The Astrophysical Journal, 974:169 (32pp), 2024 October 20 https://doi.org/10.3847/1538-4357/ad6a4f

Superphot+: Real-time Fitting and Classification of Supernova Light Curves

Kaylee M. de Soto[1], V. Ashley Villar[1,2], Edo Berger[1,2], Sebastian Gomez[1,3], Griffin Hosseinzadeh[4], Doug Branton[5], Sandro Campos[6], Melissa DeLucchi[6], Jeremy Kubica[6], Olivia Lynn[6], Konstantin Malanchev[6], and Alex I. Malz[6]
[1] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138-1516, USA
[2] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, USA
[3] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[4] Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA
[5] DiRAC Institute and the Department of Astronomy, University of Washington, 3910 15th Avenue NE, Seattle, WA 98195, USA
[6] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
Received 2024 March 13; revised 2024 June 24; accepted 2024 July 12; published 2024 October 10

Abstract

Photometric classifications of supernova (SN) light curves have become necessary to utilize the full potential of large samples of observations obtained from wide-field photometric surveys, such as the Zwicky Transient Facility (ZTF) and the Vera C. Rubin Observatory. Here, we present a photometric classifier for SN light curves that does not rely on redshift information and still maintains comparable accuracy to redshift-dependent classifiers. Our new package, Superphot+, uses a parametric model to extract meaningful features from multiband SN light curves. We train a gradient-boosted machine with fit parameters from 6061 ZTF SNe that pass data quality cuts and are spectroscopically classified as one of five classes: SN Ia, SN II, SN Ib/c, SN IIn, and SLSN-I. Without redshift information, our classifier yields a class-averaged F_1-score of 0.61 ± 0.02 and a total accuracy of 0.83 ± 0.01. Including redshift information improves these metrics to 0.71 ± 0.02 and 0.88 ± 0.01, respectively. We assign new class probabilities to 3558 ZTF transients that show SN-like characteristics (based on the ALeRCE Broker light-curve and stamp classifiers) but lack spectroscopic classifications. Finally, we compare our predicted SN labels with those generated by the ALeRCE light-curve classifier, finding that the two classifiers agree on photometric labels for $82\% \pm 2\%$ of light curves with spectroscopic labels and $72\% \pm 0\%$ of light curves without spectroscopic labels. Superphot+ is currently classifying ZTF SNe in real time via the ANTARES Broker, and is designed for simple adaptation to six-band Rubin light curves in the future.

Unified Astronomy Thesaurus concepts: Supernovae (1668); Nested sampling (1894); Light curve classification (1954); Transient detection (1957)

Materials only available in the online version of record: machine-readable table

1. Introduction

Currently, ∼10,000 supernova-like (SN-like) transients are photometrically detected every year by the Zwicky Transient Facility (ZTF; Bellm et al. 2019), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016), and the Asteroid Terrestrial-impact Last Alert System (Tonry et al. 2018), among other facilities. Global resources can spectroscopically follow up on ∼10% of these transients. Wide-field surveys planned for this decade, including the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; Tyson 2002) and the Nancy Grace Roman Space Telescope High Latitude Time Domain Survey (Rose et al. 2021), are expected to increase annual SN detections by a factor of ∼100. With spectroscopic resources not expected to increase exponentially in the same time frame, 99.9% of new SNe light curves will lack traditional *spectroscopic* classifications (see, e.g., Filippenko 1997 for review).

In response to this limitation, several works have implemented algorithms that classify SNe using only photometric information (Muthukrishna et al. 2019; Villar et al. 2019, 2020; Hosseinzadeh et al. 2020; Boone 2021; Sánchez-Sáez et al. 2021), or a combination of photometry and host galaxy information (Gomez et al. 2020a, 2023a, 2023b; Gagliano et al. 2023; Kisley et al. 2023; Sheng et al. 2024). Many of these classifiers show successful performance with simulated light curves (e.g., RAPID. Muthukrishna et al. 2019; ParSNIP, Boone 2021). However, simulated light curves typically lack the observed population diversity in real data sets; it is therefore challenging to predict classifier performance on real data (Aleo et al. 2023). Boone (2021) highlights a particular failure mode, in which a classifier is able to distinguish between SNe II simulated from discrete models. Among classifiers that do train on real data (e.g., Superphot, Villar et al. 2019; Hosseinzadeh et al. 2020 on Pan-STARRS data; SuperRAENN, Villar et al. 2020, on Pan-STARRS data; FLEET, Gomez et al. 2020a, 2023a, 2023b, on ZTF and Open Supernova Catalog data; GHOST, Gagliano et al. 2023, on SDSS-II, ESSENCE, and SNLS data; Kisley et al. 2023, on THEx data), only the Automatic Learning for the Rapid Classification of Events (ALeRCE; Sánchez-Sáez et al. 2021) pipeline, Fink (Leoni et al. 2022), and the NEural Engine for Discovering Luminous Events (NEEDLE; Sheng et al. 2024), all trained on ZTF data, are currently being run in real time with publicly accessible predictions. Furthermore, the latter two are tuned to isolate specific classes of interest (SNe Ia for Fink, tidal disruption events, hereafter TDEs, and superluminous SNe for NEEDLE).

Thus, we aim to design a publicly available, multiclass classifier that is trained on real data. This pipeline should be

designed for easy adaptation to Rubin light curves in the future. Of particular concern is the current requirement for spectroscopic redshift information for the vast majority of photometric classifiers (with the exception of FLEET, NEEDLE, and both of ALeRCE's classifiers). Even with new spectroscopic surveys (e.g., 4MOST; De Jong et al. 2019), which will obtain spectroscopic redshifts for millions of galaxies, only a small fraction of the $\sim 10^{10}$ galaxies detected by LSST (LSST Science Collaboration et al. 2009) will have associated spectroscopic redshift information. Additionally, according to the LSST Science Requirements Document (Ivezić & LSST Science Collaboration 2018), Rubin is not expected to meet its minimum target criteria for accurate photometric redshifts within its first 3 yr of operation (Graham et al. 2018; Kessler et al. 2019). This lack of reliable redshifts for early Rubin observations, along with a preference for very dim host galaxies among exotic SNe (e.g., Type I SLSNe; Hsu et al. 2024), necessitates an SN classifier that does not use redshift information. Instead, we rely only on the light-curve shape and color to differentiate between SN classes.

Here, we present the SN classification pipeline Superphot +, which (1) empirically fits SN light curves to a parametric model and (2) trains a gradient-boosted machine (GBM) classifier on the best-fit model parameters. Superphot+ improves on the Superphot (Villar et al. 2019; Hosseinzadeh et al. 2020) pipeline by accelerating fitting for real-time light-curve processing, improving class reweighting, and enabling classification without redshift information. In this work, we train and apply Superphot+ to ZTF light curves observed through 2023 October, although we emphasize Superphot+'s adaptability for other photometric data sets. This paper is organized as follows. In Section 2, we describe the selection and pruning of our training and test data sets. We describe the details of the light-curve fitting and choice of sampling algorithm in Section 3. In Section 4, we describe feature selection (including and excluding redshift information), optimization of the classifier architecture, and oversampling of the training data set. In Section 5, we summarize multiclass and binary classifier performance with and without redshift information, emphasizing accuracy as a function of classification confidence. We also consider performance on partial light curves for real-time classification through the ANTARES Broker (Narayan et al. 2018; Matheson et al. 2021). In Section 6, we compare Superphot+'s performance without redshift information to that of the ALeRCE (Sánchez-Sáez et al. 2021) light-curve classifier, one of the only comparable redshift-independent classifiers currently available in the literature. We also compare our training results using redshifts with those from previous pipelines that require redshift. In Section 7, we use our trained classifier to assign photometric labels to 3558 ZTF SN-like transients, which lack spectroscopic classification but were labeled likely SNe by ALeRCE's light-curve and stamp classifiers. We compare our photometric predictions with those from ALeRCE. Finally, we discuss conclusions and avenues for future work in Section 8. Our code is publicly available on GitHub[7] and the Python Package Index as `superphot-plus`. All data sets are available on Zenodo (DOI:10.5281/zenodo.12519870; de Soto et al. 2024).

[7] https://github.com/VTDA-Group/superphot-plus

2. Data Set Generation

2.1. Photometric Data from the Zwicky Transient Facility

This work uses light-curve data from ZTF (Bellm et al. 2019), a wide-field survey conducted with a 48 inch telescope located at Palomar Observatory. ZTF photometrically identifies ∼5000–10,000 new (likely) extragalactic transients every year, ∼20% with spectroscopic classifications. Both ZTF public surveys (the Northern Sky Survey and the Galactic Plane Survey) image a combined ≈25,000 deg^2 of the northern sky at a high cadence of ∼2 days, in both the g and r bands.

To train our photometric classifier, we first collate SNe that (1) have associated ZTF light curves of sufficient quality (according to metrics detailed in Section 2.3) and (2) are spectroscopically labeled as one of the output classes listed in Section 2.4. We refer to this set as the training or "spectroscopic" data set interchangeably throughout the paper. To create this training set, we first query the Transient Name Server (TNS; Gal-Yam 2021) for all spectroscopically classified transients with ZTF internal names and photometry.[8] This query yields 9526 events that were registered between 2013 March 19 and 2023 October 11. While most of this set are SNe, there are also non-SN transients, like active galactic nuclei (AGN) and TDEs, which are pruned as described in Section 2.4. We keep all TNS classes in our data set for the time being to determine the classes (both SN and non-SN) that are not used directly for training but have sufficient "high-quality" light curves to merit further analysis in Section 5.4. We then download the g- and r-band light curves for these events through a Python API maintained by ALeRCE (Förster et al. 2021)'s LSST (currently ZTF) alert broker.

2.2. Data Preprocessing

Superphot+'s parametric function captures light curves in flux (rather than magnitudes). We thus convert our photometry from magnitudes to fluxes using $f = 10^{-0.4(m-\mathrm{zp})}$ and a constant zero-point of $\mathrm{zp} = 26.3$, which is halfway between the r-band and g-band median zero-points as stated by the ZTF collaboration (Masci et al. 2018, 2020). We correct all resulting light curves for Milky Way extinction. To do so, we adopt $E(B-V)$ from the dust maps provided by Schlegel et al. (1998) and Schlafly & Finkbeiner (2011), and use a Fitzpatrick & Massa (2007) extinction model with $R_V = 3.1$. We neglect host galaxy extinction, which is difficult to calculate from light-curve data alone.

ZTF alert data are calculated from difference imaging and point-spread function photometry to measure the time-variable flux of transients. However, poor template subtraction can lead to a complex background, causing light curves to asymptote well above zero flux and creating spurious detections after the transient events (and potentially before, if forced photometry is included). Our ZTF light curves do not include forced photometry, so we only see false detections at the tail end of the light curves. While we could add a constant term to our parametric model to account for a flux offset, we find that including such an offset leads to oversubtraction among SNe with long plateaus. Therefore, we instead filter these spurious data points before fitting. Following the procedure described in

[8] We note that some spectroscopic classifications may differ between TNS and subsequent analysis papers; for simplicity, we only use the primary label from TNS.

Appendix A, we clip the tail ends of 4324 (45%) light curves to some extent.

Because we do *not* utilize redshift in the primary version of Superphot+, we do not convert light curves into rest frame. When incorporating redshift information into our classifier in Section 5.5, we add redshifts and *k*-corrected peak absolute magnitudes as additional input features rather than altering the light curves or their derived fit parameters.

2.3. Data Set Pruning

To refine our spectroscopic data set before model fitting and training, we exclude light curves that fail to satisfy certain criteria. First, we keep only light curves with at least five points of signal-to-noise ratio (SNR) greater than 3 in each of the *g* and *r* bands (*after* clipping light-curve tails). This number is selected so all light curves have either (1) somewhat constrained fit parameters across the entire light curve or (2) strongly constrained fits in at least one portion of the light curve, depending on the sampling cadence. This cut on the number of observations removes 2488 (26%) light curves from the data set, leaving 7038 events. Additionally, we include only light curves whose brightness variability in both bands cannot solely be attributed to measurement error. Quantitatively, we remove transients whose maximum amplitude in either band is less than 3 times that band's mean flux uncertainty. We also remove transients in which the standard deviation of all fluxes in a single band is less than that band's mean flux uncertainty. These two cuts eliminate 486 (7%) light curves from the remaining spectroscopic set. This is a smaller fraction than is removed by the observational cut, consistent with ZTF registering an event as a transient only after sufficient brightening relative to the template flux.

A summary of our data quality cuts on all TNS classes in our data set is shown in Table 1. We are left with 6552 transients from all spectroscopic classes. We note that longer-duration transients (such as SLSNe and TDEs) have much smaller fractions of their total sample pruned, as there is more time for ZTF to sufficiently sample each light curve before they fade in brightness. In contrast, very rapid transients such as M dwarf stellar flares or cooler transients such as SNe Ib/c are more heavily pruned.

