MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Maven: a multimodal foundation model for supernova science

To cite this article: Gemma Zhang *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 045069

View the article online for updates and enhancements.

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 045069 https://doi.org/10.1088/2632-2153/ad990d

MACHINE
LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
9 September 2024

REVISED
19 November 2024

ACCEPTED FOR PUBLICATION
28 November 2024

PUBLISHED
19 December 2024

PAPER

Maven: a multimodal foundation model for supernova science

Gemma Zhang[1,2,7,*], Thomas Helfer[3,7,*], Alexander T Gagliano[1,4,5], Siddharth Mishra-Sharma[1,2,6,8] and V Ashley Villar[1,5]

[1] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Boston, MA, United States of America
[2] Department of Physics, Harvard University, Cambridge, MA 02138, United States of America
[3] Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY 11794 United States of America
[4] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America
[5] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, MS-16, Cambridge, MA 02138, United States of America
[6] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America
[7] Both authors contributed equally to this work. Authors may be listed in either order.
[8] Currently at Anthropic; work performed while at MIT/IAIFI.
[*] Authors to whom any correspondence should be addressed.

E-mail: yzhang7@g.harvard.edu and thomashelfer@live.de

Keywords: foundation model, Contrastive learning, supernovae, astrophysics, deep learning, multimodal

Abstract

A common setting in astronomy is the availability of a small number of high-quality observations, and larger amounts of either lower-quality observations or synthetic data from simplified models. Time-domain astrophysics is a canonical example of this imbalance, with the number of supernovae observed photometrically outpacing the number observed spectroscopically by multiple orders of magnitude. At the same time, no data-driven models exist to understand these photometric and spectroscopic observables in a common context. Contrastive learning objectives, which have grown in popularity for aligning distinct data modalities in a shared embedding space, provide a potential solution to extract information from these modalities. We present Maven, the first foundation model for supernova science. To construct Maven, we first pre-train our model to align photometry and spectroscopy from 0.5 M synthetic supernovae using a contrastive objective. We then fine-tune the model on 4702 observed supernovae from the Zwicky transient facility. Maven reaches state-of-the-art performance on both classification and redshift estimation, despite the embeddings not being explicitly optimized for these tasks. Through ablation studies, we show that pre-training with synthetic data improves overall performance. In the upcoming era of the Vera C. Rubin observatory, Maven will serve as a valuable tool for leveraging large, unlabeled and multimodal time-domain datasets.

1. Introduction

The discovery rate of supernovae (SNe) has grown exponentially over the past four decades, thanks in large part to wide-field, untargeted optical surveys (e.g. All Sky Automated Survey for SuperNovae (ASAS-SN; Shappee *et al* 2014), ATLAS (Tonry *et al* 2018), the Zwicky transient facility (ZTF; Bellm *et al* 2018) and the young supernova experiment (YSE; Jones *et al* 2021). Today, well over ten-thousand SNe are discovered annually. The upcoming legacy survey of space and time (LSST; Ivezić *et al* 2019), conducted by the Vera C. Rubin Observatory, is expected to commence in 2025 and will continue for ten years. LSST will enable the photometric discovery of over one million SNe annually, in addition to millions of other non-SN variable phenomena (including flaring stars and active galactic nuclei). We expect a small fraction of LSST SNe–no more than 1%–to be observed spectroscopically.

SNe can be characterized by a spectral energy distribution (SED), the energy emitted by the event over wavelength and time. Data from this SED at a fixed time is observed as spectroscopy, and for a fixed

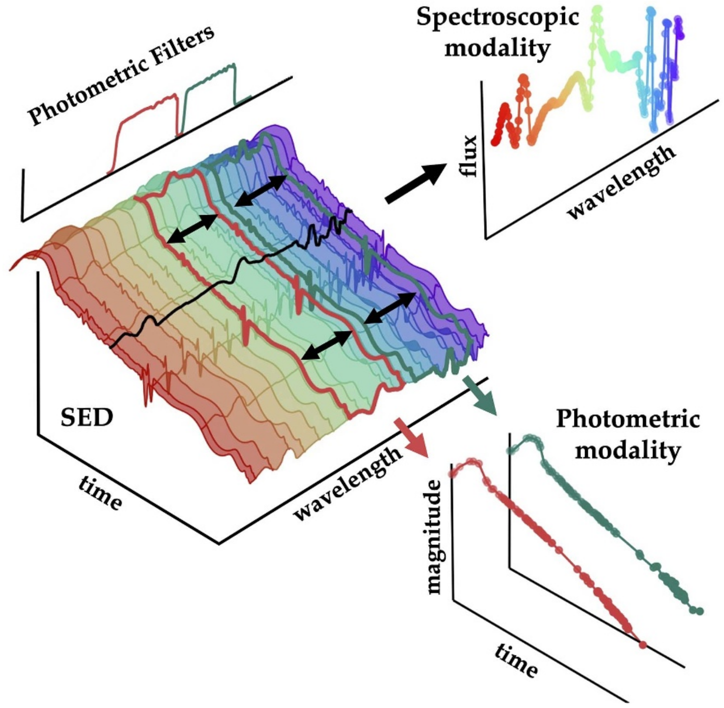

Figure 1. Illustration of the mutual information between modalities considered in this work. Supernovae are characterized by a spectral energy distribution (SED) varying in wavelength and time. Photons are collected through broad-band filters (ZTF-g and ZTF-r with transmission curves shown in the top left) at multiple epochs during explosion to construct multi-band light curves (bottom right). At a single epoch, supernova light can also dispersed with a spectrograph (SEDM in this work) to obtain a spectrum (top right). Both modalities offer a complementary, but limited, view of an SN's underlying SED.

wavelength range is observed as photometry. This photometry is collected over time to construct a SN's light curve (see figure 1). While photometry is easily obtained, spectroscopy is significantly more time-consuming to acquire (long integration times are needed to build up sufficient signal across a spectrograph). This challenge has catalyzed research into techniques to infer the underlying physics of an explosion directly from photometric observations, including the classification of SN types (e.g. Villar *et al* 2019, Muthukrishna *et al* 2019a, Möller and de Boissière 2020, Boone 2021, Gagliano *et al* 2023, Rehemtulla *et al* 2024, de Soto *et al* 2024) and inference of SN redshifts (Mitra *et al* 2023, Qu and Sako 2023). In this context, supervised machine learning has dominated the training of models for the classification of SN types and the estimation of SN redshift. The labels used in the supervised training scenario must be first extracted from spectra, demanding large spectroscopic datasets for sufficient model performance. To overcome this issue, researchers have begun to explore self-supervised learning to leverage the structure of unlabeled photometric datasets, by training a feature extraction network and generating a low-dimensional latent space (Richards *et al* 2012, Villar *et al* 2020). The learned latent space can then be used to classify events using supervised methods.

Self-supervised representation learning for time-domain astrophysics is appealing for multiple reasons. Pre-trained models have been shown to produce latent data representations that are more robust against distribution shifts than their supervised counterparts (Goyal *et al* 2022, Shi *et al* 2022). Distribution shift is a common obstacle when applying models trained on bright, spectroscopically-confirmed low-redshift transients to fainter, more distant phenomena that are underrepresented in the training data. Self-supervised learning may also be less sensitive to the class imbalances observed in transient science (Yang and Xu 2020): labeled SN samples are dominated by type Ia SNe due to their high luminosities relative to other classes. The generalizability of learned representations (Kim *et al* 2021, Ericsson *et al* 2022) also offers the potential for using a pre-trained model for multiple inference tasks and across diverse time-domain surveys, with only minimal fine-tuning.

Contrastive learning has emerged as an effective pre-training objective for combining multiple data modalities. Radford *et al* (2021) present an embedding scheme called contrastive language-image pre-training (CLIP) for aligning natural language and associated images in a shared latent space. Following this example, domain-specific 'foundation models' are beginning to appear in the literature. Parker *et al*

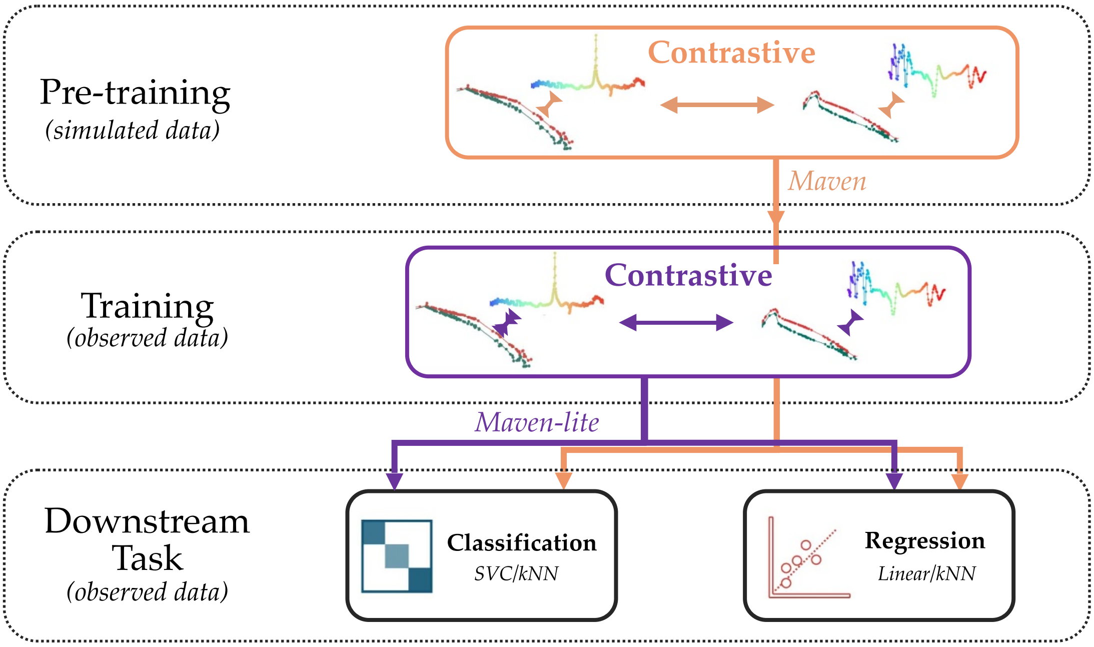

Figure 2. Overview of our training workflows. We first pre-train on a large simulated data set using contrastive methods (using light curves and spectra). We follow up by training on the observational ZTF dataset and then use a simple model to translate these embeddings to downstream tasks. Different colors indicate different first training steps and their subsequent arrows indicate subsequent training steps.

