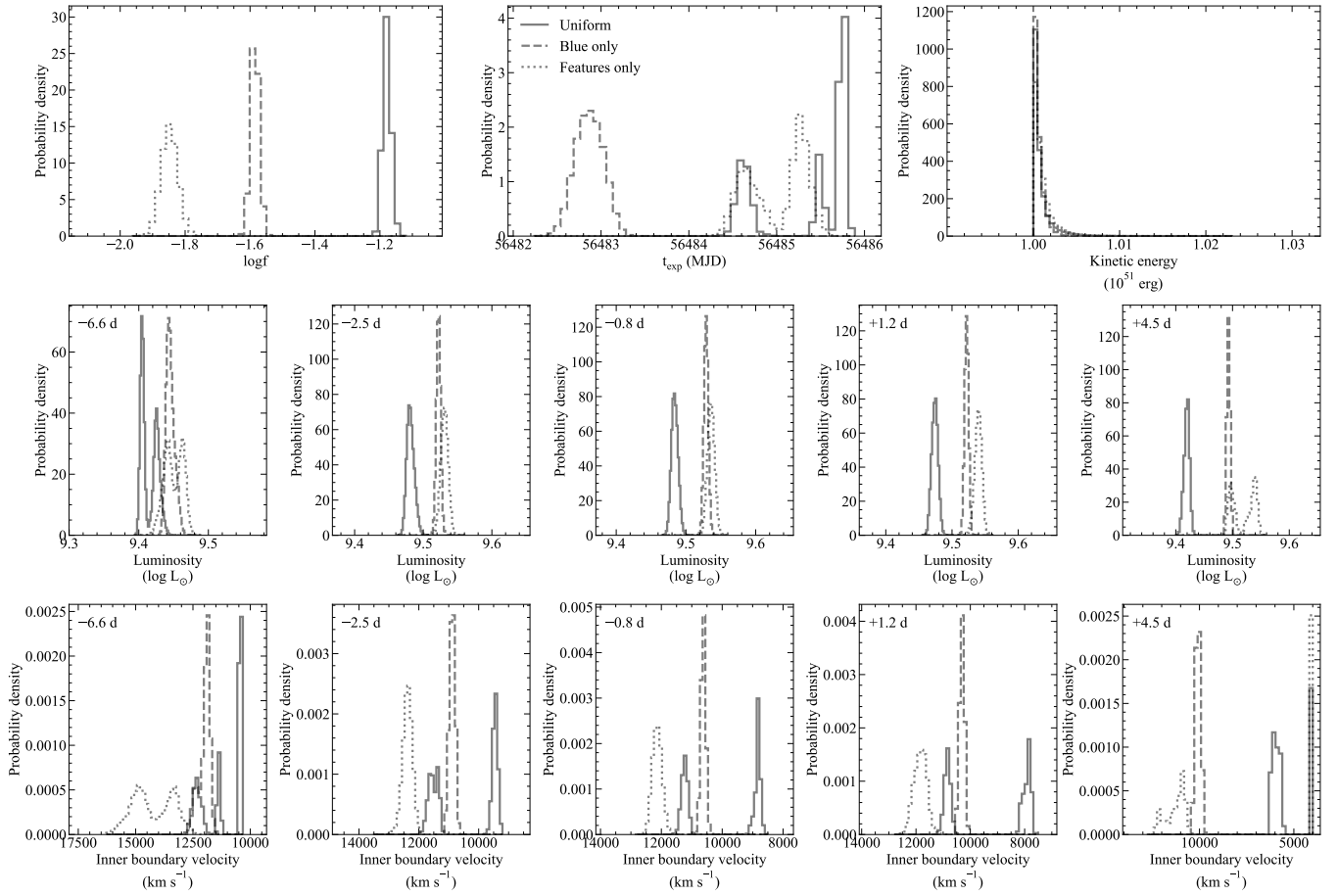
Figure D4. As in Fig. D2 for SN 2013dy and N100 models.