2.4. Class Selection for Training

After pruning poor-quality light curves from our data set, we filter the remaining sample to only include SNe spectroscopically classified as either SN Ia, SN Ib/c, SN II, SN IIn, or SLSN-I (following Hosseinzadeh et al. 2020; Villar et al. 2020), including rarer subtypes as detailed below:

1. *Type Ia SNe (SNe Ia).* SNe Ia have distinct Si spectroscopic features (while lacking those of H/He) near peak. They often exhibit secondary peaks in the near-infrared (which can appear in the *r* band; Kasen 2006). Their progenitors are usually white dwarfs that experience thermonuclear runaway as they exceed the Chandrasekhar limit, although diversity in progenitor scenarios exists (Blondin et al. 2012). Due to their high intrinsic rates and bright peak magnitudes ($M_B \simeq -19.3$ mag), which make them observable at higher redshifts for a fixed magnitude limit, Type Ia SNe dominate our data set. In addition to Branch normal SNe Ia (Branch et al. 1993), we have included SNe Ia-91T-like, SNe Ia-CSM, and SNe Ia-91bg-like in this category.
2. *Type Ib/c SNe (SNe Ib/c).* Type Ib/c SNe are core-collapse (CC) SNe without H spectroscopic features. SNe Ic additionally lack He lines. SNe Ib/c tend to be optically redder at peak compared to SNe Ia. The progenitor stars of SNe Ib/c have likely been stripped of their H/He envelopes, potentially by a binary companion (Filippenko 2005) or strong stellar winds (Conti & Niemelä 1976). The optical light curves are predominantly powered by the radioactive decay of ^{56}Ni and ^{56}Co. We include broad-lined SNe Ic (SNe Ic-BL) and calcium-rich SNe Ib (SNe Ib-Ca-rich) in this category. We do not include the one calcium-rich SN Ic event in this data set.
3. *Type II SNe (SNe II).* SNe II are CC SNe with H spectroscopic features. They are primarily powered by H recombination following collapse of a red supergiant (or potentially a blue supergiant), creating a postpeak plateau in their light curves. We include both SNe IIL and SNe IIP subtypes in this category, as the distinction is based on solely photometric features, and there is debate whether these subclasses are truly distinct (Anderson et al. 2014; Sanders et al. 2015; Rubin et al. 2016).
4. *Type IIn SNe (SNe IIn).* SNe IIn are primarily characterized by narrow H emission lines during the photospheric phase (Smith 2014). Their light curves are extremely heterogeneous, in overall brightness, duration, and shape (see Nyholm et al. 2020 for a recent sample). They are primarily powered by shocks arising from the interaction of the SN ejecta and preexisting circumstellar material (CSM), which was likely deposited by luminous blue variable (LBV) progenitors. Because Type II superluminous SNe (SLSNe-II) are not numerous enough to form their own output class, we include them as a subset of SNe IIn. Their hydrogen emission lines can strongly resemble those of SNe IIn, although a fraction of our SLSNe-II more closely resemble SNe II in their emission (Kangas et al. 2022) or are potential TDE/AGN misclassifications (Stein et al. 2023; Ridley et al. 2024). It is uncertain whether IIn-like SLSNe-II are part of the IIn continuum or a distinct class (e.g., see Gal-Yam 2012; and Pessi et al. 2023). We choose not to group SLSNe-II with SLSNe-I because the latter group generally lacks H emission. While both types are exceptionally bright, we avoid using absolute magnitudes to inform our output classes, keeping in line with our goal of redshift-independent classification.
5. *Type I superluminous SNe (SLSNe-I).* SLSNe-I are exceptionally bright SNe ($M_B \lesssim -20$ mag) that lack signatures of H/He/Si in their near-peak spectra. Their light curves tend to be bluer and longer in duration compared to SNe Ia. Their exact progenitor channel is uncertain. Some evidence (Nicholl et al. 2017) suggests that they are powered by the rapid spindown of a newly born magnetar (Metzger et al. 2013); other works propose large amounts of ^{56}Ni decay (Kasen et al. 2011; Gal-Yam 2019). Potential signs of CSM interaction have also been noted (Hosseinzadeh et al. 2022).

We remove 429 objects that do not belong to these five spectroscopic classes, 38 of which arguably belong to the above classes but are marked as peculiar ("pec"). This selection

Table 1
Data Pruning Summary

Object Type	Original Num. %	N -Obs Cut %	Variability Cut %	Num. Remaining %	Percent Removed by Cuts %
			Used in Training		
SLSN-I	101	5	13	83	17.8
SLSN-II	55	3	4	48	12.7
SN II	1326	283	136	907	31.6
SN IIL	2	1	0	1	50.0
SN IIP	126	22	34	70	44.4
SN IIn	265	35	21	209	21.1
SN Ia	6128	1585	198	4345	29.1
SN Ia-91T-like	219	50	9	160	26.9
SN Ia-91bg-like	59	34	2	23	61.0
SN Ia-CSM	20	2	0	18	10.0
SN Ib	157	59	8	90	42.7
SN Ib-Ca-rich	6	5	1	0	100.0
SN Ib/c	40	19	4	17	57.5
SN Ic	170	59	9	102	40.0
SN Ic-BL	82	23	9	50	39.0
			Excluded from Training, Analyzed in Section 5		
AGN	57	6	4	47	17.5
LBV	10	3	0	7	30.0
SN IIb	121	53	8	60	50.4
SN Iax	18	5	1	12	33.3
SN Ibn	33	13	0	20	39.4
TDE	64	4	5	55	14.1
			Excluded from Training		
CV	215	106	3	106	50.7
Galaxy	19	17	0	2	89.5
ILRT	3	2	0	1	66.7
LRN	3	1	0	2	33.3
M dwarf	6	5	0	1	83.3
Nova	36	20	0	16	55.6
QSO	5	0	1	4	20.0
SN	28	14	1	13	53.6
SN I	29	12	2	15	48.3
SN II-pec	10	3	2	5	50.0
SN IIn-pec	2	0	1	1	50.0
SN Ia-Ca-rich	1	1	0	0	100.0
SN Ia-SC	5	1	0	4	20.0
SN Ia-pec	46	12	6	28	39.1
SN Ib-pec	5	1	1	3	40.0
SN Ibn/Icn	2	0	0	2	0.0
SN Ic-Ca-rich	1	0	0	1	0.0
SN Ic-pec	1	0	0	1	0.0
SN Icn	5	3	0	2	60.0
Varstar	19	11	0	8	57.9
Other	26	10	3	13	50.0

Note. Results of data quality cuts on each transient type from our original spectroscopic TNS sample. The first cut ("N-Obs Cut") requires at least five data points of SNR $\geqslant 3$ in each band, while the second cut ("variability cut") ensures the amplitude and flux variations in each band sufficiently exceed the average flux uncertainty.

leaves a training sample of 6123 SNe. From Table 1, 201 pruned events not used in training are AGN, LBVs, SNe IIb, SNe Ibn, SNe Iax, or TDEs. We apply Superphot+ to these events in Section 5.4, as they are not prevalent enough to justify separate output labels, but can still impact the purity of Superphot+'s output predictions. We exclude cataclysmic variables (CVs) from this subsequent analysis, as we assume they will be separated from our SN-like data set by a more general classifier (e.g., the ALeRCE light-curve classifier).

2.5. Photometric Data Set

In addition to a "spectroscopic" training data set, we also collate a "photometric" data set. This consists of ZTF light curves that are high quality and SN-like in behavior, but do not have a spectroscopic classification. Our photometric set will serve as our test set and be classified by Superphot+.

To collate this data set, we use ALeRCE's (Sánchez-Sáez et al. 2021) two "top-level" classifiers. One of these classifiers uses two-band light curves to distinguish between SNe,

stochastic phenomena (e.g., AGN, CVs), and periodic variables (the "light-curve" classifier; Sánchez-Sáez et al. 2021). Sánchez-Sáez et al. (2021) find that ALeRCE's light-curve classifier is highly successful, with an F_1-score of 0.97 and SN completeness of 100% (these metrics are defined in Section 4.2). The other classifier uses image cutouts (the "stamp" classifier; Carrasco-Davis et al. 2021) to potentially label objects as asteroids or bogus in addition to SNe, stochastic variables, or periodic variables. Carrasco-Davis et al. (2021) report 87% SN completeness for the stamp classifier at the time of training, with a ≃5% false-positive rate.

First, we gather 21,781 light curves that are photometrically classified as an SN-like transient by ALeRCE's light-curve classifier with 50% or greater confidence but do not have associated spectroscopic labels (as of 2023 October). We then clip spurious light-curve tails as described in Appendix A, followed by the same observational and variability cuts that we applied to the spectroscopic data set. These cuts prune 5805 and 5854 light curves, respectively, leaving 10,122 high-quality light curves without spectroscopic labels.

Through visual inspection, we find that these cuts do not sufficiently eliminate non-SNe transients from our photometric set. Non-SNe sources include bogus detections, AGN-like variables, and very noisy or low-amplitude variable stars. Therefore, we also remove any light curve not marked as an SN by the ALeRCE stamp classifier. This leaves 3973 events, which means that less than 50% of the events labeled "SN-like" by ALeRCE's light-curve classifier are also labeled as SN-like by the stamp classifier. This result is surprising and suggests either a much higher false-positive rate than reported for ALeRCE's light-curve classifier or a much lower SN completeness than reported for the stamp classifier. Investigation into the performance of top-level classifiers is left to other work.

2.6. Properties of the Reduced Data Sets

After pruning our data sets, we are left with 6123 light curves in the spectroscopic training set, and 3973 light curves in the photometric test set. The breakdown of the spectroscopic set is as follows:

1. SN Ia, 4546 (75.0%);
2. SN II, 978 (16.1%);
3. SN Ib/c, 259 (4.3%);
4. SN IIn, 257 (4.2%);
5. SLSN-I, 83 (1.4%).

We find that 66% of these light curves are also in the ZTF Bright Transient Survey (ZTF BTS; Fremling et al. 2020) catalog, which aims to spectroscopically classify all light curves brighter than 19 mag at peak that pass certain quality cuts.[9] Of our light curves not in ZTF BTS's catalog, 91% are brighter than 19 mag but do not pass ZTF BTS's quality cuts. Our class fractions approximately match those from the entire ZTF BTS data set (Perley et al. 2020), with the exception of SLSNe-I. Our cuts yield an SLSN-I fraction (1.4%) almost double that of ZTF BTS (0.75%). This is not unexpected, as (1) programs targeting SLSNe-I (e.g., FLEET, Gomez et al. 2023a) increase their prevalence in TNS, and (2) SLSNe-I have longer timescales that allow for more observations at a fixed cadence; they therefore pass our first quality cut more frequently than the other SN classes.

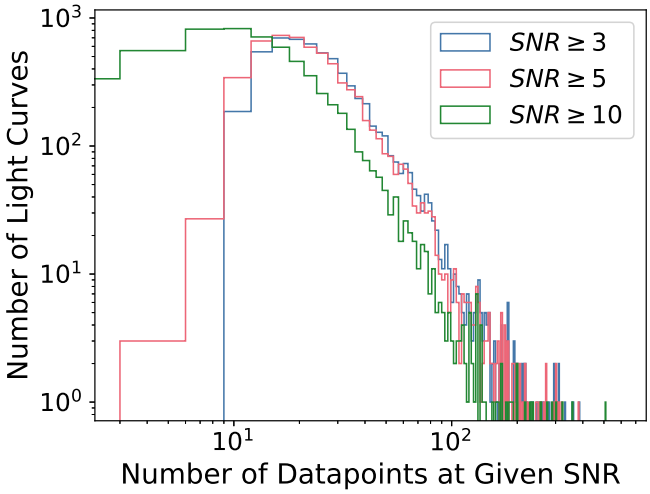


Figure 1. The distribution of spectroscopically classified SNe in our data set as a function of the number of data points above a given SNR. Any light curves with fewer than 10 SNR ⩾ 3 data points are removed during our pruning process. Most light curves that pass our quality cuts (69%) have at least 20 SNR ⩾ 3 data points, and most SNR ⩾ 3 points (93%) also have SNRs ⩾ 5. This highlights the high quality of our remaining light curves.

The distribution of the number of observations (in either band) per light curve above various SNRs is shown in Figure 1. Because we require at least five points of SNR ⩾ 3 per band, there are no events in our pruned data set with fewer than 10 combined observations above this SNR. Most light curves in our pruned data set (69%) have ⩾20 data points with SNR ⩾ 3, and most points with SNR ⩾ 3 (93%) also have SNR ⩾ 5. This demonstrates that the quality of the average light curve in our training set is significantly higher than our minimum quality cuts, and our cuts only remove events that are of significantly lower quality than the rest of the data set.

The final peak r-band apparent magnitude distribution of both the spectroscopic and photometric data set is shown in Figure 2. The distribution of spectroscopically classified SNe matches our expectations, as it exhibits a single peak at approximately 18.5 mag, where ZTF BTS attempts to enforce high spectroscopic completeness. The photometric data set includes a tail of events much brighter than any spectroscopically classified light curve; these events are likely remnant bogus objects or non-SNe. After fitting our pruned data sets, we only classify light curves with sufficiently good quality fits (determined by a reduced chi-squared metric). This cut is further detailed in Section 3.1, and removes most of the very bright outliers within the photometric data set. The remaining photometric events share a similar distribution to the spectroscopic data set, except that the distribution peaks around 19 mag. The number of events in both data sets dimmer than 19.75 mag falls off rapidly, consistent with the ZTF limiting magnitude of ∼20.5 and our signal-to-noise and amplitude constraints.

We note that no constraints are imposed on the temporal coverage of the light curve, so our final spectroscopic sample includes 91 light curves (1.5%) with only prepeak observations, and 503 light curves (8.3%) with only postpeak observations. These partial light curves, while less informative, will be common in real-time classification and are thus crucial

[9] ZTF BTS requires light curves (1) have two constraining measurements within 7.5 days of the brightness peak, and (2) do not set within a month after maximum light.

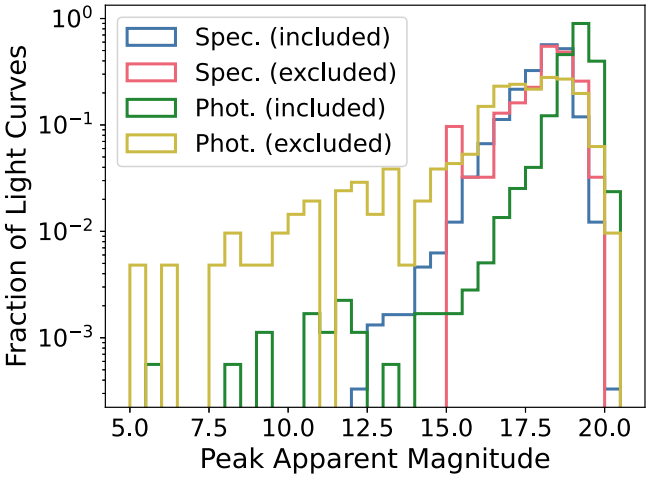


Figure 2. The distribution of peak r-band apparent magnitude (m_r) of the spectroscopic (blue and pink) and photometric (green and yellow) data sets. The spectroscopic and photometric data sets are preprocessed identically. The pink and yellow histograms (62 and 415 objects) correspond to poorly fitted light curves that are cut from the final sample, whereas the blue and green light curves pass the fit quality cuts. Extremely bright sources within the photometric set ($m_r < 10$ mag) and other probable false detections are successfully excluded by the reduced chi-squared cut. The remaining photometric set has a similarly shaped apparent magnitude distribution to the spectroscopic set but shifted toward fainter magnitudes, which reflects ZTF BTS's (and therefore our spectroscopic sample's) brighter limiting magnitude compared to ZTF's (and thus the photometric sample's) limiting magnitude.

to keep in our training set. Similarly, our photometric data set includes 135 (3.4%) prepeak and 267 (6.7%) postpeak light curves. We explore partial light curves more thoroughly in Section 5.2.