(2024) recently introduced a cross-modal foundation model using galaxy spectroscopy and images. After independently embedding galaxy images and spectra into low-dimensional latent spaces, they use contrastive training to align the embeddings into a joint latent space. They find that such a model can achieve state-of-the-art performance on the inference of various physical properties (including redshift, mass, and age). Similarly, Slijepcevic *et al* (2024) leveraged contrastive learning with instance differentiation, and created a foundational model for radio galaxies by augmenting and aligning unimodal data instances via simple transforms such as rotations. Their resulting model is able to perform accurate morphological classification with fewer labels than supervised methods.

Here, we present Maven, the first multimodal foundation model for SNe. In contrast to previous models for SN classification and redshift inference, our model is constructed using spectroscopic and photometric information simultaneously. Motivated by previous work in synthetic pre-training, we first train Maven by aligning simulated light curve-spectrum pairs via contrastive learning, and fine-tune it on a small sample of observed data using the same approach (see figure 2). Our final model achieves state-of-the-art performance on multiple downstream tasks. We also train a model with only observed data, called Maven-lite, to quantify the impact of synthetic pre-training. Though we limit our analysis to classification and redshift (two popular inference tasks in SN science), the model is a milestone toward general-purpose training for a range of downstream tasks.

Our paper is organized as follows. In section 2, we describe the simulated and observed data used in this work. In section 3, we describe the architectures of our photometric and spectroscopic encoder models, the contrastive learning objective used to pre-train and fine-tune Maven, and the downstream tasks we use to evaluate Maven's performance. We present our results in section 4, and compare our model to baseline transformer models optimized explicitly for the explored tasks. We further compare our results to existing transformer-based models from the literature. We conclude by discussing the value of contrastive pre-training in astronomy and potential future research directions in sections 5 and 6.

2. Datasets and simulations

In this study, we utilize two datasets: a simulated dataset for pre-training and a dataset of observations for subsequent fine-tuning and validation[7]. We describe the details of each below.

[7] All data are available at https://huggingface.co/datasets/thelfer/multimodal_supernovae.

2.1. Simulating supernovae with the SNANA simulation code

We generate synthetic SN samples using the SNANA simulation code (Kessler *et al* 2009). SNANA mimics the observing process beginning from a rest-frame SED of an astrophysical transient. A volumetric rate is chosen and the sky is populated at random with transients. A survey strategy, detection efficiency, and the survey's estimated noise properties (zeropoint and sky noise) are provided to construct synthetic observations.

We simulate observations of the ZTF (Bellm *et al* 2018) using the framework described in Aleo *et al* (2023), which approximately matches the redshift distribution of the SNe in our observed sample (described in the following section 2.2). We simulate 500 000 total events evenly split between five different SN classes, using SED models from the Photometric LSST Astronomical Time-Series Classification Challenge (Kessler *et al* 2019): SNe Ia (using the `SALT2` model; Guy *et al* 2007); SNe Ib/c (`SNIbc-Templates`; Kessler *et al* 2010); SLSNe-I (using the model `SLSNI-MOSFIT`; Villar *et al* 2017); and SNe II (`SNII-Templates`; Kessler *et al* 2010), which includes both SNe IIP/IIL; and SNe IIn (`SNIIn-MOSFIT`; Villar *et al* 2017). We use the same volumetric rates for SNE II, SNe IIn, and SNe Ib/c as in the PLAsTiCC challenge (Strolger *et al* 2015), re-scaled to match the fractional rates presented in Shivvers *et al* (2017). The volumetric rate for SNe Ia is taken from Hounsell *et al* (2018), and that for SLSNe-I traces the star-formation history parameterized in Madau and Dickinson (2014). Our simulations mimic the ZTF survey strategy, filter transmissions, and reported sky noise. This results in a similar selection function favoring low-redshift ($z < 0.1$) SNe as our observed sample, although we do not explicitly define a brightness threshold for photometry as is done with the BTS sample (Fremling *et al* 2020) and our sole quality cut is removing events with fewer than four total photometric observations. As a result, our simulated events are intrinsically fainter and lower-quality than our observed events.

In addition to the previously-developed simulations, we define a spectrograph object in SNANA with wavelength bins corresponding to the wavelength coverage of the ZTF SED machine (Blagorodnova *et al* 2018), with which the vast majority of our observed SNe were classified. To mimic the stochasticity inherent to SN classification in practice, we allow synthetic spectra to be obtained randomly from explosion to peak light, and with sufficient exposure time to achieve S/N of 5 within an arbitrary wavelength window. Galactic extinction is applied to both modalities at the simulated SN location following the extinction law from Cardelli *et al* (1989).

We then pre-process all spectra in the same manner as in Muthukrishna *et al* (2019b): we apply low-pass median filtering to remove high-frequency noise, re-bin the data to log-wavelength space, and estimate the flux continuum using a polynomial fit and divide it out. While this continuum-division step removes color information, it has been shown that it has a negligible impact on redshift estimation (Blondin and Tonry 2007). The spectra are kept in the observer frame (not redshift-corrected).

2.2. The ZTF bright transient survey

Since 2019, the ZTF (Bellm *et al* 2018) has conducted a wide-field public survey consisting of photometry obtained with the Palomar 48-inch Schmidt telescope at a cadence of roughly two nights. The telescope observes in three photometric filters: ZTF-*g*, ZTF-*r*, and ZTF-*i*. Photometry is promptly reduced and streamed to alert brokers including ANTARES (the Arizona-NOIRLab Temporal Analysis and Response to Events System; Matheson *et al* 2021). For non-Galactic transients observed at or expected to peak brighter than an apparent magnitude of ∼18.5, a classification spectrum is automatically obtained using the SED machine (SEDM; Ben-Ami *et al* 2012, Blagorodnova *et al* 2018, Rigault *et al* 2019), a low-resolution spectrograph mounted on the Palomar 60-inch telescope (Cenko *et al* 2006). SEDM spectra are then uploaded to the transient name server and the Weizmann interactive supernova data repository (WISeREP; Yaron and Gal-Yam 2012). 5377 SNe have been spectroscopically confirmed at the time of writing as part of this bright transient survey.

We obtain metadata for 4702 spectroscopically-classified SNe on June 18th, 2024 from the ZTF bright transient survey (Fremling *et al* 2020) after applying all quality and purity cuts available on the ZTF BTS webpage[8] (described in detail in Perley *et al* 2020). The subsequent SNe have photometric coverage before and after peak light, good visibility throughout the photospheric phase, an uncontaminated reference image, and occurred in low extinction fields. We consolidate our resulting sample to only include events spectroscopically classified as 'normal' SN Ia, SN Ib/c, SN II, SLSN-I, and SN IIn.

Next, we use the `Python` client of the `antares` alert broker (Matheson *et al* 2021) to consolidate difference photometry for all SNe in ZTF-*g* and ZTF-*r* (ZTF-*i* observations are mainly private, comprising ∼10% of all observations; Aleo *et al* 2023), and download their associated SEDM spectra from the transient

[8] https://sites.astro.caltech.edu/ztf/bts/bts.php.

name server[9] and WISEReP[10, 11]. We pre-process the observed spectra following the same procedure as our synthetic ones.

Next, we augment our observational data with noise. In each training iteration, we apply Gaussian noise to the photometric and spectroscopic observations with mean zero and standard deviation equal to the magnitude of the reported observational errors. This acts to both increase our training set and to make our model more robust to typical observational noise.

3. Methodology

3.1. Contrastive representation learning

Contrastive learning is a type of self-supervised learning based on the existence of associations between data samples. It encourages corresponding data pairs to develop similar representations while separating unassociated pairs in representation space. For multimodal datasets, contrastive learning has been a common approach for aligning data pairs across modalities. Here, our goal is to build a shared representation space using photometric and spectroscopic data from the same event, and to explore the predictive properties of these representations for downstream tasks.

For both pre-training and fine-tuning, we use the standard softmax-based bidirectional variant of the InfoNCE (Oord *et al* 2018) contrastive loss function introduced for training . Given a minibatch $\mathcal{B}$ of $|\mathcal{B}|$ associated pairs $\{(X_i, Y_i)\}_{i=1}^{|\mathcal{B}|}$(in this work, the $X_i \in I$ represents the light curve and $Y_i \in T$ represents spectrum of a single SN), the goal is to align the learned representations of corresponding (positive) pairs (X_i, Y_i) while repelling the representations of unaligned (negative) pairs $(X_i, Y_{j\neq i})$:

$$\mathcal{L}\left(\mathcal{B}\right) = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\log\frac{\mathrm{e}^{x_i\cdot y_i/\tau}}{\sum_{j=1}^{|\mathcal{B}|}\mathrm{e}^{x_i\cdot y_j/\tau}} + \log\frac{\mathrm{e}^{x_i\cdot y_i/\tau}}{\sum_{j=1}^{|\mathcal{B}|}\mathrm{e}^{x_j\cdot y_i/\tau}}\right) \tag{1}$$

where $x_i = f(X_i)/\|f(X_i)\|$ and $y_i = g(Y_i)/\|g(Y_i)\|$ are the normalized representations of the ith data pairs associated with each other, and τ is a learnable hyperparameter. Encoders $f: I \to \mathbb{R}^{d_{\mathrm{emb}}}$ and $g: T \to \mathbb{R}^{d_{\mathrm{emb}}}$ map the two modalities to an embedding space of dimension d_{emb}. Transformer-based encoders are chosen to capture and aggregate the temporal correlations of our light curve data and the wavelength correlations of our spectroscopic data. We describe these encoders in more detail in the next section. This loss treats the two representations symmetrically, thus ensuring that the two modalities are equally weighted.