3. Parametric Model and Fitting Procedure

After each light curve is preprocessed, we fit the SN flux in each band to a piecewise parametric model introduced by Villar et al. (2019) and Hosseinzadeh et al. (2020):

$$F(t) = \frac{A}{1 + \exp\left[\frac{-(t - t_0)}{\tau_{\rm rise}}\right]} \times \begin{cases} 1 - \beta(t - t_0), & \text{if } t - t_0 < \gamma \\ (1 - \beta\gamma)\exp\left[\frac{\gamma - (t - t_0)}{\tau_{\rm fall}}\right], & \text{otherwise} \end{cases}. \quad (1)$$

This model has seven fit parameters and describes a rise in brightness followed by an approximately linear plateau, which then switches to an exponential decline γ days after t_0. This general form captures the main characteristics of both CC SNe and SNe Ia. The effect of each parameter on the model is illustrated in Figure 2 of Villar et al. (2019). A is the amplitude of the model, t_0 is roughly the phase of peak brightness, and $\tau_{\rm rise}$ and $\tau_{\rm fall}$ are the exponential timescales for the rise and decline of the light curve, respectively. β and γ represent the slope (relative to the amplitude) and duration of the plateau following peak, respectively. There are two versions of each parameter, as shown in Table 2, corresponding to the fits separately derived from the set of g-band observations and the set of r-band observations. Finally, each band has an associated $\sigma_{\rm extra}$ parameter, which serves as an extra uncertainty added in quadrature to each of the flux uncertainties. This extra uncertainty accounts for the limitations in the empirical model itself.

We expect the g- and r-band light curves for a given event to be correlated. Previous works correlate modeled fluxes across filters using 2D Gaussian processes (Boone 2019; Kornilov et al. 2023) or blackbody approximations (Russeil et al. 2024). Within our Bayesian framework, we introduce correlations by expressing the g-band priors relative to the r-band priors for each parameter. We designate the r band as the "reference" band since most light curves have better coverage in the r band. Sampled g-band parameters, denoted by the "g" subscript, are assumed to be the log of the ratio between the actual g- and r-band fit parameters. We choose to sample all of our parameter ratios in log space because an equal shift in either direction of the log parameter corresponds to an equal but inverse multiplicative scaling of the g- over r-band parameter ratio. The only exception is $t_{0,g}$, which is expressed as the difference (time delay) between the g- and r-band t_0 parameters. All sampled g-band ratios are then combined with the sampled r-band parameters before being used in our parametric model. This formulation correlates fit parameters across bands and constrains multiband fits in regions sampled in only one band. It also leads to very narrow and informative priors, as we find that g- to r-band parameter ratios are quite similar across SN-like light curves. Our choice of parameterization could bias fits for parameters where interband correlations are less physically justified. Examples include the amplitude ratio of SNe affected by extreme host reddening, parameters affected by higher-order features our model cannot capture (such as secondary r-band peaks in SNe Ia), or parameters derived from partial light curves (which would just reflect our fitting priors). We note that the latter issue would only arise among the 594 partial light curves (9.8%) in our spectroscopic sample.

The above choice of priors necessitates that we fit g and r bands simultaneously in a 14-dimensional parameter space. This differs from the method used in Villar et al. (2019) and Hosseinzadeh et al. (2020), where each band is fit independently twice, and the second iteration's prior distribution is the average of each band's posterior distribution from the first iteration. While both strategies encourage similar fits across bands of the same light curve, ours allows us to define the expected variation between bands a priori. Furthermore, the final fits are less likely to be skewed by bands with fewer data points. However, fitting 14 parameters simultaneously is more difficult and computationally expensive than fitting seven parameters.

To efficiently explore this combined high-dimensional space, we turn to a variety of modern sampling techniques as discussed in Appendix B. We ultimately use nested sampling to fit the archival light curves used in this work, and stochastic variational inference (SVI) for real-time light-curve fitting. The efficiency of both of these samplers at early iterations relies heavily on the size of the prior volume; therefore, we spend significant effort refining the fit priors to best mirror the expected best-fit parameters of ZTF SNe. To achieve this, we start with broad, uniform priors, and then iteratively alternate between fitting our data set and replacing our priors with the marginal posterior distributions combined from all our light curves.[10] This process continues until the priors and population-level posteriors are sufficiently similar. All final priors are

[10] Note that, before combining posteriors, we oversample our fits to balance class prevalence, as described in Section 4.1.

Table 2
Model Fit Priors

Parameter	Distribution	Mean	Standard Deviation	Truncated Min	Truncated Max
A	Log-Gaussian	0.096	0.058	−0.3	0.5
β	Gaussian	8.3×10^{-3}	3.9×10^{-3}	0	0.03
γ	Log-Gaussian	1.43	0.31	0	3.5
t_0	Gaussian	$t(F_{\max}) - 17.9$	9.9	$t(F_{\max}) - 100$	$t(F_{\max}) + 30$
$\tau_{\rm rise}$	Log-Gaussian	0.67	0.43	−2	4
$\tau_{\rm fall}$	Log-Gaussian	1.53	0.30	0	4
$\sigma_{\rm extra}$	Log-Gaussian	−1.66	0.34	−3	−0.8
A_g	Log-Gaussian	−0.08	0.11	−1	1
β_g	Log-Gaussian	−0.21	0.27	−2	1
γ_g	Log-Gaussian	−0.05	0.17	−1.5	1.5
$t_{0,g} - 1$	Log-Gaussian	−3.4	4.4	−50	30
$\tau_{{\rm rise},g}$	Log-Gaussian	−0.15	0.19	−1.5	1.5
$\tau_{{\rm fall},g}$	Log-Gaussian	−0.15	0.26	−1.5	1.5
$\sigma_{{\rm extra},g}$	Log-Gaussian	−0.15	0.25	−1.5	1.0

Note. The prior distributions for each fit parameter, which are sampled to explore the posterior probability space during nested sampling. For the log-Gaussian distributions, the provided mean, standard deviation, and truncated limits are of the underlying Gaussian distribution before exponentiation.

truncated Gaussians or truncated log-Gaussians, as detailed in Table 2. The priors for A_r and $\sigma_{{\rm extra},r}$ are expressed relative to the maximum r-band flux value of each light curve. We can see that our final priors mirror the data set's combined marginal posterior distributions in Figure 3.

For each light curve, our nested sampler returns a set of several hundred posterior samples (the exact number varying per light curve). The resulting fits for six representative SNe are shown in Figure 4. Note that our fits are tightly constrained for very well-sampled light curves, while the fits from poorly sampled light curves (e.g., those with only rise or decline information) are more likely to be prior dominated. We will use our posteriors to estimate uncertainties in our final classifications.

We note that the same parametric model (that of Villar et al. 2019) is used as part of a larger feature set in Sánchez-Sáez et al. (2021), although Sánchez-Sáez et al. (2021) fit each band independently and do not use the $\sigma_{\rm extra}$ parameters. Sánchez-Sáez et al. (2021) find that the resulting best-fit parameters are more effective than other extracted light-curve features in differentiating between SN classes. However, ALeRCE uses the Levenberg–Marquardt fitting algorithm, implemented in the Python package `scipy` (Virtanen et al. 2020) as `curve_fit`. As we explore in Appendix B, gradient-descent based minimization algorithms like Levenberg–Marquardt can often lead to poor optimal model fits.

3.1. Fit Quality Metrics

To evaluate the quality of our model fits, we calculate a modified reduced chi-squared value for each fit per light curve:

$$\chi^2_{\rm red} = \frac{1}{N}\sum \frac{|\boldsymbol{f}_{\rm obs} - \boldsymbol{f}_{\rm model}|^2}{\boldsymbol{\sigma}^2_{\rm eff}}, \quad (2)$$

where N is the number of data points, and $\boldsymbol{\sigma}^2_{\rm eff} = \boldsymbol{\sigma}^2 + \sigma^2_{\rm extra}$. This differs slightly from the traditional reduced chi-squared definition in that we do not augment the denominator to reflect the number of fit parameters (degrees of freedom). We opt for our modified metric as opposed to the traditional reduced chi-squared value because the latter breaks down for light curves with fewer than eight data points per band; the influence of priors would prevent the best fit from perfectly passing through the data points even for these sparser light curves, yielding infinite reduced chi-squared values.

We calculate the median reduced chi-squared values across all fits per light curve, and we find that 6061 (∼99%) light curves in our spectroscopic data set are fit with a median reduced chi-squared value less than 1.2. We then apply a chi-squared cut to remove the remaining 62 light curves. This set includes 44 SNe Ia, 12 SNe II, 3 SNe IIn, and 3 SNe Ib/c. From these events, we see that 48 cannot be fit adequately by our empirical model as a result of either poor template subtraction, extreme outlier points, or extreme secondary peaks either before or after the primary peak.[11] An additional four light curves could have been fit better to lower the chi-squared value, leaving about 10 well-fit light curves that are unnecessarily removed. However, not applying a chi-squared cut or using a more lenient threshold lets significantly more bogus fits through, degrading the quality of our training set.

We also apply a $\chi^2_{\rm red} \leqslant 1.2$ cut to our photometric data set, which removes 415 light curves. From both Figure 2 and visual inspection, we see that this cut successfully removes both abnormally bright light curves that are likely bogus, and light curves that are clearly not from SNe. We are left with a final photometric data set of 3558 (89.6%) light curves.

We see the effect of our fit quality cut in Figure 5, where we also overlay classification accuracy as a function of fit $\chi^2_{\rm red}$. This accuracy is calculated from the final training results in Section 5. We see in this plot that a cutoff value of 1.2 preserves most of our spectroscopic data set and prevents accuracy from dropping at high $\chi^2_{\rm red}$ values due to poorly estimated fit parameters. Among the photometric data set, the peak of the $\chi^2_{\rm red}$ distribution is also well below the cutoff value.

[11] The samples with poor template subtraction pass the data pruning process due to the estimated flux baseline dramatically changing over time, causing the rise of the light curve to end up dimmer than the decline region in the subtracted light curve. This mode of incorrect subtraction mainly appears among pre-2019 light curves.

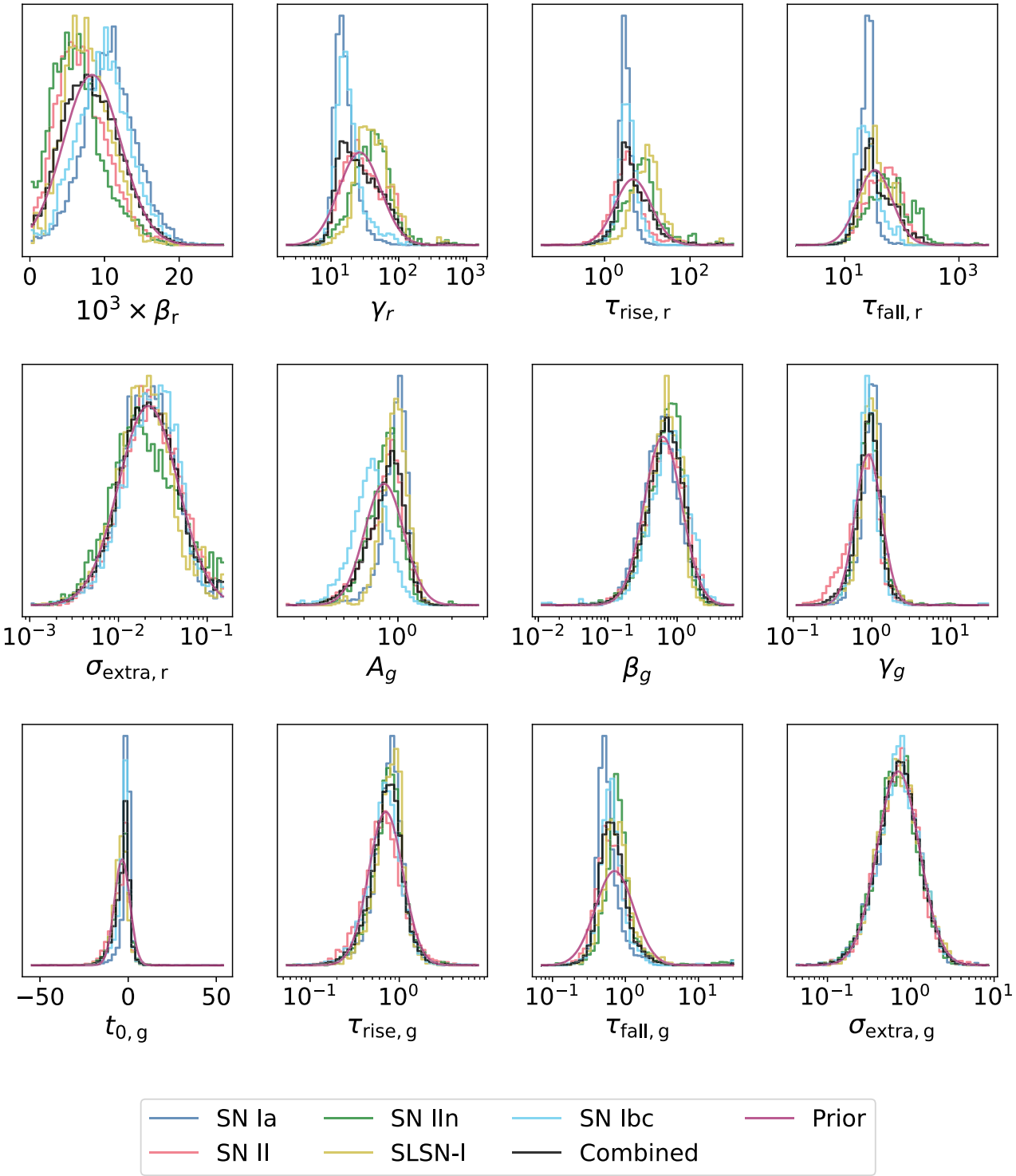


Figure 3. Marginal distributions of each fit parameter in the oversampled (equal contribution from each class) data set. We exclude the r-band A and t_0 parameters as they are not used as inputs for our classifier without redshift information; the g-band versions of these parameters are log ratios between the two bands. Some parameters (e.g., β_g, γ_g, $\tau_{\mathrm{rise,g}}$) are distributed similarly for each SN type, whereas others (e.g., γ_r, τ_{fall}) are clearly more distinguishing. Final priors, shown in purple, are set as the Gaussians or log-Gaussians that most closely match the combined population distributions.