3.2. Modality encoders

The encoders $f: I \to \mathbb{R}^{d_{\mathrm{emb}}}$ and $g: T \to \mathbb{R}^{d_{\mathrm{emb}}}$ are designed to efficiently extract information from high-dimensional data for the two considered modalities. Both light curve and spectrum encoders are based on the transformer architecture (Vaswani *et al* 2017). In this section, we describe the architecture and explore how common representation / pre-training approaches impact downstream task performance (see figure 3 for an overview).

The transformer-based light curve encoder processes magnitude measurements and their corresponding observation times. Given an input sequence of magnitude-time pairs $X_i = ((m_1, t_1), \ldots, (m_n, t_n)) \in I$, where t_j is defined as the number of days from the first observation, the normalized magnitudes are initially linearly projected to the d_{model}-dimensional embedding space of the transformer. Each transformer layer applies multi-head self-attention (with n_{heads} heads acting separately). Here we define $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(QK^T/\sqrt{d_k}\right)V$ where Q, K, and V are linear projections of the input representing queries, keys, and values, and $d_k = d_{\mathrm{model}}/n_{\mathrm{heads}}$. A two-layer feedforward network, $\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$, is applied to each sequence element (a magnitude-time pair) separately. Layer normalization and residual connections are applied after attention as well as the feedforward layer.

To account for the temporal nature of light curves, we use sinusoidal time encodings to project the times t_i to a higher-dimensional space,

$$\mathrm{TE}\left(t_i, j\right) = \begin{cases} \sin\left(t_i/n_t^{2j/d_{\mathrm{model}}}\right) & \text{if } i \text{ is even} \\ \cos\left(t_i/n_t^{2j/d_{\mathrm{model}}}\right) & \text{if } i \text{ is odd} \end{cases}, \tag{2}$$

[9] www.wis-tns.org/.

[10] www.wiserep.org/.

[11] Despite spectroscopic classifications being available on the ZTF website for all listed SNe, SEDM spectra could not be found for a few objects. When an SEDM spectrum is not available, we instead use the first reported spectrum. A positional encoding is used for the wavelengths of each spectrum, so in principle our spectrum encoder has the capacity to generalize to multiple spectrographs.

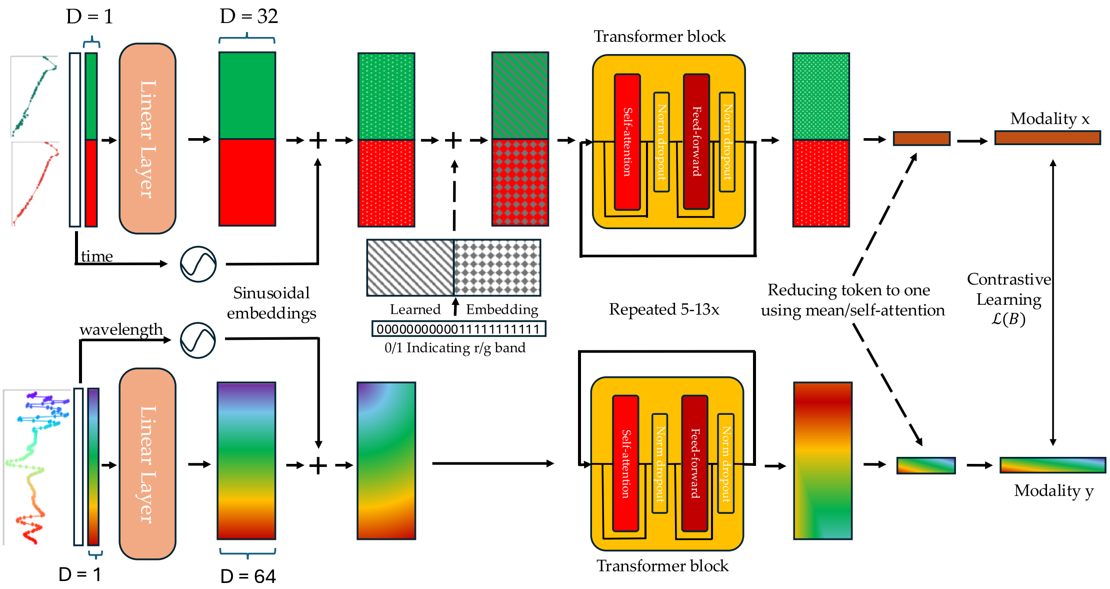

Figure 3. Overview over model architecture. We start by using a linear layer to increase the dimensionality of the features from one to the chosen embedding dimension. Next, we encode both time and wavelength using sinusoidal embeddings. For light curves specifically, we additionally include embeddings to represent the band information. These combined tokens are then processed through a series of transformer blocks. To reduce the output to a single token, we use a simple mean or self-attention mechanism. Finally, we project this token to match the desired length of the embedding space.

where j is the time embedding index, t_i are the input times, and n_t is a hyperparameter governing the periodicity of the time encodings. This encoding allows the model to capture both absolute and relative timing of observations across a wide range of timescales.

To incorporate light curve measurements from multiple photometric filters, we concatenate all measurements for each SN and add an additional band encoding. Different bands are one-hot encoded with integers and then added to light curve magnitude embeddings before being passed into the transformer encoder.

In contrast, the spectrum encoder processes flux measurements across multiple wavelengths. It utilizes a similar transformer-based architecture to that of the light curve encoder, but interprets the input sequence as $((f_1, \lambda_1), \ldots, (f_n, \lambda_n))$, where f_i represents the flux at observer-frame wavelength λ_i. The positional encoding for wavelengths follows the same sinusoidal pattern as the light curve encoder, but with λ in place of t. This approach allows the model to capture both local and global spectral correlations.

For both the light curve and spectrum encoders, in addition to deterministic aggregate e.g. mean or max pooling, we consider attention-based learnable aggregation to convert the per-sequence representation to a 1-D representation vector. This enables the model to learn a data-dependent aggregation scheme, potentially better capturing correlations in the data. We initialize a learnable query vector $Q_{\text{learned}} \in \mathbb{R}^{d_{\text{emb}}}$, where d_{emb} is the embedding dimension. A projection of the encoded sequence after the final transformer layer, $X_{\text{final}} \in \mathbb{R}^{n_{\text{seq}} \times d_{\text{seq}}}$ gives the keys and values for the attention mechanism. We use a multi-head attention architecture with two heads to then get $x_{\text{agg}} = \text{Attention}(Q_{\text{learned}}, K_{\text{final}}, V_{\text{final}}) \in \mathbb{R}^{d_{\text{emb}}}$ as desired. This attention-based pooling allows the model to focus on the most relevant parts of the sequence when creating the final embedding. We treat the aggregation method as a hyperparameter: in the hyperparameter tuning process discussed in section 3.5, we consider both mean and attention-based aggregation.

3.3. Transfer learning and fine-tuning

After pre-training some of our models on the simulations discussed in section 2.1, we fine-tune all weights on the small set of ZTF BTS measurements discussed in section 2.2. We define our best-performing hyperparameter-optimized pre-trained model as 'Maven', and our best-performing hyperparameter-optimized model without pre-training as 'Maven-lite' (see figure 2).

3.4. Stratified k-fold cross-validation

To quantify uncertainties for both end-to-end and fine-tuned models, we perform a five-fold cross-validation, in which we split the ZTF dataset into five unique train-test splits. All five folds share the

Table 1. Table of hyperparameters that were optimized in a random hyperparameter search. Parameters in the leftmost column were optimized for end-to-end models while parameters in the rightmost column were optimized for all models (including end-to-end training and finetuning).

Light curve / spectrum encoder	Metadata encoder[a]	Optimizer and training
Number of transformer blocks	Dimension of class label embedding	Learning rate
Number of attention heads	Number of layers in MLP	Dropout rate
Normalization factor in time encoding (n_t)	MLP hidden layer dimension	Weight decay
Aggregation method		Logit temperature in loss function (τ)
Embedding vector dimension		Batch size

[a] In addition to light curve and spectra, we explored using metadata as an additional modality. Details about the metadata encoder and results are shown in appendix.

Table 2. Overview of model hyperparameters. The column starting with 'sp' refer to the spectral transformer parameters, while those starting with 'lc' refer to the light curve transformer parameters. The term 'blocks' indicates the number of transformer blocks, 'head' refers to the number of attention heads, and 'emb' specifies the embedding dimension. The 'agg' column describes the aggregation method of tokens at the end, where 'mean' denotes computing the mean over tokens and 'attn' indicates using self-attention. For full reproducibility, the YAML files defining the models are available in the GitHub repository along with pre-trained models (https://github.com/ThomasHelfer/multimodal-supernovae).

Model	sp-blocks	sp-head	sp-emb	sp-agg	lc-blocks	lc-head	lc-emb	lc-agg
Maven	13	2	32	mean	5	8	64	mean
Maven-lite	13	2	32	mean	5	8	64	attn
Baseline classification					9	2	32	mean
Baseline regression					9	2	32	mean

same distribution of SN classes. The results in subsequent sections are the mean and standard deviation from these runs. To avoid added computational overhead, we do not perform cross-validation on the much larger simulation-based pre-training dataset.

3.5. Hyperparameter optimization

To determine hyperparameter values for model architecture and training, we perform a hyperparameter search for our end-to-end baseline and CLIP models using Weights & Biases (Biewald 2020). Table 1 provides a summary of the hyperparameters tuned in this process. A list of parameter values in our search are provided in configuration files in our public code repository[12]. An overview over some of the hyperparameters can be found in table 2.

In each hyperparameter sweep, we choose the set of parameter values that result in the lowest validation loss on our holdout dataset. Due to the high computational cost associated with hyperparameter tuning, we employ a random train-test split on our dataset instead of carrying out k-fold cross-validation. In addition, we reuse the optimal hyperparameters of the Maven-lite model for Maven instead of performing a separate hyperparameter search. In the transfer learning stage of the pre-trained models discussed in section 3.3, we only tune the hyperparameters shown in the optimizer and training column of table 1, while other hyperparameters are fixed to their pre-train values.

3.6. Downstream tasks

We evaluate the performance of Maven and Maven-lite on two primary downstream tasks: classification and regression.

Classification of SNe from photometry *alone* has been an area of active study due to the long integration times required to build up sufficient signal-to-noise with spectroscopy and the subsequent rise of wide-field photometric surveys. SN classes are highly imbalanced in observed samples, due to a combination of different intrinsic volumetric rates and a steep selection function toward brighter classes (SNe Ia). We separately consider both five-way (SN Ia, SN II, SN Ib/c, SLSN-I, SN IIn) and three-way classification (SN Ia, SN II, SN Ib/c), considering in the latter case only the three most commonly-observed classes.

In addition to classification, we attempt to predict the redshift of each SN (which we call our 'regression' task). Redshift estimation using spectroscopic and photometric SNe Ia is a fundamental tool for cosmological analyses. Although non-Ia classes are significantly more observationally diverse (e.g. Modjaz

[12] https://github.com/ThomasHelfer/multimodal-supernovae.

Table 3. Classification performance for three classes by model configuration: This table presents the classification performance of various models using light curve data from the ZTF dataset. The models are categorized based on whether they utilized simulation pre-training (‘pre-trained’), the type of last layer added to embedding models (‘last-layer’). The modalities taken into account when training on the real ZTF dataset are indicated in ‘real-pre’ (lc—light curve, sp—spectrum, m—metadata) as well as whether a SVC or *k*NN. Performance metrics include macro-F1 (mac-f1), micro-F1 (mic-f1), macro-precision (mac-p), and macro-recall (mac-r). The results are presented as mean ± standard deviation, calculated over five folds. Baseline models, trained in an end-to-end supervised fashion using only the ZTF data, are included for comparison.

pre-trained	last-layer	real-pre	mac-f1	mac-p	mac-r
No	end-to-end baseline		0.7011 ± 0.0303	0.6934 ± 0.0360	0.7527 ± 0.0247
Yes	*k*NN	lc-m	0.6920 ± 0.0217	0.7286 ± 0.0377	0.6721 ± 0.0183
Yes	*k*NN	lc-sp	0.6874 ± 0.0342	0.8041 ± 0.0833	0.6516 ± 0.0216
Yes	*k*NN	lc-sp-m	0.6849 ± 0.0194	0.7280 ± 0.0334	0.6643 ± 0.0161
Yes	SVC	lc-m	0.6747 ± 0.0297	0.8026 ± 0.0257	0.6435 ± 0.0257
Yes	SVC	lc-sp-m	0.6522 ± 0.0237	0.7892 ± 0.0975	0.6247 ± 0.0215
No	*k*NN	lc-sp-m	0.6268 ± 0.0251	0.7204 ± 0.0701	0.6000 ± 0.0199
No	*k*NN	lc-sp	0.6265 ± 0.0231	0.6670 ± 0.0532	0.6119 ± 0.0121
No	*k*NN	lc-m	0.6249 ± 0.0228	0.7309 ± 0.0661	0.6035 ± 0.0184
Yes	SVC	lc-sp	0.6195 ± 0.0190	0.7753 ± 0.0994	0.6056 ± 0.0172
No	SVC	lc-m	0.5971 ± 0.0220	0.7871 ± 0.1858	0.5842 ± 0.0163
No	SVC	lc-sp-m	0.5938 ± 0.0156	0.7892 ± 0.1873	0.5802 ± 0.0077
No	SVC	lc-sp	0.5749 ± 0.0099	0.5857 ± 0.0126	0.5686 ± 0.0102

Table 4. Regression performance by model configuration: this table presents the regression performance of various models using light curve data from the ZTF dataset. The models are categorized based on whether they utilized simulation pre-training (‘pre-trained’), the type of last layer added to embedding models (‘last-layer’). The modalities taken into account when training on the real ZTF dataset is indicated in ‘real-pre’ (lc—light curve, sp—spectrum, m—metadata) as well weather we use a linear or *k*NN layer to translate our embedding to a redshift prediction (‘last-layer‘). Performance metrics include the coefficient of determination (R^2), L1 loss, and L2 loss. The results are presented as mean ± standard deviation, calculated over five folds. Baseline models, trained in an end-to-end supervised fashion using only the ZTF data, are included for comparison.

pre-trained	last-layer	real-pre	R^2	L1	L2
Yes	*k*NN	lc-m	0.6543 ± 0.0280	0.0094 ± 0.0005	0.0152 ± 0.0010
Yes	Linear	lc-sp-m	0.6513 ± 0.0440	0.0096 ± 0.0005	0.0152 ± 0.0016
Yes	*k*NN	lc-sp	0.6496 ± 0.0398	0.0095 ± 0.0004	0.0152 ± 0.0014
Yes	*k*NN	lc-sp-m	0.6470 ± 0.0422	0.0094 ± 0.0006	0.0152 ± 0.0012
Yes	Linear	lc-sp	0.6386 ± 0.0447	0.0099 ± 0.0003	0.0155 ± 0.0016
Yes	Linear	lc-m	0.6345 ± 0.0444	0.0100 ± 0.0006	0.0156 ± 0.0014
No	*k*NN	lc-m	0.6150 ± 0.0294	0.0103 ± 0.0003	0.0160 ± 0.0012
No	end-to-end baseline		0.6129 ± 0.0245	0.0104 ± 0.0004	0.0160 ± 0.0010
No	*k*NN	lc-sp-m	0.6090 ± 0.0464	0.0102 ± 0.0005	0.0161 ± 0.0015
No	*k*NN	lc-sp	0.6078 ± 0.0408	0.0103 ± 0.0006	0.0161 ± 0.0014
No	Linear	lc-sp	0.5948 ± 0.0402	0.0107 ± 0.0007	0.0164 ± 0.0015
No	Linear	lc-sp-m	0.5938 ± 0.0450	0.0108 ± 0.0004	0.0164 ± 0.0016
No	Linear	lc-m	0.5927 ± 0.0399	0.0107 ± 0.0004	0.0165 ± 0.0015

et al 2019), estimating SN redshift remains critical for estimating the intrinsic properties of an explosion (luminosity from photometry and chemical composition from spectroscopy).

To transform our contrastive-trained light curve embeddings into classification predictions, we explore both support vector classification (SVC) and *k*-nearest neighbors classification (*k*NN). SVC works by finding an optimal hyperplane to separate classes. Here, we use a linear kernel with `scikit-learn` default parameters. *k*NN classification, in contrast, classifies SNe based on the similarity of their feature embedding to other latent-space neighbors.

For redshift regression, we explore both linear regression and *k*NN regression. The former uses linear transformation of the embeddings to estimate redshift, while the latter estimates redshift based on the average (or median) redshift of closest training examples in the latent space.

In our comparison, we find *k*NN to be best-performing for both regression and classification. A more comprehensive comparison over our experiments can be found in tables 3 and 4. In the following sections, we mainly quote results from the best performing *k* value for brevity but note that we also experimented with multiple *k*NN classifiers.

Lastly, we train transformer-based supervised models directly on the observational ZTF dataset as our baseline models. For the classification baseline model, we optimize for the multi-class cross-entropy loss and take the class with highest pseudo-probability score as the prediction for each event in the validation set. The regression baseline model outputs a single value and is optimized using the mean squared error (MSE) loss. The hyperparameters of our baseline models are given in table 2.

4. Results

In this section, we present results from Maven, Maven-lite, and our baseline models on the downstream tasks.

4.1. t-SNE visualization of latent spaces

To explore the impact of contrastive pre-training on the latent space of our Maven models, we visualize a sample of embedded light curves. We compute Maven and Maven-lite embeddings of our five dominant classes for both the synthetic and observed samples: SNe Ia, SNe II, SNe Ib/c, SLSNe I, and SNe IIn. Similar to Slijepcevic *et al* (2024), we first reduce the dimensionality of our latent space using principal component analysis from the encoder output of 128 features to 50 features. This allows us to explore the impact tSNE perplexity on the observed latent structure while managing computational overhead. We have confirmed that the 50 resultant principal components retain >99.999% of the variance in the original embeddings. We then produce two-dimensional representations of these embeddings using the t-distributed stochastic neighbor embedding tool (t-SNE; Van der Maaten and Hinton 2008). Our results are presented in figure 4 for Maven-lite (left column) and Maven (right column), where the embeddings are colored by class in the top row and shaded by redshift in the bottom row.

Significant differences are visible between the two latent spaces. Considering the Maven-lite embeddings, only the synthetic SLSN-I light curves (blue) are well-separated from the other classes; the core-collapse (SN Ib/c, II, IIn) and thermonuclear (Ia) events show significant overlap. Observed Ia and II light curves (outlined in black) show similar embeddings independent of class, and little consistency with the synthetic embeddings: the majority of observed SN Ia and SN II lie at the boundary between synthetic SLSN-I and SN II/SN IIn embeddings.

In our Maven embeddings, we observe both clear separation of classes and consistent redshift gradients across our embedded light curves. The simulated SNe Ia appear best-organized by redshift, consistent with their photometric homogeneity. The redshift gradient across observed SNe Ia is also well-aligned with that of the synthetic sample, whereas a similar distribution is not observed in the Maven-lite embeddings. Synthetic SNe Ib/c appear strongly mixed with both SNe Ia and SNe II, indicative of the photometric degeneracies between these classes.

Interestingly, although observed SN Ia and SN II embeddings lie closest to the synthetic events of the same class, the overlap between synthetic and observed data remains low. We attribute this to a distributional shift between synthetic and observed data. Observed events are prioritized for spectroscopic confirmation if they are brighter than (or expected to brighten above) $m < 18.5$th magnitude, and additional quality and purity cuts are imposed (see section 3 for details). While a detailed comparison between synthetic and observed events is beyond the scope of this work, this separation may also reflect the simplistic nature of our simulations relative to reality, and emphasizes the need for significantly larger observed SN samples for effective pre-training.