4. Classifier Details

4.1. Balancing the Training Set

As detailed in Section 2.6, our spectroscopic data set is heavily imbalanced across classes, with there being over 50 times as many SN Ia (our majority class) as SLSN-I events (smallest class). Machine learning algorithms perform less efficiently on imbalanced training sets, preferring to over-classify the majority class. Therefore, we need to (1) make the most of every light curve in our minority classes and (2) rebalance the training set as part of our training process.

To accomplish the former, we use stratified K-fold cross-validation to calculate performance metrics from every sample in our spectroscopic set. In this work, we use $K = 10$ folds,

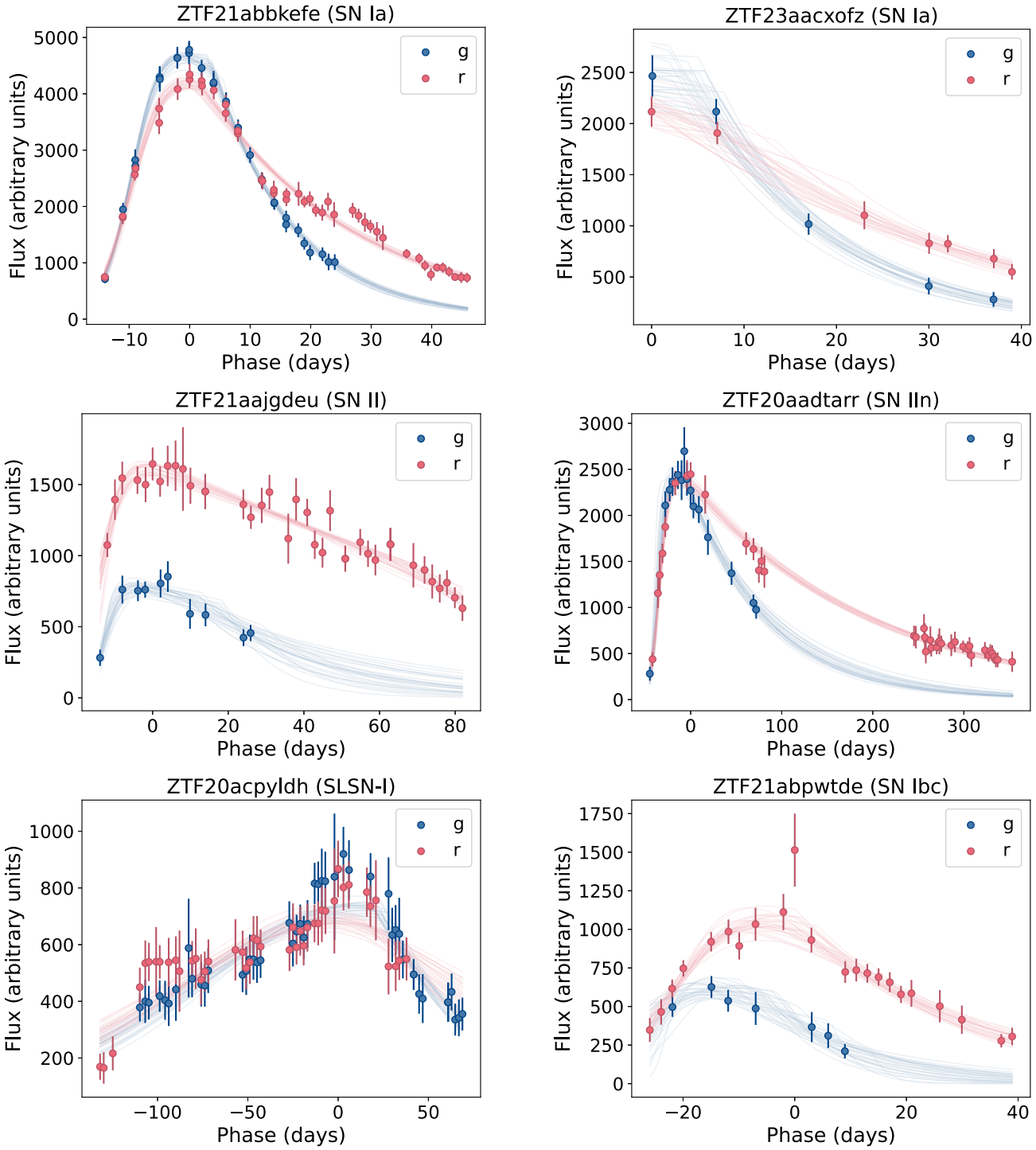


Figure 4. Representative model fits for six classified SN light curves. We show two SNe Ia (top row), an SN II (middle left), an SN IIn (middle right), an SLSN-I (bottom left), and an SN Ib/c (bottom right). Thirty posterior draws are shown for each light curve, capturing the uncertainties and scatter in our model fits.

meaning the data set is split into 10 groups, with equal class fractions in each group, and each group is used as the test set for a separately trained classifier. We find that, while fewer folds degrades classifier performance, more folds increases performance variation across folds. The remaining 90% of each fold not used in the test set is further split 90–10 into a training and validation set. These two sets are oversampled independently for class balancing and then used to train the corresponding classifier.

We rebalance our training and validation sets by oversampling multiple model fits per minority-class light curve; the classifier treats each fit as a separate input. This procedure is discussed in detail, and compared to traditional oversampling techniques, in Appendix C.

4.2. Classification Metrics

Before exploring classifier architectures, we define four metrics to evaluate classification performance. The per-class

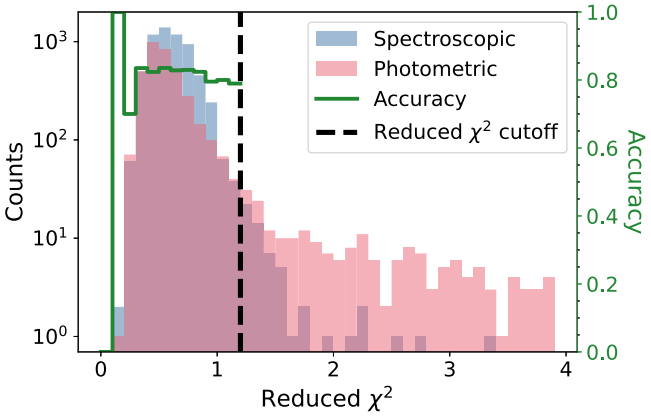


Figure 5. The distributions of reduced chi-squared values from the light-curve model fits for the spectroscopically and photometrically classified samples. Because a very small fraction of the spectroscopic set has a reduced chi-squared above 1.2, we can reasonably assume that most light curves with values above 1.2 are either not SNe or SNe of much lower data quality. We therefore exclude them from our final data sets. The right axis shows classification accuracy as a function of reduced chi-squared for the pruned spectroscopic set, which further supports our choice of cutoff.

completeness is the fraction of samples that belong to one class that are correctly classified as that class. The *accuracy* is the microaveraged completeness, or the fraction of light curves in the entire data set that are classified correctly. The per-class *purity* is the fraction of light curves classified as one class that are actually of that class. The F_1-*score* is the harmonic mean of the completeness and purity. Purity, completeness, and F_1-score are calculated separately for each SN class. Each performance metric is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP}}{N} \tag{3}$$

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$F_1 = \frac{2 \times \text{Purity} \times \text{Completeness}}{\text{Purity} + \text{Completeness}}, \tag{6}$$

where N is the total number of light curves in the data set. *TP* is the true-positive count, which is the number of light curves within one class that are correctly predicted as that class. Likewise, *FP* is the false-positive count (the number of samples classified as one type that are not actually that type), and *FN* is the false-negative count (the number of samples of one class that are misclassified). All metrics are expressed in our results as 80% confidence intervals, or the median value bounded by the second-highest and second-lowest value among our 10 *K*-folds.

Because our data set is highly imbalanced, quantifying classifier performance by accuracy will bias it toward correctly classifying SNe Ia at the expense of the rarer SN types. On the other hand, macroaveraged (i.e., class-averaged) statistics equally penalize low performance within any SN type. Because we value both per-class completeness and purity, as each has its respective science cases, we optimize our parameters by maximizing the macroaveraged F_1-score; this gives equal importance to all SN classes and balances completeness and purity values.

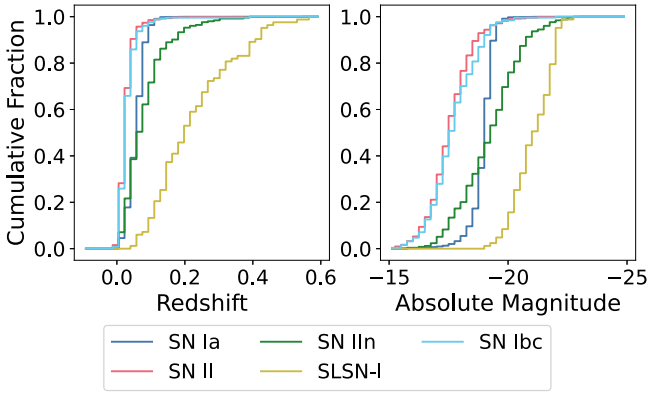


Figure 6. Cumulative distributions of the redshifts and corresponding absolute peak *r*-band magnitudes of our spectroscopic subset with redshift information. As expected, some SLSNe-I have brighter absolute magnitudes and therefore can be detected at farther redshifts. The bulk of our data set is low-redshift light curves ($z < 0.2$) with absolute magnitudes between -16 and -20.

4.3. Classifier Feature Selection and Architecture

Photometric redshift estimates ("photo-zs") are notably more common than spectroscopic estimates in current and future (e.g., LSST) SN data sets. Each photo-z is calculated from the broadband spectral energy distribution of the SN's host galaxy, and associating SNe with the correct host galaxy is not a trivial task. Sources highly offset from their closest galaxy, sources at high redshift, or sources near more than one galaxy can be attributed no or incorrect photometric redshift (see, e.g., Gagliano et al. 2021). Therefore, we first train our classifier on features that do not require redshift estimates. This is straightforward since we do not use redshifts during the preprocessing or fitting steps of Superphot+. We construct our redshift-independent classifier inputs from 12 out of the 14 fit parameters, excluding both A_r and $t_{0,r}$. Neither A_r nor $t_{0,r}$ are intrinsic SN properties; the former is dependent on the phases in which each SN is observed (and associated apparent fluxes), while the latter is dependent on the absolute MJD of each SN.

ZTF has more complete and reliable redshift estimates compared to what we expect from the first few years of Rubin (as detailed in Section 1). Therefore, we present a second version of our classifier that does use redshift information and is only trained on light curves with associated host galaxy redshifts. From our spectroscopic data set, 16 light curves are missing redshift estimates on TNS. We exclude these light curves when training the redshift-inclusive classifier.

We also readd the *r*-band amplitude (A_r) as an input feature for the redshift-inclusive classifier. Additionally, we include the redshift z and the (cosmological *k*-corrected) *r*-band absolute magnitude M_r, calculated from the brightest measured *r*-band flux. The cumulative distributions of these two new inputs are plotted in Figure 6. We see that most events in our training set are at $z < 0.2$, with the exception of some SLSNe-I that have brighter absolute magnitudes. Because A_r is the amplitude of the modeled peak relative to the brightest measured flux (which can be significantly offset for partial light curves), combining it with M_r quantifies the absolute magnitude of the *modeled* light curve at peak.

The shape parameters β, γ, τ_{rise}, and τ_{fall} all have a first-order $(1+z)$ redshift dependence resulting from time dilation. When using redshifts as input features, we assume our classifier can learn to correct for this dilation. However, for our redshift-independent classifier, not correcting for this effect could skew

fit parameters for farther SNe (such as many SLSNe-I) and in turn affect classifier performance. For our ZTF SN data set, the majority of redshifts are less than 0.2, and less than 0.6 among SLSNe-I, which would not shift fit parameters enough to mimic other SN classes (as supported by Figure 3 and our use of log-normal priors). Therefore, we do not consider this a significant source of classification error within our redshift-independent classifier. However, for surveys that can observe SNe at farther redshifts (e.g., Rubin), time dilation can stretch light curves by much larger factors, potentially impacting classification. For these surveys, one can alternatively replace the four aforementioned input features with $\theta_1 = \beta\tau_{\rm rise}$, $\theta_2 = \tau_{\rm rise}/\tau_{\rm fall}$, and $\theta_3 = \tau_{\rm rise}/\gamma$, which would reduce our input vector's length by 1 and cancel out the $(1+z)$ multiplicative factors. We find that this alternate feature set reduces the F_1-score by $\sim$10% for our ZTF data set, but may improve performance for deeper surveys.

The output of each classifier is a vector of five values representing the "pseudoprobability" of the input event belonging to each SN class. These are not true probabilities because, while the vector elements sum to unity, they are not calibrated (i.e., confidence values do not match the fraction of true samples within events assigned that confidence). For our multiclass problem, the assigned label for each light curve is the class with the highest output pseudoprobability. The classification "confidence" is this highest probability. For single-class variants, we instead assign positive labels to an event if the positive pseudoprobability exceeds a prespecified confidence threshold. We provide a more detailed discussion on the calibration of these pseudoprobabilities and confidence thresholds in Section 5.3.

We explore two different classifier architectures, neural networks and GBMs, in Appendix D. We find that GBMs, constructed with the `LightGBM` package, yield optimal classifier performance across K-folds. We train separate GBMs to classify events with and without redshift information.

5. Classification Results

In this section, we test multiple variants of our classifier and quantify the efficacy of our pipeline. These variations include multiclass classification, single-class classification, and training with and without redshift information.

5.1. Multiclass Classification without Redshift Information

First, we study the results of our five-way GBM. This classifier does not use any redshift information and is the "default" classification mode in Superphot+. Our trained model classifies the spectroscopic data set with a F_1-score of 0.61 ± 0.02 and an accuracy of 0.83 ± 0.01. The associated confusion matrices are shown in Figure 7. The class-averaged completeness is 0.73 ± 0.08. SNe Ia unsurprisingly have the highest completeness at 0.87 ± 0.01 due to their high prevalence in the spectroscopic data set. On the other hand, SNe IIn are prone to the highest fraction of misclassifications, with a completeness of 0.52 ± 0.07. This may be because SNe IIn are a highly diverse class of SNe in terms of observational properties (see, e.g., Nyholm et al. 2020, for a recent sample from the Palomar Transient Factory).

The class-averaged purity is 0.58 ± 0.04, with SNe II and SNe Ia having the highest purities at 0.84 ± 0.03 and 0.97 ± 0.01, respectively. In contrast, SLSNe-I and SNe Ib/c have notably lower purities (both $\simeq$0.25–0.30). This is likely reflective of the imbalances in our training set; even a small fraction of true SNe Ia or SNe II can heavily contaminate the small samples of predicted SLSNe-I or SNe Ib/c. For example, only 9% of SNe Ia are misclassified as SNe Ib/c, but this small fraction still translates to 420 SNe Ia. These contaminants account for over half of the 713 total predicted SNe Ib/c. We note that this difficulty is also observed in Hosseinzadeh et al. (2020), again due to a significant class imbalance (albeit less extreme than our data set's imbalance).

Feature importance analysis shows that Superphot+ relies most on fall timescales, followed by plateau durations, peak band ratios (a proxy for color), and rise timescales. It is then no surprise that Superphot+ most commonly misclassifies light curves of similar timescales. SNe Ia are most likely to be misclassified as SNe Ib/c (and vice versa), and SLSNe-I are most likely to be misclassified as SNe IIn (and vice versa). SNe II are misclassified about equally across other classes. We also find that very long-lived light curves, such as those of SNe IIn and SLSNe-I, are sometimes wrongly attributed long plateaus (causing SN II misclassifications), or SNe II without post-plateau sampling are fitted with no plateaus and slow fall timescales (causing SLSN-I misclassifications). We can clearly see overlaps in fitting parameters between classes in Figure 3, each potentially impacting classifier performance.

Examples of misclassified light curves, along with their classification probabilities, are shown in Figure 8. By manual inspection, the main confounding factors among misclassified light curves of each type are as follows:

1. *SNe Ia.* Most SNe Ia misclassified as SNe II, SNe IIn, or SLSNe-I have partial light curves and thus some unconstrained fit parameters. Many SNe Ia that are wrongly labeled as SNe Ib/c have secondary r-band peaks that cannot be fit by our model, or $g-r$ colors similar to those of SNe Ib/c (i.e., redder than typical SNe Ia). We note that, since we only correct for Milky Way extinction, SNe Ia belonging to host galaxies with exceptionally high extinction will appear redder and more like SNe Ib/c.
2. *SNe II.* SNe II with missing observations that prevent plateau constraints are more likely to be misclassified. This is especially problematic for long-lived events, which are often misclassified as SNe IIn.
3. *SNe IIn.* The SN IIn class, in general, is particularly heterogeneous. Misclassified events are most often classified as SLSNe-I (similar timescales) or SNe II (fit with a plateau).
4. *SLSNe-I.* Misclassified SLSNe-I often have slow declines, which are best fit as plateaus in our empirical model, meaning that most misclassified SLSNe-I are assigned high SN IIn and SN II probabilities.
5. *SNe Ib/c.* Misclassified SN Ib/c light curves are often sparse or noisy. Some events are particularly blue around peak and thus misclassified as SNe Ia.

We next investigate light-curve properties associated with especially poor classifier performance. First, we explore the impact of number of observations and light-curve SNR on classification accuracy. We calculate the latter using the top 90th percentile of all the data points' SNRs within the light curve. Both metrics show weak positive correlations with classification accuracy, as evident in Figure 9. More data points correlate more strongly with SN II and SN IIn accuracy. This

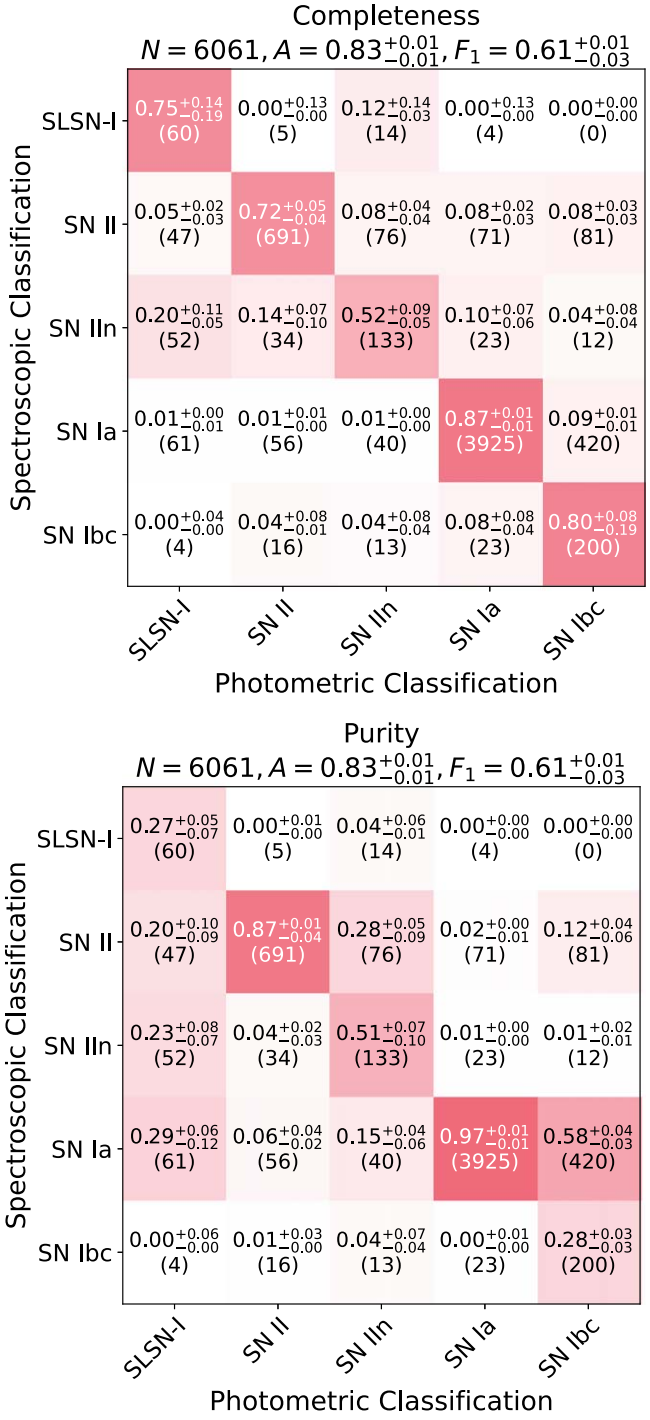


Figure 7. The confusion matrices of the trained classifier on the training set, using K-fold cross-validation. Here, we set $K = 10$ folds. The completeness matrix is normalized to sum to one across each true class, while the purity matrix is normalized along each predicted class. Our classifier achieves high completeness across the five classes. Note that the most populous classes leak into the rarer predicted sample sets, reducing the associated purities.

makes sense, as light curves with many high-SNR observations are more likely to have well-sampled plateaus. Interestingly, SLSN-I and SN Ib/c classification worsens beyond ∼30 data points, as does SLSN-I and SN IIn classification at a 90th percentile SNR above ∼20. Most of the incorrectly classified light curves with over 30 observations are exceptionally long lived and/or have secondary behavior beyond what our empirical model can capture, such as additional peaks, high variance within the fall region, or preexplosion variations that are unable to be removed through template subtraction. The best way the model can capture this anomalous behavior is with extended rise times (resembling SLSNe-I and SNe IIn) or longer plateaus (resembling SNe II and SNe IIn). A small fraction (<10 objects) of SNe Ib/c with >50 observations show dramatically inconsistent template subtraction across the light curve, preventing clean light-curve tail clipping, but this is not the main culprit of SNe Ib/c misclassification. Among SLSNe-I of SNR above 20, we find either (1) shorter, Ia-like evolution timescales, or (2) exceptionally long-lived light curves with II-like plateaus. Most misclassified, SNR > 20 SNe IIn have incomplete light curves and are classified with lower confidence, with the remainder exhibiting either shorter (Ia-like) or longer (SLSN-like) timescales than expected.

We additionally consider our classifier's performance within its most confident predictions. We regenerate our confusion matrices in Figure 10 including only the 3200 (52.7%) events classified with confidence greater than 0.7 (a cutoff chosen to match that from Hosseinzadeh et al. 2020; and Villar et al. 2020). Our performance metrics improve substantially, with a new F_1-score, class-averaged purity, and class-averaged completeness of 0.86 ± 0.05, 0.81 ± 0.12, and 0.92 ± 0.10, respectively. Completeness increases most substantially for SLSNe-I (from 0.75 ± 0.17 to 1.00 ± 0.20) and SNe IIn (from 0.52 ± 0.07 to 0.80 ± 0.13). These classes also significantly improve in purity, with SLSN-I's increasing from 0.27 ± 0.06 to 0.83 ± 0.28 and SN IIn's increasing from 0.51 ± 0.09 to 0.76 ± 0.19. However, 75% of both classes are removed by the high-confidence cut (compared to 43% of SNe Ia), leaving only a handful of very confidently labeled events. We note that SNe Ib/c are, again, labeled with the lowest purity at 0.50 ± 0.13 (with many labeled SNe Ib/c being true SNe Ia).

5.2. Classification of Partial Light Curves

Next, we explore the efficacy of Superphot+ as a real-time classifier by analyzing its performance on partial light curves. This is important to explore, as one major benefit of Superphot + is computationally efficient fitting that can keep up with both ZTF and expected LSST alert streams, enabling real-time classification.

First, we consider that not all fit parameters will be informative for early-phase light curves. If the light curve's final observation is before the end of a plateau or peak, then the $t_0 > \gamma$ function of our piecewise model will be completely unconstrained, and the best-fit values for γ and $\tau_{\rm fall}$ (for both bands) will reflect their priors. These features may skew our classifier toward incorrect labels for these partial light curves. Therefore, we train an alternate version of our GBM classifier without γ, $\tau_{\rm fall}$, or the corresponding g-/r-band ratios, calling this our "early-phase" classifier. We compare this classifier's performance to our "full-phase" classifier described in the previous section.

To compare the real-time performance of our early-phase and full-phase classifiers (both without redshift information), we randomly select up to 20 SNe per K-fold per class, and then truncate these selected light curves at a series of increasing phases, where phase is defined as the time after peak r-band brightness. Each truncated light curve, if it has data points, is fit and classified, and the completeness, purity, and F_1-score are calculated at each phase for each SN type for each fold. These macroaveraged metrics for both the early-phase (dashed) and

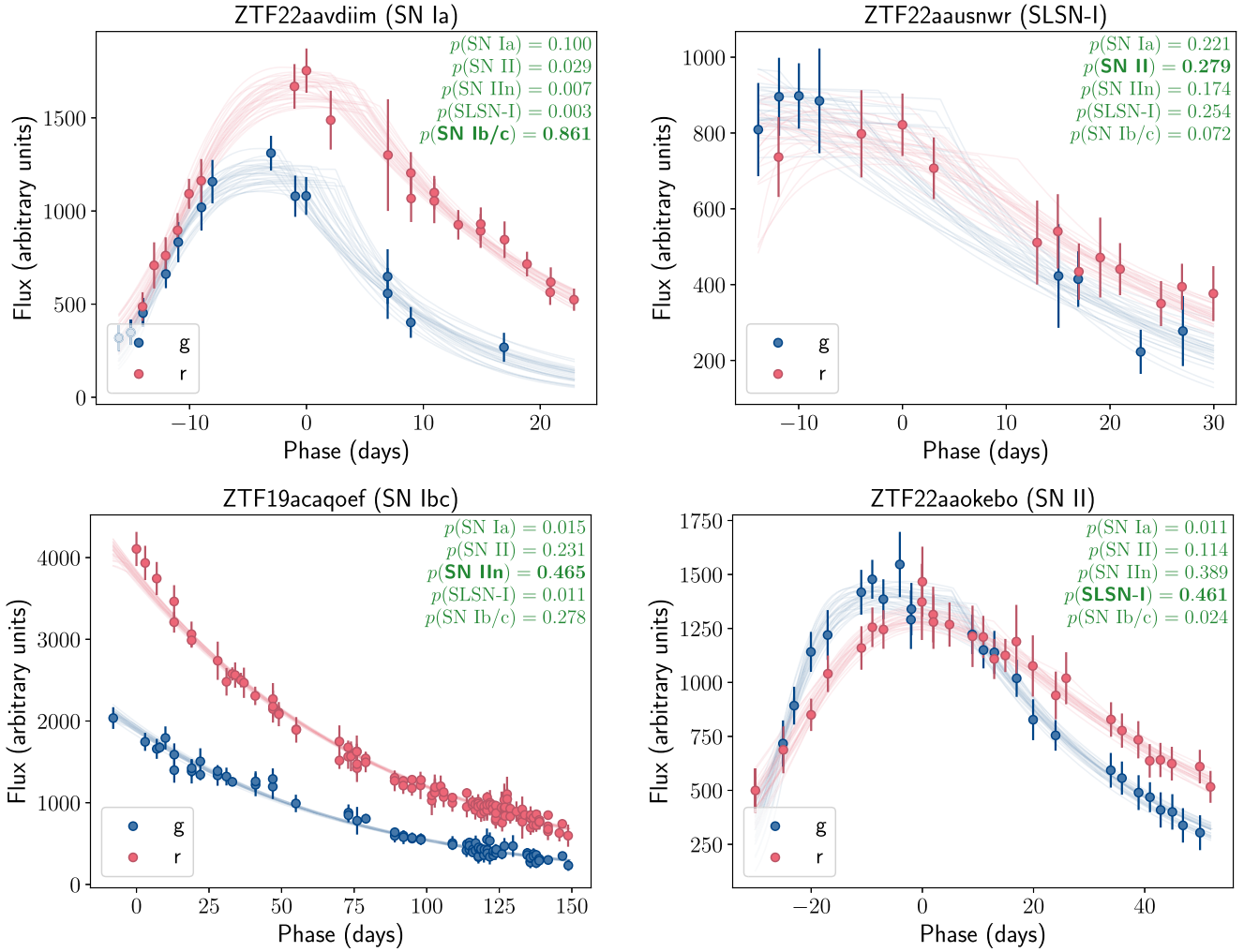


Figure 8. Examples of light curves that are misclassified by Superphot+, where the spectroscopic labels are specified in the titles, and the probabilities output by Superphot+ are in green. Probable causes of misclassification include redder color (top left), unsampled rises (top right and bottom left), and unusual timescales (long-duration tail for bottom left, short plateau for bottom right).

full-phase (solid) classifiers are shown in Figure 11, with uncertainty margins representing the full-phase 1σ uncertainties across 10 K-folds. We find that the early-phase classifier outperforms the full-phase classifier for cutoff phases before $\sim$20 days, after which the full-phase classifier performs better. Note that metrics for very early phases have high variance from small number statistics; few events in the data set are observed at such early phases, most being SLSNe-I. Trends for individual SN classes are detailed in Appendix E.