4.2. Classification performance

Our results are visualized using a set of confusion matrices for our three-way classification task in figure 5. We show the confusion matrices for precision (normalized by predicted class) and recall (normalized by true class) for our models. Precision is the proportion of true positives out of all positive predictions made by the model, while recall is the proportion of true positive predictions out of all true positive instances. We note higher recall by Maven on the two dominant classes in our sample: 0.79 for SNe II and 0.99 for SNe Ia, compared to 0.74 for SNe II and 0.91 for SNe Ia with the baseline model. We observe poorer recall with the minority SN Ib/c class, which comprises ∼5% of the observed sample: 0.18 with simulated pre-training compared to 0.61 for the baseline. We predict that the baseline model is better able to outline the decision boundaries for this class.

We observe the opposite results on the minority class when considering class precision. Our Maven model achieves comparable precision to the baseline for SNe II and SNe Ia but substantially higher precision for SNe Ib/c, 0.58 compared to 0.28. We note that, with substantially higher discovery rates of rare classes anticipated with the Vera C. Rubin observatory, classification precision is essential for obtaining spectroscopic follow-up observations of relevant events. We have explored the misassociation rate as a function of event peak brightness, but identify no obvious correlations.

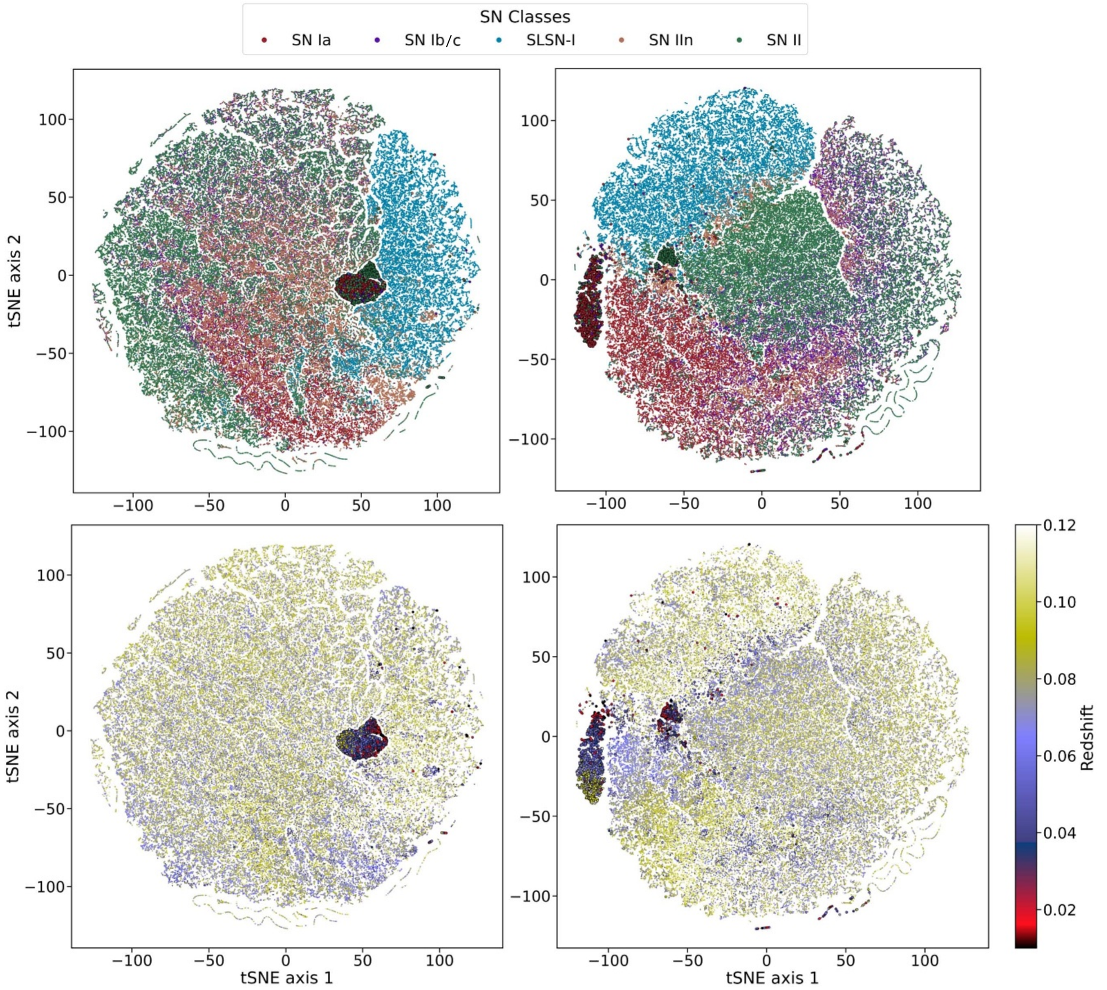

Figure 4. Visualization of synthetic and observed light curves embedded by Maven-lite (left column) and Maven (right column). Points in top row are colored by SN class, and points in bottom row are shaded by spectroscopic redshift. Observed data are outlined in black.

A common metric in classification tasks is the F_1 score, which increases to 1 in the limit of perfect classification. For a class C, F_1 is defined as the harmonic mean between the class's recall r and precision p:

$$F_{1,C} := 2\frac{p_C \times r_C}{p_C + r_C}. \tag{3}$$

We calculate for each model both the micro-averaged F_1 score, which averages performance across all events irrespective of class; and the macro-averaged F_1 score, which averages the F_1 score computed independently for each class. The macro-averaged F_1 score is a valuable indicator for our use case given the significant class imbalance, as the micro-F_1 can approach unity when all events are labeled as the dominant class. We present these results, along with the macro-averaged precision and recall ('mac-p' and 'mac-r') in table 5. We further show the macro-F_1 score of each model as a bar plot in figure 6.

We observe macro-F_1 scores within 1-σ of the baseline model for the majority of the pre-trained kNN classifiers that we experimented with (see section 3.6), from a score of 0.6874 ± 0.0342 for Maven compared to a baseline of 0.7011 ± 0.0303. The scores for the pre-trained models are systematically higher than those without pre-training: we found an average F_1 score of 0.68 for all pre-trained kNN classifiers compared with an average of 0.63 for the kNN classifiers trained with only observed data.

We have also calculated the performance of our models for the five-way classification task, which additionally considers the rarer classes SN IIn and SLSN I. Here, we observe a marginally higher average F_1 score for the synthetic pre-trained contrastive models than the baseline, though the results are consistent to within one standard deviation (0.50 ± 0.03 for the best model compared to 0.49 ± 0.04). As with the three-way classification results, the macro-averaged precision of our pre-trained models is on average higher than the end-to-end baseline, with the best model achieving a score of 0.58 ± 0.03 compared to the baseline of 0.50 ± 0.09.

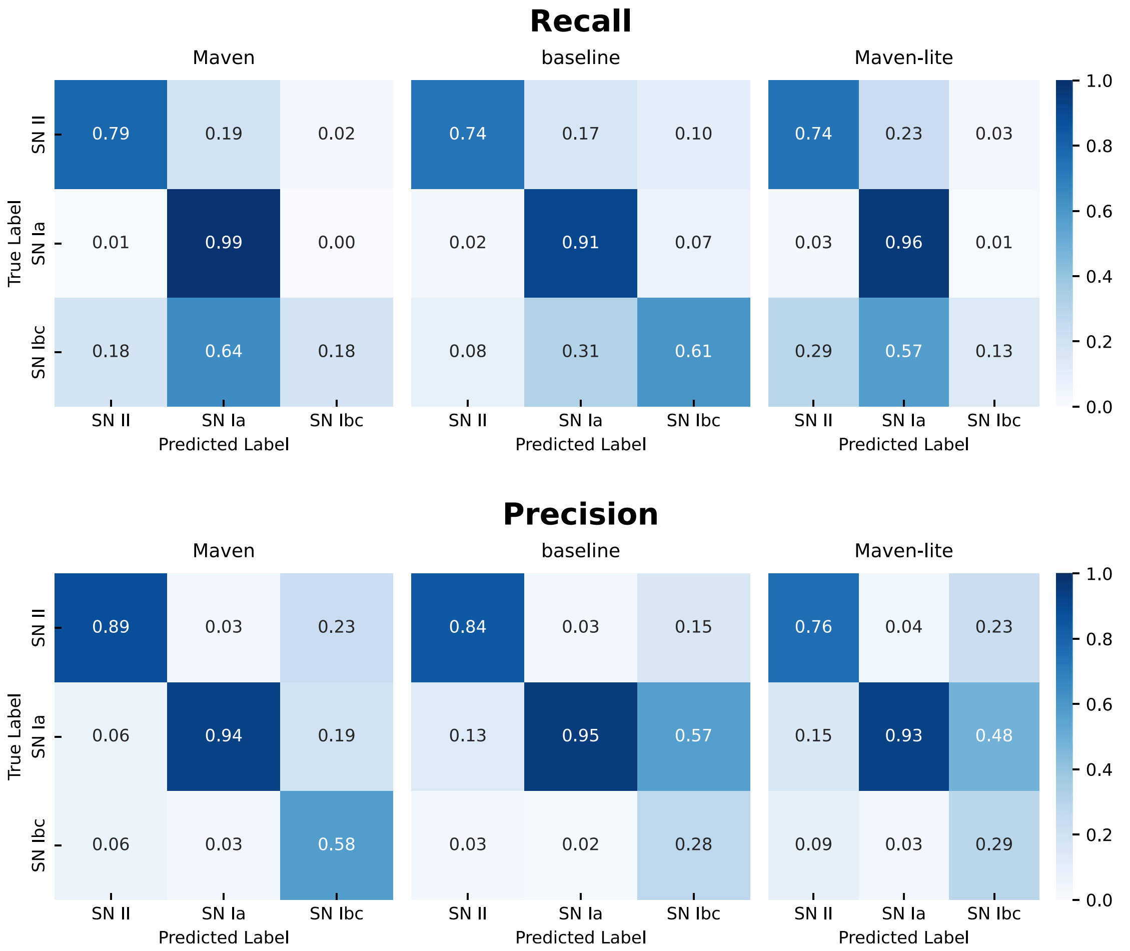

Figure 5. Normalized precision and recall confusion matrices for supernova classification across different models and modalities. The models compared are: (a) contrastive learning with real data and pre-trained on simulated data (Maven), (b) baseline using a supervised model approach, and (c) contrastive training with real data and no pre-training (Maven-lite). The classes included are SN II, SN Ia, and SN Ib/c.

Table 5. Overview of classification model performance. We present three classification models: the baseline only trained on the ZTF dataset, Maven-lite without synthetic pre-training, and Maven with synthetic pretraining and observed fine-tuning. A more comprehensive overview over the runs performed in this paper can be found in table 3.

Name	pre-trained	*k*NN	mac-F_1	mic-F_1	mac-p	mac-r
Baseline	No	—	**0.7011 ± 0.0303**	0.8728 ± 0.0205	0.6934 ± 0.0360	**0.7527 ± 0.0247**
Maven	Yes	8	0.6874 ± 0.0342	**0.9247 ± 0.0070**	**0.8041 ± 0.0833**	0.6516 ± 0.0216
Maven-lite	No	3	0.6265 ± 0.0231	0.8943 ± 0.0110	0.6670 ± 0.0532	0.6119 ± 0.0121

Note: Bold indicates best performing model for each performance metric.

4.3. Comparison to three-way photometric classifiers on ZTF SNe

Next, we compare our multimodal model to photometric classifiers in the literature that have been validated on ZTF light curves. de Soto *et al* (2024) developed a gradient-boosted machine trained on best-fit features from a flexible piecewise parametric light curve model. The resulting classifier, Superphot+, was trained on ZTF photometry for 6123 spectroscopically confirmed SNe. Variants of the classifier trained with and without redshift information were considered. 66% of these events pass the ZTF-imposed quality cuts and are also used in this work; the class breakdown of both samples are comparable for SNe Ia, II, and Ib/c. de Soto *et al* (2024) report (from their figure 18, in the way of four-way classification including SLSNe I) mean recall values of 0.77 for SNe II, 0.88 for SNe Ia, and 0.86 for SNe Ib/c when considering redshift information. By comparison, Maven achieves mean recall values of 0.79, 0.99, and 0.18 for the same respective classes. From our confusion matrices in figure 5, we observe that our lower recall on the minority class (SN Ib/c) is due to a systematic misclassification of these events as SNe Ia. The explicit inclusion of redshift information is likely to bring performance gains distinguishing these populations, as SNe Ia are ∼2 magnitudes brighter at peak Richardson *et al* (2014). de Soto *et al* (2024) further report mean precision values of 0.88 for SNe II,

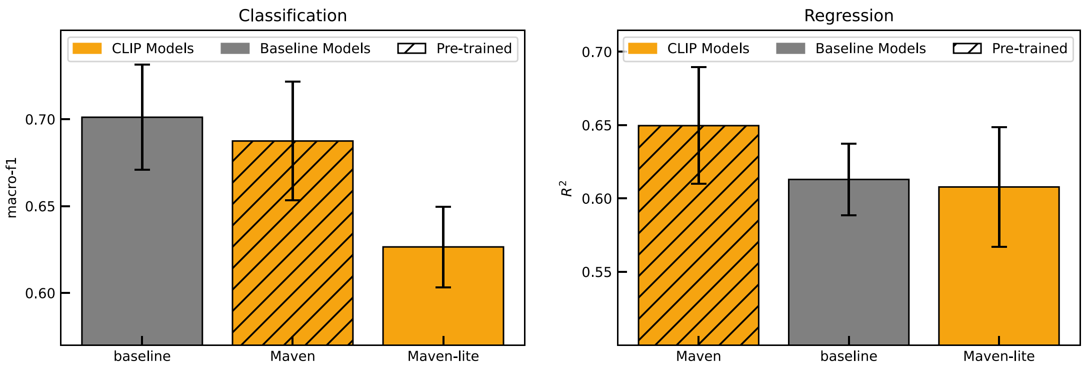

Figure 6. Final performance metrics for Maven, Maven-lite, and baseline models for on downstream classification and regression tasks.

Table 6. Overview of regression model performance. We present three regression models: the baseline only trained on the ZTF dataset, a contrastive model trained only on the ZTF dataset (Maven-lite) and a contrastive model pre-trained on simulated data (Maven) and then subsequently trained on ZTF. A more comprehensive overview over the runs performed in this paper can be found in table 4.

Name	pre-trained	*k*NN	R^2	L1	L2	OLF
Maven	Yes	9	$\mathbf{0.6496 \pm 0.0398}$	$\mathbf{0.0095 \pm 0.0004}$	$\mathbf{0.0152 \pm 0.0014}$	0.0002 ± 0.0005
Baseline	No		0.6129 ± 0.0245	0.0104 ± 0.0004	0.0160 ± 0.0010	0.0002 ± 0.0005
Maven-lite	No	9	0.6078 ± 0.0408	0.0103 ± 0.0006	0.0161 ± 0.0014	0.0002 ± 0.0005

Note: Bold indicates best performing model for each performance metric.

0.97 for SNe Ia, and 0.30 for SNe Ib/c; The classification precision for Maven is comparable for SNe II (0.89) and marginally lower for SNe Ia (0.94), and twice as high for SNe Ib/c (0.58).

The results of Pimentel *et al* (2023) are more directly comparable to this work. Pimentel *et al* (2023) present a transformer model for ZTF photometry in which the time of each observation is encoded as the phase from first detection using a Fourier decomposition-based temporal modulation, with noise added to the values in training to prevent overfitting. In a two-stage pre-training scheme with both synthetic and observed events, the optimization problem is defined with reconstruction and cross-entropy regularization terms to preserve class-specific information in the encoded light curves. The resulting 'S-TimeModAttn' model is trained and validated on g and r-band light curves from the ZTF, with presumably substantial overlap with the observational dataset considered in this work. Though Pimentel *et al* (2023) is trained on substantially fewer events (1978 compared to our 4702), we have confirmed that the population of ZTF SNe discovered before 2023 as might have been used for training is indistinguishable from our larger sample in the distribution of both redshift and photometric properties (rise time, peak apparent brightness, and decline time). Pimentel *et al* (2023) report a macro-F_1 of 0.614 ± 0.036 in the task of four-way classification (Ia, II, Ib/c, and SLSN), compared to our 0.6874 ± 0.0342 for three-way classification; and a macro-precision of 0.598 ± 0.030 compared to our 0.804 ± 0.083. A macro-recall (also referred to as completeness) score of 0.72 for three-way classification can be inferred from their confusion matrices, compared to our lower 0.6516 ± 0.0216 in figure 5. Class-specific F_1 scores and precisions (also referred to as purity) are not reported.

Cabrera-Vives *et al* (2024) apply a custom transformer model (ATAT) to synthetic photometry and metadata from the extended LSST astronomical time-series classification challenge (ELAsTiCC[13]). The ATAT model consists of separate transformers, one which encodes light curves with a temporal encoding based on Fourier series and a quantile tokenizer for extracted photometric features (including the number and phases of non-detections and the flux characteristics of detections). The dataset used to train ATAT is distinct from the dataset considered here, preventing a direct comparison of classification performance.

4.4. Regression performance

We next consider the task of redshift estimation. We quantify the performance of our models with the coefficient of determination R^2, the L1 and L2 error, and the outlier fraction 'OLF', defined as $|z_{\mathrm{pred}} - z_{\mathrm{true}}|/(1 + z_{\mathrm{true}}) > 0.15$. We report these values for contrastive pre-trained models in table 6.

[13] https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/.

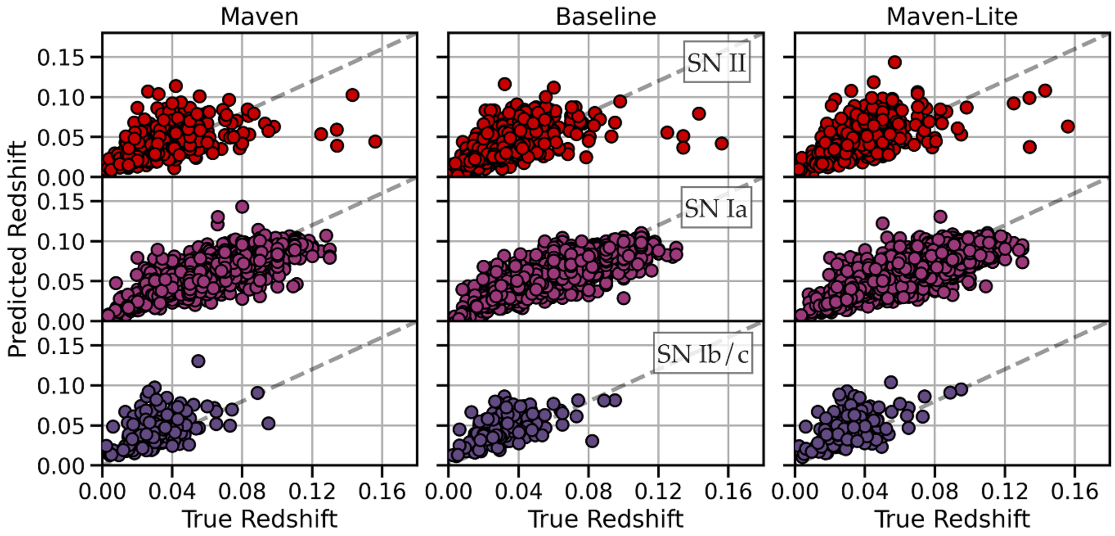

Figure 7. Predicted versus true redshift for three different models: our pre-trained Maven (left), a baseline model using supervised methods, and our Maven-lite model trained end-to-end on ZTF observations. The data is segregated into different SN classes. The SN classes are displayed in separate columns, and the models are shown in separate rows for clear comparison.