Both the early-phase and full-phase classifiers are currently integrated into an ANTARES (Saha et al. 2014) filter, labeling events from the ZTF Alert Stream in real time.[12] The early-phase classifier is applied to light curves with last observation earlier than 15 days after peak, while the full-phase classifier is applied to light curves with later observed phases.

5.3. Single-class Performance and Calibration

High-purity samples of a singular SN class are often required for population-specific studies (e.g., increasing the sample of spectroscopically classified SLSNe), at the cost of completeness. Here, we consider the performance of Superphot+ when optimized for binary (single-class) classification problems. We can reuse our trained multiclass GBM by selecting a target class and compressing all probabilities outside of the target class into a single "negative" probability. The problem simplifies from assigning an object one of five class labels to assigning one of two: positive and negative. In the multiclass problem, the assigned label is determined by the highest probability, which differs for every event. Setting a minimum confidence threshold would result in some objects receiving no assigned label, which we do not allow in our multiclass framework. For binary classification, we can adjust the confidence threshold required for a positive label; this choice inversely impacts purity and completeness of the target class. For example, requiring a very high classification confidence before assigning a positive label will lead to a smaller predicted data set of that class but also fewer contaminants from other spectroscopic classes.

Before considering different confidence thresholds, we first consider whether the pseudoprobabilities from our classifier are well calibrated. The calibration curve, shown in Figure 12, examines whether the pseudoprobabilities assigned by the classifier for a specific class is an overestimate or underestimate of the true probability. Pseudoprobabilities are "well calibrated" if the reported classifier probability matches the

[12] https://antares.noirlab.edu/

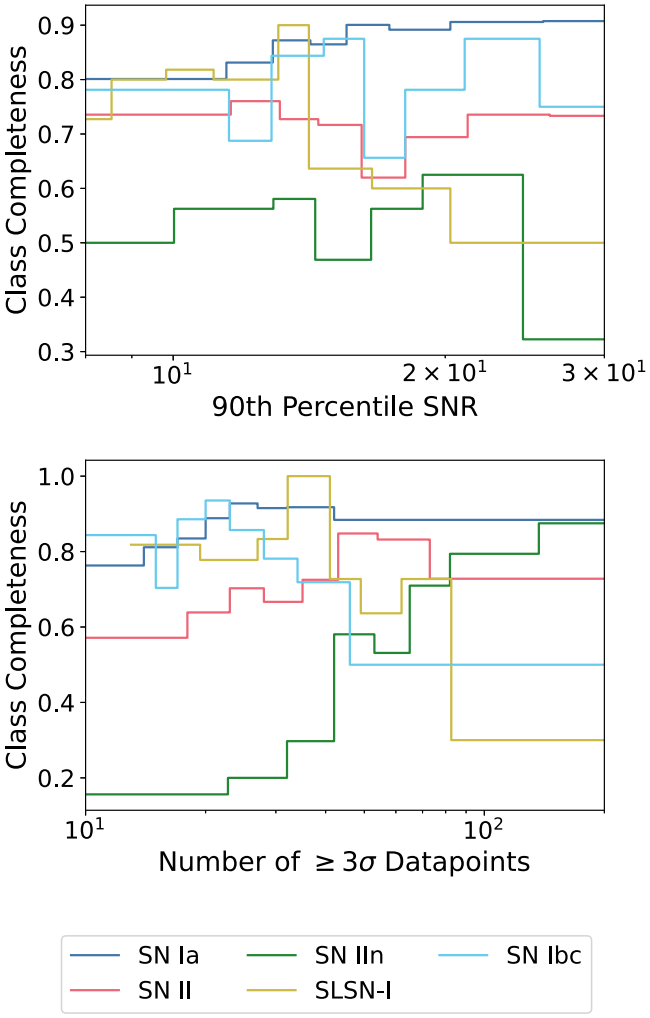


Figure 9. The spectroscopic set's binned classification accuracies as a function of 90th percentile signal-to-noise ratio (top) and number of data points with SNR $\geqslant 3$ (bottom). The bin widths are chosen to contain an equal number of events per bin within each class. There is a weak increase in accuracy as a function of 90th-percentile SNR that plateaus around 20. The exception is SLSN-I, whose highest SNR light curves are longer lived and therefore misclassified as SNe II or SNe IIn. For SNe II and SNe IIn, more data points correlate with higher classification accuracies, likely due to better sampled plateaus, which make them easier to classify. SLSNe-I and SNe Ib/c both show decreased classifier performance with excessive number of data points.

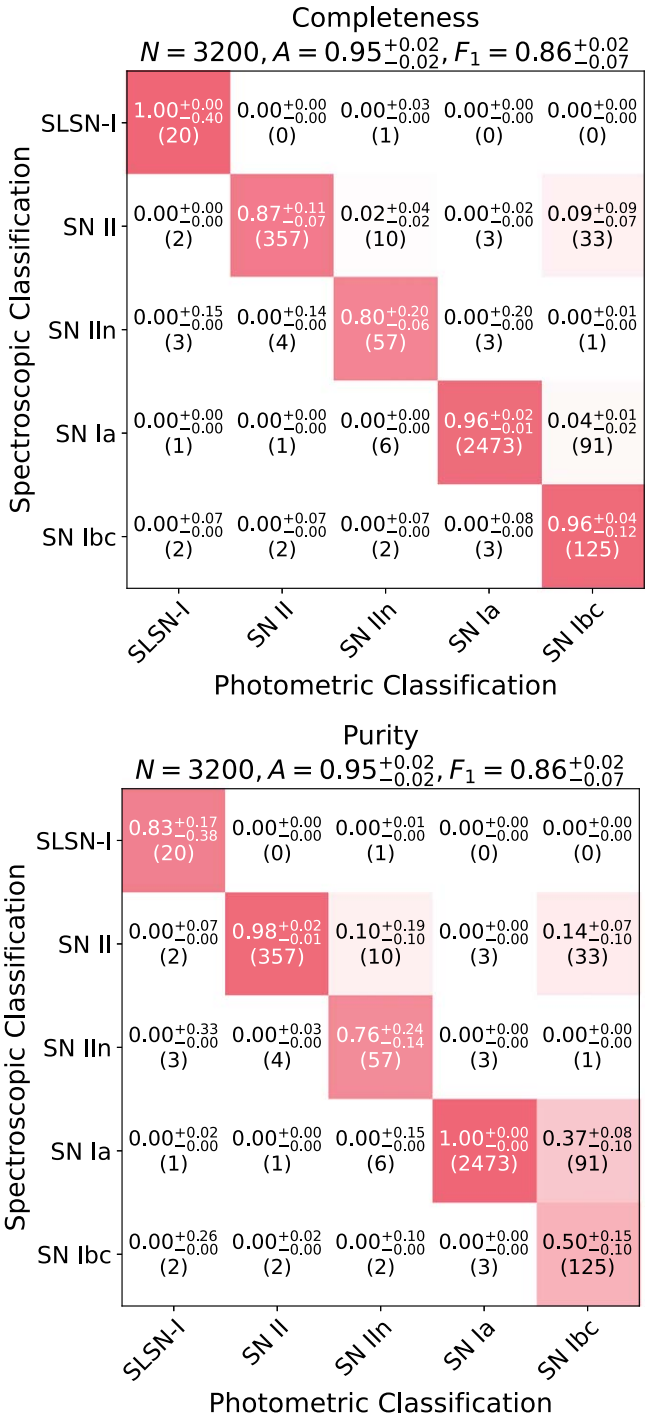


Figure 10. Same as Figure 7, but only including light curves classified with confidence greater than or equal to 0.7. This includes 3200 of the 6061 total events, and both the completeness and purity confusion matrices show significant improvement. The F_1-score increases from 0.61 ± 0.02 to 0.86 ± 0.05, and accuracy increases from 0.83 ± 0.01 to 0.95 ± 0.02. These plots especially highlight the difficulty with discerning SNe IIn from SLSNe-I and SNe II without redshift information (much fewer events included), and in maintaining a high-purity set of Type Ib/c SNe with our imbalanced data set.

fraction of events correctly classified at that confidence. An ideal calibration curve perfectly follows a $y = x$ line for all classes. For Superphot+, the classifier assigns overconfident SN Ib/c pseudoprobabilities, but underconfident SN Ia values. This likely reflects the classifier's balance between SN Ib/c purity and SN Ia completeness. The other three class probabilities do not show strong biases.

Keeping in mind that our classifier is uncalibrated, we can now explore the effect of single-class confidence thresholds on the performance metrics detailed in Section 4.2. While receiver operator characteristic curves (Bradley 1997) are commonly generated to summarize a classifier's performance, they tend to be overly optimistic for highly imbalanced data sets, such as our SN data set. Therefore, we instead rely on purity–completeness curves (i.e., precision–recall curves in machine learning literature) to explore binary classifier performance (Davis & Goadrich 2006; Saito & Rehmsmeier 2015). The purity–completeness curve for each class in our data set is shown in Figure 13, with 1σ uncertainties calculated across K-folds. A perfect classifier for an SN type follows the top right corner, where both the purity and completeness are 1.0. A completely random classifier follows a horizontal line aligned with the target class's prevalence in the data set. The diamonds

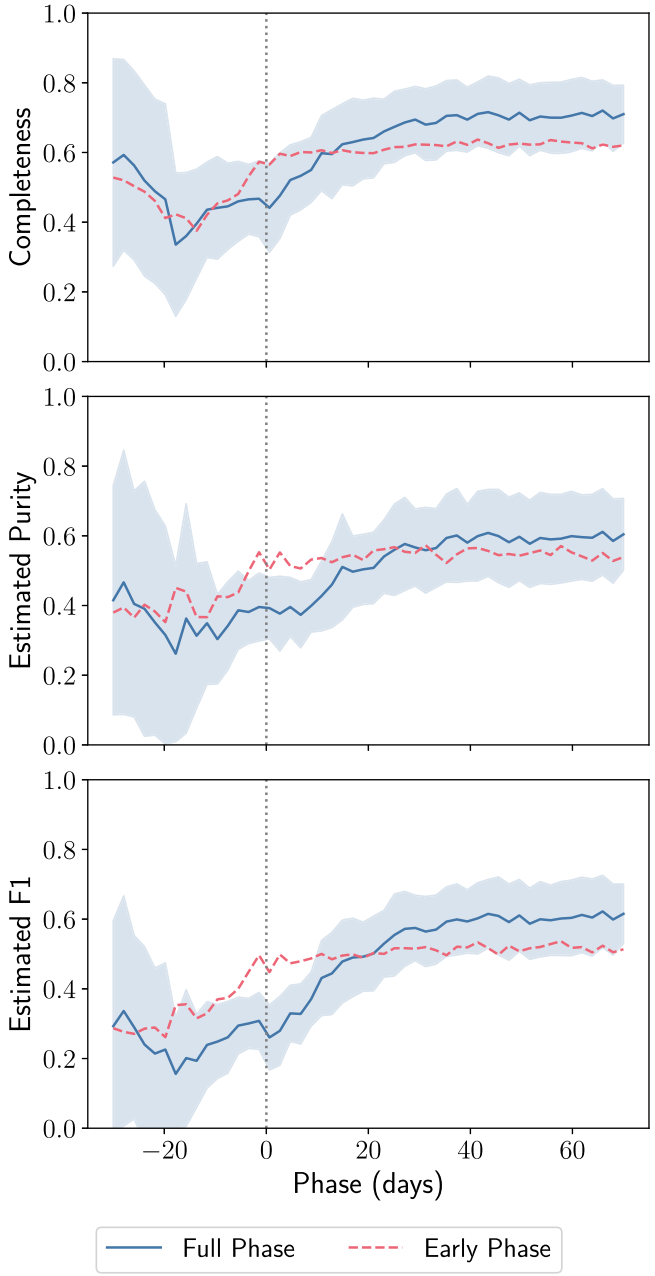


Figure 11. Classification completeness, purity, and F_1-score as a function of light-curve phase, macroaveraged across SN types. The dashed line corresponds to the early-phase classifier, whereas the full-phase classifier is represented by the solid line and associated 1σ K-fold uncertainty margins. Here, phase is relative to peak magnitude in r band. The purities and F_1-scores are estimated from class prevalences and completeness values. Twenty light curves of each type per fold are randomly sampled to generate this data, so final accuracies differ slightly from those of the entire spectroscopic set. The early-phase classifier outperforms the full-phase classifier before phase $\sim$20 days. Similar plots are shown for each SN class separately in Appendix E.

correspond to a confidence cutoff of 0.5, where the assigned label corresponds to the highest (binary) pseudoprobability.

The area under the purity–completeness curve (AUPC) quantifies binary classifier performance. It benefits from not relying on choice of confidence threshold, unlike the binary F_1-score. We use Figure 13 to directly compare Superphot+'s classification of SLSNe-I with that of FLEET (Gomez et al. 2020a, 2023a), a binary classifier designed to isolate a high-purity SLSN-I (or TDE) data set. FLEET's AUPC value is

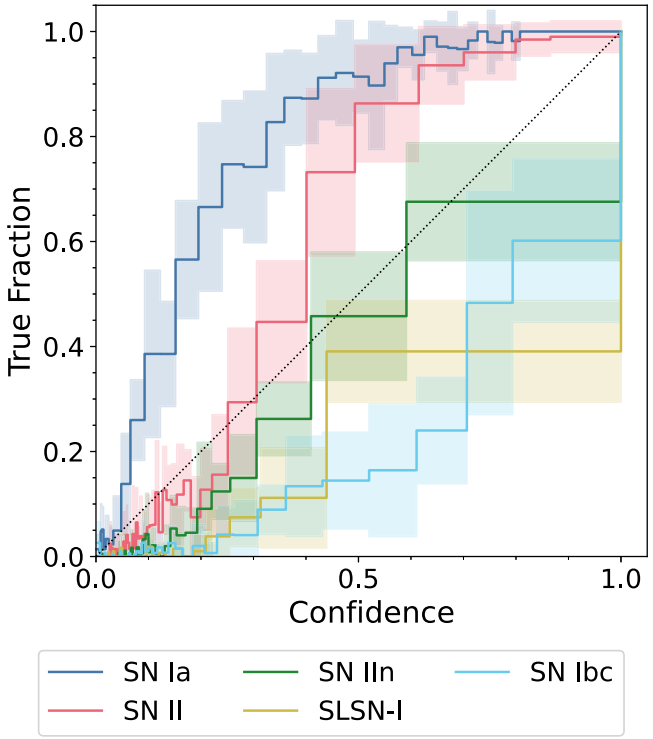


Figure 12. The calibration curve for each SN class, treating Superphot+ as a binary classifier. A calibration curve plots the fraction of events correctly classified at different confidence levels. A well-calibrated classifier would follow the $y = x$ diagonal for each class. We see that Superphot+ is overconfident about SN Ib/c probabilities and underconfident about SN Ia probabilities, while fairly well calibrated for all other classes.

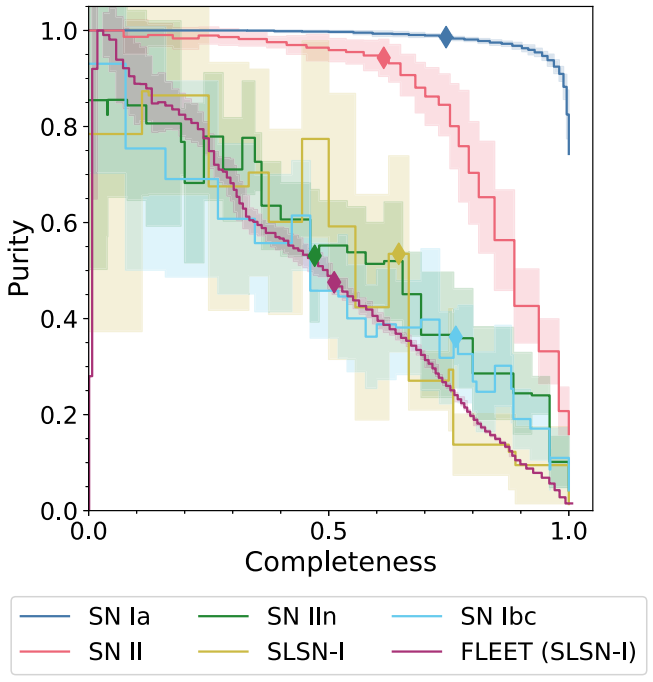


Figure 13. The purity–completeness curve for each SN class, treating Superphot+ as a binary classifier. We calculate single-class metrics by compressing all probabilities from classes other than the target class into a single probability. Each point along the curve corresponds to a different confidence threshold for classification. A "perfect" classifier (i.e., one with 100% confidence and accuracy) would follow the top right corner of the axes, and a completely random classifier follows a horizontal line scaled to the target class's prevalence.

0.49 ± 0.03, with the uncertainty resulting from different random seed initializations. Superphot+'s SLSN-I AUPC value is 0.52 ± 0.19, where the larger uncertainty propagates from variance across K-folds. These overlapping AUPC values are promising, as it shows that Superphot+, which must balance the performance of multiple classes, boasts comparable binary performance to pipelines optimized for binary classification.

We can then choose a confidence threshold to optimize F_1-score for each target class. As an example, we retrain our GBM to output either SN Ia or SN-Other labels, the latter of which includes our other four classes. Generating high-purity (Branch normal) SN Ia data sets is crucial for cosmological studies (Jones et al. 2017), although we note our Type Ia sample is only 96.7% Branch normal; 91bg-like SNe Ia (23), 91T-like SNe Ia (159), and SNe Ia with CSM interaction (18) are also grouped into this classification. We see in Figure 14 that our GBM's optimal confidence threshold is at $p = 0.175$, with a maximal F_1-score of 0.90 ± 0.01 and accuracy of 0.93 ± 0.01. This matches our knowledge that our models return underconfident SN Ia pseudoprobabilities. We also show the corresponding purity matrix; SN Ia purity is high at 0.94 ± 0.01, at the cost of a lower SN-Other purity. The contamination from non-Ia SNe is approximately double that in the curated SN Ia data set from Jones et al. (2017), but a confidence cut can be applied to increase our classifier's SN Ia purity. Additionally, only $4\% \pm 1\%$ of SNe Ia are classified as non-Ia SNe.

5.4. Classification of Excluded Transient Types

Superphot+'s output classes exclude rarer transient types that have SN-like light curves but lack the prevalence to constitute additional output classes. These classes include AGN, SNe Iax, Ibn, IIb, TDEs, and LBVs/other massive star outbursts. We expect many light curves from these classes to pass our data quality and fitting chi-squared cuts and thus contaminate our predicted data sets. We first fit light curves of these classes from our pruned spectroscopic set, and we remove the 15 events with median reduced chi-squareds above 1.2. We then classify the remaining 189 objects with our five-output GBM model to determine which labels they would likely be assigned, and summarize the results in Table 3.

Most SNe Iax are not labeled as SNe Ia, but rather as SNe Ib/c. This is perhaps not surprising as Type Iax SNe tend to be redder than SNe Ia (Foley et al. 2013). SNe IIb are also primarily labeled as SNe Ib/c because they tend to be redder and faster evolving than SNe II without clear plateaus (Claeys et al. 2011). In contrast, most SNe Ibn are classified as SNe Ia or SNe II, since their light curves evolve over faster timescales but are too blue to be mistaken for SNe Ib/c. TDEs are most commonly classified as SLSNe-I or SNe IIn since their light curves are bluer and decay over very long timescales (although many partial light curves are labeled as SNe Ia or SNe II). LBV eruptions are predominantly labeled as SNe II; while their timescales are shorter than those of SLSNe or SNe IIn, many light curves exhibit postpeak variability that is incorrectly fit as plateaus. AGN are mainly labeled as SNe IIn as both can occupy sparser regions of the joint parameter space; furthermore, AGN stochastic behavior is often modeled as extremely gradual declines aligned with some SNe IIn.

We can examine Superphot+'s photometric predictions for its five output classes and determine the level of contamination from the classes listed here. From Table 3, we see that the most common label is SNe Ib/c, with 45 (out of 189) rare-type transients predicted to be SNe Ib/c. While this is significant compared to the 259 true SNe Ib/c in our spectroscopic data set, it is only 5.9% of the 713 events labeled as SNe Ib/c by Superphot+, the majority being SN Ia misclassifications. Therefore, we find that rare-type transients are a minor contaminant for predicted SNe Ib/c. Across all classes, misclassifications from within our five main spectroscopic labels have the greatest impact on class purity. This supports our decision not to include the rarer classes listed here as additional output labels. However, we expect orders of magnitude more events from these classes to be detected by Rubin; a Rubin-tailored Superphot+ variant may include these classes as additional outputs.

5.5. Inclusion of Redshift Information

Here, we train a variant GBM that does use redshift information, as described in Section 4.3; this GBM includes redshifts and brightest absolute magnitudes as additional features. The resulting confusion matrices are shown in Figure 15. By including redshift information, the accuracy increases from 0.83 ± 0.01 to 0.88 ± 0.01, and the F_1-score increases from 0.61 ± 0.02 to 0.71 ± 0.02. The most significant improvement is the SLSN-I purity from 0.27 ± 0.06 to 0.58 ± 0.14, where the contamination from true SNe Ia and SNe II is dramatically reduced. Without redshift information, partial SLSN-I light curves can be mistaken for SNe Ia or SNe II depending on how much of the rise and fall regions are missing; the disparity in peak absolute magnitudes fixes most of these misclassifications. We additionally see improved SN Ib/c purity from 0.28 ± 0.03 to 0.38 ± 0.06, as brighter SNe Ia are less frequently mislabeled as SNe Ib/c.

We also consider the subsample classified with confidence above 0.7 in Figure 16. Of the 6045 events with associated redshifts, 4485 (74%) are classified with confidence above 0.7, with a F_1-score of 0.81 ± 0.03. This F_1-score is actually lower than that of the GBM trained without redshift information when the same confidence cut is applied. However, the GBM with redshift information is generally more confident; a much lower fraction of light curves (25% versus 47% for the redshift-independent classifier) is removed by the $p > 0.7$ cut. We conclude that including redshift information makes our classifier more confident when labeling events, but *less* accurate among highly confident predictions. To determine why, we generate performance metrics among the most confident 3200 events (chosen to match the number of events remaining after the redshift-independent high-confidence cut). In this case, the performance metrics are nearly identical, but we see higher SLSN-I and SN IIn completeness when not using redshift information. This, combined with the number of high-confident SLSNe-I approximately doubling after including redshift information, leads us to conclude that redshift information is biasing our classifier to label high-redshift events as SLSNe-I by default.

Finally, we train an SN Ia versus "other" SN model that uses redshift information, and once again optimize the confidence threshold. With an optimal threshold of $p = 0.160$, the model returns $F_1 = 0.92 \pm 0.01$ and an SN Ia purity of 0.95 ± 0.01, both of which slightly exceed the corresponding binary metrics when not using redshift information.

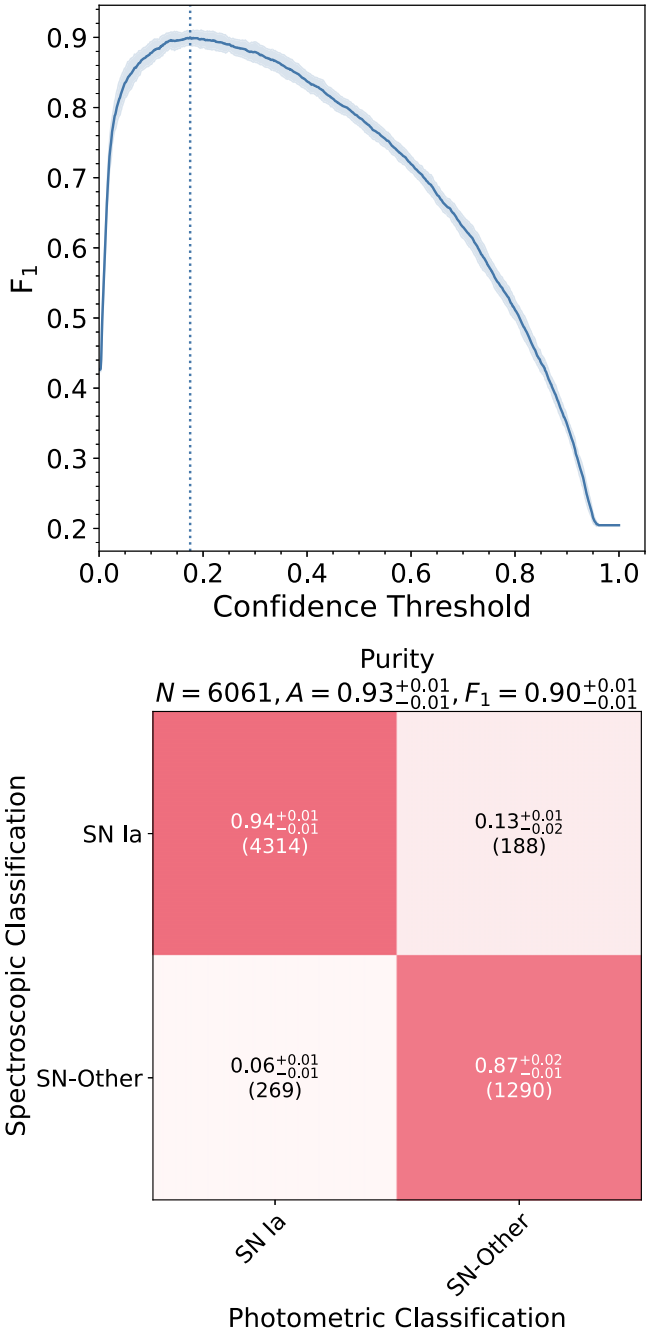


Figure 14. Top: The binary F_1-score of our SN Ia vs. core-collapse SN classifier as a function of confidence threshold. Higher thresholds mean the classifier must return a higher pseudoprobability for SN Ia for us to assign the object an SN Ia photometric label. We find that F_1 is optimized for an SN Ia threshold of $p \geqslant 0.175$, indicated by the vertical dotted line. Bottom: The corresponding purity matrix when we use the optimal confidence threshold. We see very successful differentiation between SNe Ia and our other SN classes, with an F_1-score of 0.90 ± 0.01 and an SN Ia purity of 0.94 ± 0.01.

6. Comparison with Other Classifiers

We compare Superphot+'s performance (both including and excluding redshift information) with those of three state-of-the-art pipelines from previous works: Superphot (Hosseinzadeh et al. 2020), SuperRAENN (Villar et al. 2020), and ParSNIP (Boone 2021). Unlike Superphot+, all of these previous pipelines require redshift information, and all were originally trained on four-band Pan-STARRS Medium Deep Survey (PS1-MDS) light curves. Both the original Superphot and SuperRAENN papers use random forests for classification, whereas ParSNIP uses a GBM trained with LightGBM. We regenerate light-curve encodings from each pipeline with our ZTF data set for a fair comparison, and use those features to train identical GBMs. This isolates variations in performance as resulting from better light-curve encapsulation by the selection of parametric (for Superphot+ and Superphot) or nonparametric (for SuperRAENN and ParSNIP) features. We train both multiclass and single-class (SN Ia) GBMs for each pipeline's feature set. We also train GBMs using only peak color (A_r, A_g) and redshift information (z, M_r) as a baseline, to isolate performance improvements from light-curve shape information from each classifier.

The resulting accuracies and F_1-scores are shown in Table 4, with example light curves modeled by each pipeline in Figure 17. Training with only color and redshift information yields an accuracy of 0.77 ± 0.01 and F_1-score of 0.56 ± 0.02. Superphot, the precursor to our current pipeline, yields $F_1 = 0.55 \pm 0.02$ and an accuracy of 0.80 ± 0.01. This is only marginally better than only using redshift and color features, indicating that fits suffer tremendously from broad, uniform priors. SuperRAENN yields $F_1 = 0.71 \pm 0.03$ and an accuracy of 0.89 ± 0.01, which is higher than metrics from the original paper. We find that SuperRAENN decodings from the same encoded light curve are unstable and vary depending on the requested number of decoded time stamps. ParSNIP, using a variational autoencoder, performs similarly with a $F_1 = 0.71 \pm 0.03$ and accuracy of 0.89 ± 0.02. It excels at differentiating between SNe Ib/c and SNe I/a since it can nonparametrically encode secondary r-band "bumps" present in some SN Ia light curves; neither Superphot's nor Superphot +'s empirical model can capture these additional peaks.