We also present a bar plot of the R^2 values in figure 6, and the predicted versus true redshifts for each SN class in figure 7. As expected, we observe the highest correlation between observed and predicted redshifts for SNe Ia, the most observationally homogeneous SN class considered. We calculate an R^2 value of $R^2 = 0.6496 \pm 0.0398$ for Maven compared to the end-to-end baseline performance of $R^2 = 0.6129 \pm 0.0245$. The L1 and L2 errors are also lower on average for Maven than for our regression baseline, while the outlier fraction is comparable. We conclude that, on average, Maven outperforms the baseline regression model. Maven-lite, our model without pre-training, achieves an R^2 value of 0.6078 ± 0.0408, lower than both Maven and the baseline model.

Though a comparable photometric redshift model for low-redshift ZTF SNe does not exist in literature, an outlier fraction of 0.004 is reported for 289 photometric SNe Ia in the supernova legacy survey (SNLS), nearly an order of magnitude higher than our best model but with a substantially higher maximum redshift $z < 1.0$ (Palanque-Delabrouille *et al* 2010). Another analytic photometric redshift estimator proposed by Wang *et al* (2015) for SNe Ia discovered by LSST finds an outlier fraction of 0.0023 over $z < 1.0$, compared to our 0.0002.

5. Discussion

We have explored the value of contrastive pre-training in constructing a foundational model for SN science. By first training with synthetic events and fine-tuning with observed events, we have constructed a model, Maven, whose performance on the downstream tasks of photometric classification and redshift is on par with models optimized end-to-end for these tasks. Maven outperforms our classification baseline model, with a micro-averaged-$F1$ score of 0.92. Similarly, Maven outperforms our baseline for redshift regression, with an L2-loss of 0.015 and minimal outlier fraction. While we have limited our study to ZTF data, adapting Maven to incorporate additional photometric filters and classes of astronomical transients would allow us to repurpose it for diverse time-domain studies with the Vera C. Rubin observatory using fewer computational resources than building multiple specialized models.

Contrastive pre-training has been proposed as a simple and effective mechanism for extracting information from multiple modalities in a single model. The following conditions need to be met for multimodal contrastive learning to be effective: that significant information content is shared across these modalities; that the mutual information is relevant for the downstream tasks; and that the shared information is the maximal information in each modality relevant for the downstream tasks. Recently, Liang *et al* (2023) formalized this picture by defining 'multi-view redundancy' as a necessary condition for effective pre-training using traditional contrastive learning. In our case, we know spectra to be highly informative for both classification (the taxonomy is *defined* by spectra obtained early after a SN's explosion, with the temporal evolution of the explosion rarely considered) and redshift inference, which is achieved primarily through the identification of spectral lines. Supernova photometry, although containing some spectral information, is significantly more lossy: the collection of photons through a broadband photometric filter destroys valuable information about a supernova's underlying SED that might otherwise be valuable for these tasks (as seen in the diagram in figure 1). For these reasons, we can characterize supernova light curves as an 'information-poor' modality and spectra as an 'information-rich' modality for our tasks. Contrastive pre-training, in this case, is unable to bring significant performance gains beyond end-to-end optimized models. This behavior persists despite aligning these modalities directly with the relevant downstream information (metadata of an event's spectroscopic classification and redshift, as discussed in appendix): *the least-informative modality sets an upper limit on the mutual information that can be extracted.*

Seen from this perspective, it is surprising that we do not observe substantial *drops* in performance relative to our baseline models. We attribute this to systematic hyperparameter tuning and synthetic pre-training, through which we are able to mitigate the negative effects of contrastive alignment. We therefore advise caution in the use of multimodal contrastive pre-training, which should be specialized for the input modalities and the anticipated downstream tasks. In our case, additional improvements may be possible with a pre-training scheme designed to preserve both mutual and unique information content relevant for classification and redshift estimation (as is proposed in Liang *et al* 2023).

6. Conclusion

We conclude by summarizing our key findings:

1. We train Maven through self-supervised contrastive learning on SN spectra and light curves. Maven is able to achieve state-of-the-art performance on both redshift estimation and SN classification.
2. We find that pre-training on a large simulated dataset significantly improves Maven's performance on downstream tasks over a contrastively-trained model on solely the observed data.
3. Maven does not dramatically outperform supervised models optimized directly for each downstream task. We hypothesize that this is due to the light curve being an information-poor modality, which limits the amount of information our unsupervised objective is able to extract.

Starting in 2025, the Vera C. Rubin observatory will initiate the 10-year legacy survey for space and time, and detect $\sim$1 M SNe yr^{-1} in *ugrizY* filters. This consistently-calibrated photometric dataset will enable self-supervised pre-training for time-domain foundation models (including variable stars, lensing events, active galactic nuclei, and other non-SN phenomena) at an unprecedented scale. However, without spectroscopy for the vast majority of detected events, the self-supervised tasks that can be applied with this data will be limited to a single modality.

On the other hand, traditional multimodal models have considered distinct representations of a single astronomical object (photometry and spectroscopy of a supernova). Where spectroscopic *and* photometric information for a transient is sparse, however, broad physical properties can be inferred from the event's host galaxy (Hakobyan *et al* 2012, Kang *et al* 2020, Schulze *et al* 2021, Chakraborty *et al* 2024). Early efforts have emphasized the value of these data for photometric classification (Gomez *et al* 2020, Carrasco-Davis *et al* 2021, Gagliano *et al* 2023, Sheng *et al* 2024). LSST data will contain photometry for tens of billions of galaxies, millions of which will be spectroscopically-confirmed through the dark energy spectroscopic instrument (DESI; Levi *et al* 2019) or 4MOST (Dumayne *et al* 2023). Additional work should be dedicated to exploring the linking of modalities spanning distinct lengthscales, which would allow for both supernova and host-galaxy data to be consolidated in a single pre-training scheme.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://huggingface.co/datasets/thelfer/multimodal_supernovae and https://github.com/ThomasHelfer/multimodal-supernovae.

Acknowledgments

This work was initiated at the IAIFI AstroML Hackathon, held at MIT in January 2024. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/) and AST-2108676. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under Grant Contract Number DE-SC0012567. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation Grant PHY-2210452. Research reported in this publication was supported by a Postdoctoral Fellowship at the

Institute for Advanced Computational Science, Stony Brook University. This work was carried out at the Advanced Research Computing at Hopkins (ARCH) core facility (rockfish.jhu.edu), which is supported by the National Science Foundation (NSF) Grant Number OAC1920103, as well as the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Software: Jupyter (Kluyver *et al* 2016), *Matplotlib* (Hunter 2007), *Numpy* (Harris *et al* 2020), PyTorch (Paszke *et al* 2019), *PyTorch lightning* (Falcon and The PyTorch Lightning team 2019), *Astropy* (Astropy Collaboration *et al* 2013, 2018, 2022), *Einops* (Rogozhnikov 2022) *Pandas* (pandas development team 2020), *scikit-learn* (Pedregosa *et al* 2011) and *wandb* (Biewald 2020).

Appendix. Metadata as modality in contrastive learning

In addition to SN spectrum and light curve measurements, we also considered SN metadata as an additional modality for training a contrastive model. The metadata modality used in our training includes supernovae redshifts and class labels. We encode each class label with a learnable embedding vector. The metadata encoder consists of a multilayer perceptron (MLP) that takes in the concatenated vector of class embedding and redshift and outputs the final embedding. The number of hidden layers and the hidden layer dimension in the MLP were chosen from a hyperparameter search, as discussed in section 3.5.

The models which directly align event photometry with relevant metadata (redshift and class) in pre-training do not significantly outperform the models in which photometry and spectroscopy alone are aligned. Considering only pre-trained models for the task of classification, we observe comparable three-way macro-F_1 scores when aligning light curves with metadata (0.692 ± 0.022), light curves with spectra (0.687 ± 0.034), and light curves with both spectra and metadata (0.685 ± 0.019). Each of our contrastive objectives featured photometry as a modality, and we predict that this more information-poor modality is driving the observed performance across each of these models, as we discuss in additional detail in section 6.

ORCID iDs

Gemma Zhang https://orcid.org/0000-0002-8019-8082
Thomas Helfer https://orcid.org/0000-0001-6880-1005
Alexander T Gagliano https://orcid.org/0000-0003-4906-8447
Siddharth Mishra-Sharma https://orcid.org/0000-0001-9088-7845
V Ashley Villar https://orcid.org/0000-0002-5814-4061

References

Aleo P D *et al* 2023 The young supernova experiment data release 1 (YSE DR1): light curves and photometric classification of 1975 supernovae *Astrophys. J.* **266** 9

Astropy Collaboration *et al* 2013 Astropy: a community Python package for astronomy *Am. Acad. Ped.* **558** A33

Astropy Collaboration *et al* 2018 The astropy project: building an open-science project and status of the v2.0 core package *Astron. J.* **156** 123

Astropy Collaboration *et al* 2022 The astropy project: sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package *Astrophys. J.* **935** 167

Bellm E C *et al* 2018 The zwicky transient facility: system overview, performance and first results *Publ. Astron. Soc. Pac.* **131** 018002

Ben-Ami S, Konidaris N, Quimby R, Davis J T, Ngeow C C, Ritter A and Rudy A 2012 *Ground-Based and Airborne Instrumentation for Astronomy IV* vol 8446 (SPIE) pp 1044–52

Biewald L 2020 Experiment tracking with weights and biases (available at: www.wandb.com/)

Blagorodnova N *et al* 2018 The SED machine: a robotic spectrograph for fast transient classification *Publ. Astron. Soc. Pac.* **130** 035003

Blondin S and Tonry J L 2007 Determining the type, redshift and age of a supernova spectrum *Astrophys. J.* **666** 1024

Boone K 2021 ParSNIP: generative models of transient light curves with physics-enabled deep learning *Astron. J.* **162** 275

Cabrera-Vives G *et al* 2024 ATAT: astronomical transformer for time series and tabular data (arXiv:2405.03078)

Cardelli J A, Clayton G C and Mathis J S 1989 The relationship between infrared, optical and ultraviolet extinction *Astrophys. J.* **345** 245

Carrasco-Davis R *et al* 2021 Alert classification for the alerce broker system: the real-time stamp classifier *Astron. J.* **162** 231

Cenko S B *et al* 2006 The automated Palomar 60 inch telescope *Publ. Astron. Soc. Pac.* **118** 1396

Chakraborty S *et al* 2024 Type Ia supernova progenitor properties and their host galaxies *Astrophys. J.* **969** 80

de Soto K M *et al* 2024 Superphot+: realtime fitting and classification of supernova light curves (arXiv:2403.07975)

Dumayne J *et al* 2023 Using 4MOST to refine the measurement of galaxy properties: a case study of supernova hosts *RAS Tech. Instrum.* **2** 453

Ericsson L, Gouk H, Loy C C and Hospedales T M 2022 Self-supervised representation learning: introduction, advances and challenges *IEEE Signal Process. Mag.* **39** 42

Falcon W and The PyTorch Lightning team 2019 PyTorch Lightning, 1.4 (available at: https://github.com/Lightning-AI/lightning)

Fremling C *et al* 2020 The zwicky transient facility bright transient survey. I. Spectroscopic classification and the redshift completeness of local galaxy catalogs *Astrophys. J.* **895** 32

Gagliano A, Contardo G, Foreman-Mackey D, Malz A I and Aleo P D 2023 First impressions: early-time classification of supernovae using host-galaxy information and shallow learning *Astrophys. J.* **954** 6

Gomez S, Berger E, Blanchard P K, Hosseinzadeh G, Nicholl M, Villar V A and Yin Y 2020 FLEET: a redshift-agnostic machine learning pipeline to rapidly identify hydrogen-poor superluminous supernovae *Astrophys. J.* **904** 74

Goyal P, Duval Q, Seessel I, Caron M, Misra I, Sagun L, Joulin A and Bojanowski P 2022 Vision models are more robust and fair when pretrained on uncurated images without supervision (arXiv:2202.08360)

Guy J *et al* 2007 SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators *Am. Acad. Ped.* **466** 11

Hakobyan A A, Adibekyan V Z, Aramyan L S, Petrosian A R, Gomes J M, Mamon G A, Kunth D and Turatto M 2012 Supernovae and their host galaxies. I. The SDSS DR8 database and statistics *Am. Acad. Ped.* **544** A81

Harris C R *et al* 2020 Array programming with NumPy *Nature* **585** 357

Hounsell R *et al* 2018 Simulations of the WFIRST supernova survey and forecasts of cosmological constraints *Astrophys. J.* **867** 23

Hunter J D 2007 Matplotlib: a 2D graphics environment *Comput. Sci. Eng.* **9** 90

Ivezić Ž *et al* 2019 LSST: from science drivers to reference design and anticipated data products *Astrophys. J.* **873** 111

Jones D *et al* 2021 The young supernova experiment: survey goals, overview and operations *Astrophys. J.* **908** 143

Kang Y, Lee Y-W, Kim Y-L, Chung C and Ree C H 2020 Early-type host galaxies of type Ia supernovae. II. Evidence for luminosity evolution in supernova cosmology *Astrophys. J.* **889** 8

Kessler R *et al* 2009 SNANA: a public software package for supernova analysis *Publ. Astron. Soc. Pac.* **121** 1028

Kessler R *et al* 2010 Results from the supernova photometric classification challenge *Publ. Astron. Soc. Pac.* **122** 1415

Kessler R *et al* 2019 Models and simulations for the photometric LSST astronomical time series classification challenge (PLAsTiCC) *Publ. Astron. Soc. Pac.* **131** 094501

Kim D, Park S, Kim J and Lee J 2021 SelfReg: self-supervised contrastive regularization for domain generalization (arXiv:2104.09841)

Kluyver T *et al* 2016 *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed F Loizides and B Schmidt (IOS Press) pp 87–90

Levi M *et al* 2019 *Bull. Am. Astron. Soc.* **51** 57

Liang P P, Deng Z, Ma M, Zou J, Morency L-P and Salakhutdinov R 2023 Factorized contrastive learning: going beyond multi-view redundancy (arXiv:2306.05268)

Madau P and Dickinson M 2014 Cosmic star-formation history *Annu. Rev. Astron. Astrophys.* **52** 415

Matheson T *et al* 2021 The ANTARES astronomical time-domain event broker *Astron. J.* **161** 107

Mitra A, Kessler R, More S and Hlozek R LSST Dark Energy Science Collaboration 2023 Using host galaxy photometric redshifts to improve cosmological constraints with type Ia supernovae in the LSST era *Astrophys. J.* **944** 212

Modjaz M, Gutiérrez C P and Arcavi I 2019 New regimes in the observation of core-collapse supernovae *Nat. Astron.* **3** 717

Möller A and de Boissière T 2020 SuperNNova: an open-source framework for Bayesian, neural network-based supernova classification *Mon. Not. R. Astron. Soc.* **491** 4277

Muthukrishna D, Narayan G, Mandel K S, Biswas R and Hložek R 2019a RAPID: early classification of explosive transients using deep learning *Publ. Astron. Soc. Pac.* **131** 118002

Muthukrishna D, Parkinson D and Tucker B E 2019b DASH: deep learning for the automated spectral classification of supernovae and their hosts *Astrophys. J.* **885** 85

Oord A v d, Li Y and Vinyals O 2018 Representation learning with contrastive predictive coding (arXiv:1807.03748)

Palanque-Delabrouille N *et al* 2010 Photometric redshifts for type Ia supernovae in the supernova legacy survey *Am. Acad. Ped.* **514** A63

Pandas development team, T. 2020 pandas-dev/pandas: Pandas, latest, *Zenodo* (available at: https://doi.org/10.5281/zenodo.3509134)

Parker L *et al* 2024 AstroCLIP: a cross-modal foundation model for galaxies *Mon. Not. R. Astron. Soc.* **531** 4990

Paszke A *et al* 2019 *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) pp 8024–35

Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825

Perley D A *et al* 2020 The zwicky transient facility bright transient survey. II. A public statistical sample for exploring supernova demographics *Astrophys. J.* **904** 35

Pimentel O, Estévez P A and Förster F 2023 Deep attention-based supernovae classification of multiband light curves *Astron. J.* **165** 18

Qu H and Sako M 2023 Photo-zSNthesis: converting type Ia supernova lightcurves to redshift estimates via deep learning *Astrophys. J.* **954** 201

Radford A *et al* 2021 *Int. Conf. on Machine Learning* (PMLR) pp 8748–63

Rehemtulla N *et al* 2024 The zwicky transient facility bright transient survey. III. `BTSbot`: automated identification and follow-up of bright transients with deep learning (arXiv:2401.15167)

Richards J W, Homrighausen D, Freeman P E, Schafer C M and Poznanski D 2012 Semi-supervised learning for photometric supernova classification *Mon. Not. R. Astron. Soc.* **419** 1121

Richardson D, Jenkins R L I, Wright J and Maddox L 2014 Absolute-magnitude distributions of supernovae *Astron. J.* **147** 118

Rigault M *et al* 2019 Fully automated integral field spectrograph pipeline for the SEDMachine: pysedm *Astron. Astrophys.* **627** A115

Rogozhnikov A 2022 *Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id = oapKSVM2bcj)

Schulze S *et al* 2021 The palomar transient factory core-collapse supernova host-galaxy sample. I. Host-galaxy distribution functions and environment dependence of core-collapse supernovae *Astrophys. J.* **255** 29

Shappee B *et al* 2014 *American Astronomical Society Meeting Abstracts* vol 223 pp 236–03

Sheng X, Nicholl M, Smith K W, Young D R, Williams R D, Stevance H F, Smartt S J, Srivastav S and Moore T 2024 NEural engine for discovering luminous events (NEEDLE): identifying rare transient candidates in real time from host galaxy images *Mon. Not. R. Astron. Soc.* **531** 2474

Shi Y, Daunhawer I, Vogt J E, Torr P and Sanyal A 2022 *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*

Shivvers I *et al* 2017 Revisiting the lick observatory supernova search volume-limited sample: updated classifications and revised stripped-envelope supernova fractions *Publ. Astron. Soc. Pac.* **129** 054201

Slijepcevic I V, Scaife A M M, Walmsley M, Bowles M, Wong O I, Shabala S S and White S V 2024 Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning *RAS Tech. Instrum.* **3** 19

Strolger L-G, Dahlen T, Rodney S A, Graur O, Riess A G, McCully C, Ravindranath S, Mobasher B and Shahady A K 2015 The rate of core collapse supernovae to redshift 2.5 from the CANDELS and CLASH supernova surveys *Astrophys. J.* **813** 93

Tonry J, Denneau L, Heinze A, Stalder B, Smith K, Smartt S, Stubbs C, Weiland H and Rest A 2018 ATLAS: a high-cadence all-sky survey system *Publ. Astron. Soc. Pac.* **130** 064505

Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605 (available at: http://jmlr.org/papers/v9/vandermaaten08a.html)

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need (arXiv:1706.03762)

Villar V A *et al* 2020 SuperRAENN: a semisupervised supernova photometric classification pipeline trained on Pan-STARRS1 medium-deep survey supernovae *Astrophys. J.* **905** 94

Villar V A, Berger E, Metzger B D and Guillochon J 2017 Theoretical models of optical transients. I. A broad exploration of the duration–luminosity phase space *Astrophys. J.* **849** 70

Villar V *et al* 2019 Supernova photometric classification pipelines trained on spectroscopically classified supernovae from the Pan-STARRS1 medium-deep survey *Astrophys. J.* **884** 83

Wang Y, Gjergo E and Kuhlmann S 2015 Analytic photometric redshift estimator for type Ia supernovae from the large synoptic survey telescope *Mon. Not. R. Astron. Soc.* **451** 1955–63

Yang Y, and Xu Z 2020 Rethinking the value of labels for improving class-imbalanced learning (arXiv:2006.07529)

Yaron O and Gal-Yam A 2012 WISeREP–an interactive supernova data repository *Publ. Astron. Soc. Pac.* **124** 668