6.1. Comparisons to the ALeRCE Light-curve Classifier

We next direct our attention to ALeRCE's light-curve classifier, which is the only other publicly running, multiclass SN classifier that does not use redshift information. Unlike Superphot+, which solely classifies within SNe, ALeRCE uses a hierarchical random forest to first distinguish between transients, stochastic sources, and periodic variables (the "top-level classifier"), and then to categorize within these broad categories (Sánchez-Sáez et al. 2021). We use the top-level classifier to select a photometric data set, as described in Section 2.5; here, we explore ALeRCE's SN-specific light-curve classifier ("ALeRCE-SN"). AleRCE-SN also uses the model fit parameters described in Section 3 as part of a larger feature set, but it uses a gradient-descent based algorithm (through Python's `scipy.curve_fit` function) rather than our Bayesian sampling techniques. While faster, this fitting algorithm can yield poorer optimal fits and in turn more misclassifications. We will compare the performance of our classifier without redshift information to that of ALeRCE-SN to demonstrate the benefit of robust fitting techniques and careful class rebalancing.

One difficulty in directly comparing Superphot+ with ALeRCE-SN is that the latter only outputs pseudoprobabilities for four labels: SN Ia, SN Ib/c, SN II, and SLSN. It is unclear how we should treat our Type IIn true and predicted labels when calculating agreement between the two classifiers. During comparison, we only consider the events in our data set also labeled by ALeRCE-SN, and we exclude all spectroscopic SNe

Table 3
Miscellaneous Transient Classifications

Object Type	SN Ia	SN II	SN IIn	SLSN-I	SN Ib/c
SN Iax (12)	0.250 (3)	0 (0)	0 (0)	0 (0)	**0.750** (**9**)
SN Ibn (18)	**0.444** (**8**)	0.278 (5)	0.056 (1)	0.222 (4)	0 (0)
SN IIb (57)	0.123 (7)	0.228 (13)	0.035 (2)	0.018 (1)	**0.596** (**34**)
TDE (51)	0.137 (7)	0.098 (5)	0.216 (11)	**0.549** (**28**)	0 (0)
LBV (6)	0.167 (1)	**0.500** (**3**)	0.167 (1)	0 (0)	0.167 (1)
AGN (45)	0 (0)	0.178 (8)	**0.600** (**27**)	0.200 (9)	0.022 (1)
Total (189)	26	34	42	42	45
Phot. Frac.	0.006	0.041	0.132	0.157	0.059

Note. Summary of how miscellaneous transients are classified by our five-class, redshift-independent classifier, with the highest assigned probabilities in bold. The absolute number of events is shown in parentheses. In general, SNe Iax and SNe IIb are labeled as SNe Ib/c. Most SNe Ibn are classified as SNe Ia. TDEs tend to be grouped with SLSNe-I. LBVs are mostly labeled as SNe II. We also show the fraction of contamination in Superphot+'s predicted data sets from these rarer classes, with the heaviest contamination at 15.7% for predicted SLSNe-I.

IIn. After these cuts, we are left with 5525 events (91.2%) from our data set. This includes events that are not spectroscopically SNe IIn but are photometrically labeled as SNe IIn by Superphot+; for these events, we instead use the label with second-highest pseudoprobability. There are 137 such light curves, with 74, 45, 10, and 10 events relabeled as predicted SNe II, SLSNe-I, SNe Ia, and SNe Ib/c, respectively. After these alterations, we can condense our five-class confusion matrices into four-class confusion matrices for direct comparison with ALeRCE-SN. Derived metrics will be underestimates compared to if we had trained a model without SN IIn labels, as our five-class model sacrifices some performance in all other classes to balance SN IIn performance.

The four-class confusion matrices for ALeRCE-SN and Superphot+ are shown in Figure 18. Using the shared data set, Superphot+ has a F_1-score of 0.66 ± 0.03, which is better than ALeRCE-SN's $F_1 = 0.62 \pm 0.04$. ALeRCE-SN tends to classify light curves with unconstrained fits as SLSNe-I, which drops its class-averaged purity. Superphot+ instead defaults to SN II or SN Ia labels when unsure.

A potential contributor to reduced ALeRCE-SN performance could be its inclusion of an absolute t_0 fit parameter, which is defined relative to a light curve's first observation. The first detection of incomplete or noisy light curves would be significantly offset from the time of SN explosion, making the corresponding t_0 values uninterpretable and therefore biasing ALeRCE-SN's classifications. Superphot+, on the other hand, does not use $t_{0,r}$ at all (only a time delay between bands), so the classifier's behavior when applied to incomplete light curves is more predictable. This effect would be magnified when classifying the photometric data set, which has more partial light curves.

7. New ZTF Photometric Classifications

7.1. Photometric Labels from Superphot+

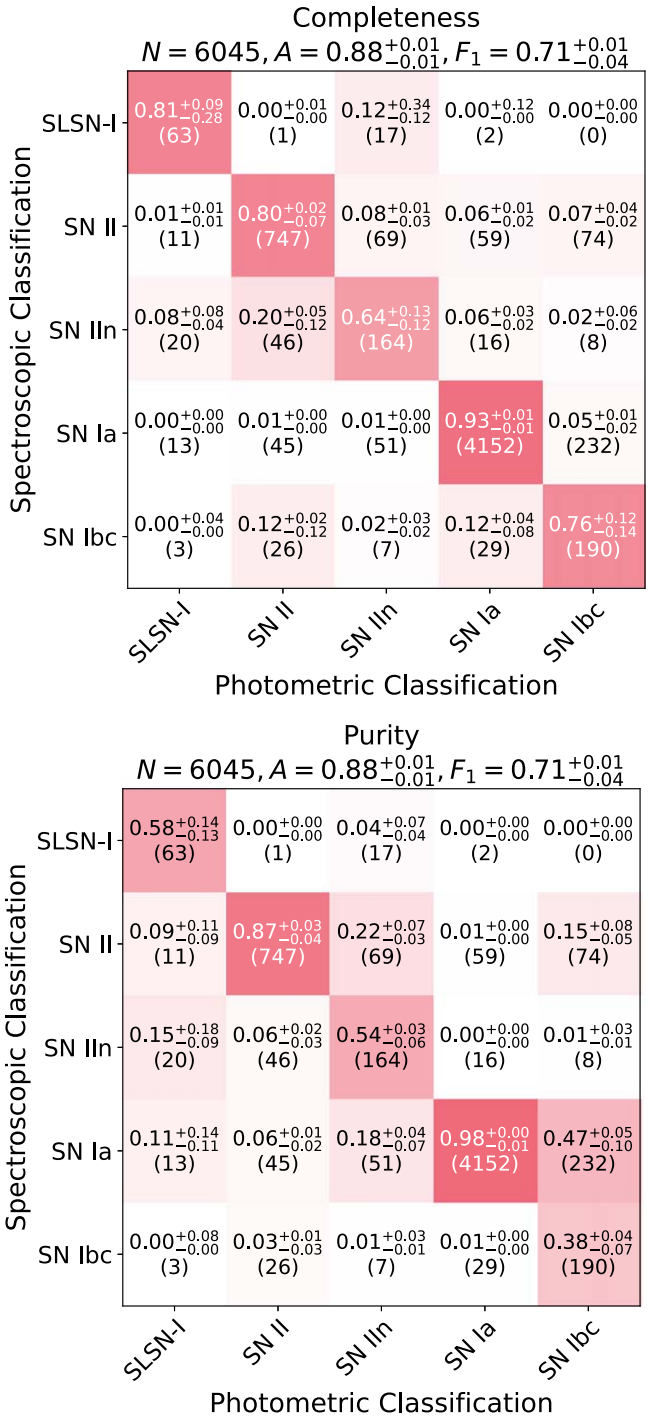
In this section, we apply Superphot+, trained without redshift information, to 3558 SN-like light curves that have not been spectroscopically classified. Light curves in this "photometric" data set, as collated in Section 2.5, pass the quality cuts described in Section 2.3 and the reduced chi-squared fit cut described in Section 3.1. The first few classification probabilities are shown in Table 5; a full version of this table is available on Zenodo.

Superphot+ classifies 58.6% of the photometric data set as SNe Ia, 13.3% as SNe II, 7.4% as SNe IIn, 8.0% as SLSNe-I, and 12.7% as SNe Ib/c. The predicted SN Ia fraction is lower, and the SLSN-I/SN IIn/SN Ib/c fractions are higher than those of the spectroscopic data set. For example, we only expect 1.4% predicted SLSNe-I and 4.3% predicted SNe Ibc in order to match the spectroscopic class fractions. If we assume the photometric data set's true class breakdown matches the spectroscopic's true class breakdown (which may not be the case), we conclude that Superphot+ classifies many true SNe Ia as other classes, lowering CC SN purities and SN Ia completeness. Including redshift information as described in Section 5.5 yields a negligible change in photometric class fractions, so we do not also analyze that variant here.

In Figure 19, we compare Superphot+'s spectroscopic and photometric class fractions with those from other SN data sets: the Pan-STARRS Medium-Deep Survey (PS1-MDS) subset used in Hosseinzadeh et al. (2020), and the Young Supernova Experiment Data Release 1 (YSE-DR1; Aleo et al. 2023), which both use measurements from the Pan-STARRS telescopes. These Pan-STARRS fractions are quite similar, with the main difference being YSE-DR1's increased SN II (16.1% versus 22.8%) and decreased SLSN-I (1.4% versus 0.4%) fractions. Because our spectroscopic sample is dominated by ZTF BTS SNe, its class breakdown is quite similar to that of ZTF BTS, as detailed in Section 2.6. Therefore, we do not include a separate column for ZTF BTS in the figure.

7.2. Comparison with ALeRCE-SN's Predictions

We also apply ALeRCE-SN to our photometric data set for comparison, adding the resulting fractions to Figure 19. Because the photometric data set is collated from ALeRCE's top-level predictions, every event in this data set has been labeled by ALeRCE-SN. ALeRCE-SN classifies 14.6% of the photometric set as SLSNe, which is a higher fraction than that from Superphot+ but about equal to Superphot+'s combined SLSN-I and SN IIn predicted fraction. This is not surprising given ALeRCE-SN's low SLSN purity within the spectroscopic data set, and ALeRCE-SN potentially labeling SN IIn contenders as SLSNe. Both photometric class compositions have fewer SNe Ia than expected in the photometric data set. ALeRCE also underclassifies objects as SNe II (9.5%) compared to the spectroscopic 16.1%.



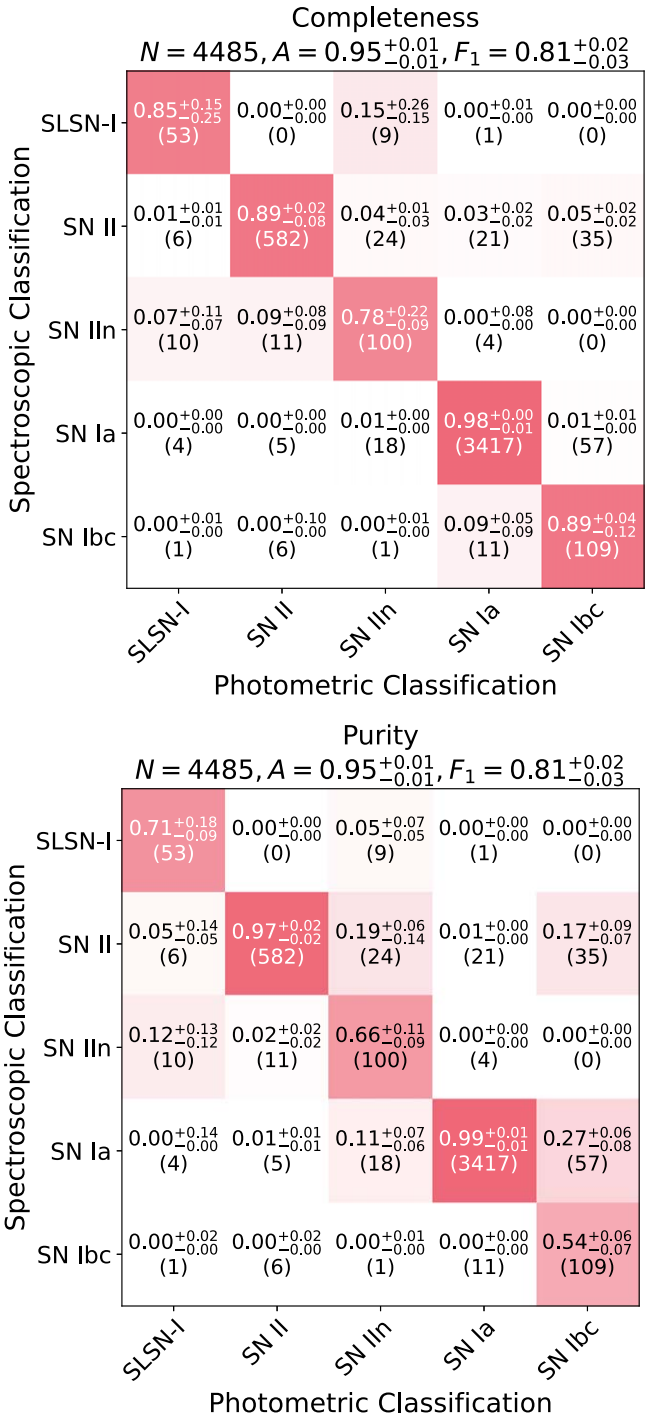


Figure 15. The completeness and purity confusion matrices for the redshift-inclusive classifier. Adding redshift information improves the F_1-score from 0.61 ± 0.02 to 0.71 ± 0.02. There is significant improvement in SLSN-I, SN II, and SN IIn classification, which demonstrates the importance of luminosity in distinguishing between light curves of these classes.

Figure 16. Same as Figure 15, but only including light curves classified with confidence greater than or equal to 0.7. While high-confidence performance is worse than when excluding redshift for the same cutoff, a much higher fraction (4485 out of 6045) of light curves are classified with high confidence compared to the redshift-exclusive classifier. This highlights the potential for inaccurate but confident classifications from redshift biases.

We can now compare Superphot+'s and ALeRCE-SN's agreement when labeling the spectroscopic and photometric data sets. First, we generate the expected agreement matrix from each classifier's spectroscopic confusion matrix, as derived in Appendix B of Hosseinzadeh et al. (2020). This predicts classification consistency assuming the two classifiers' latent spaces are completely independent:

$$A = P^T_{\text{Superphot+}} C_{\text{ALeRCE-SN}}. \quad (7)$$

Here, P is the purity matrix, and C is the completeness matrix. The expected agreement score, which is simply the fraction of all samples expected to receive the same label from both classifiers, is 0.69 ± 0.03.

Next, we generate true agreement matrices from our reduced spectroscopic and photometric data sets, which compare how both classifiers actually labeled events. The agreement matrices

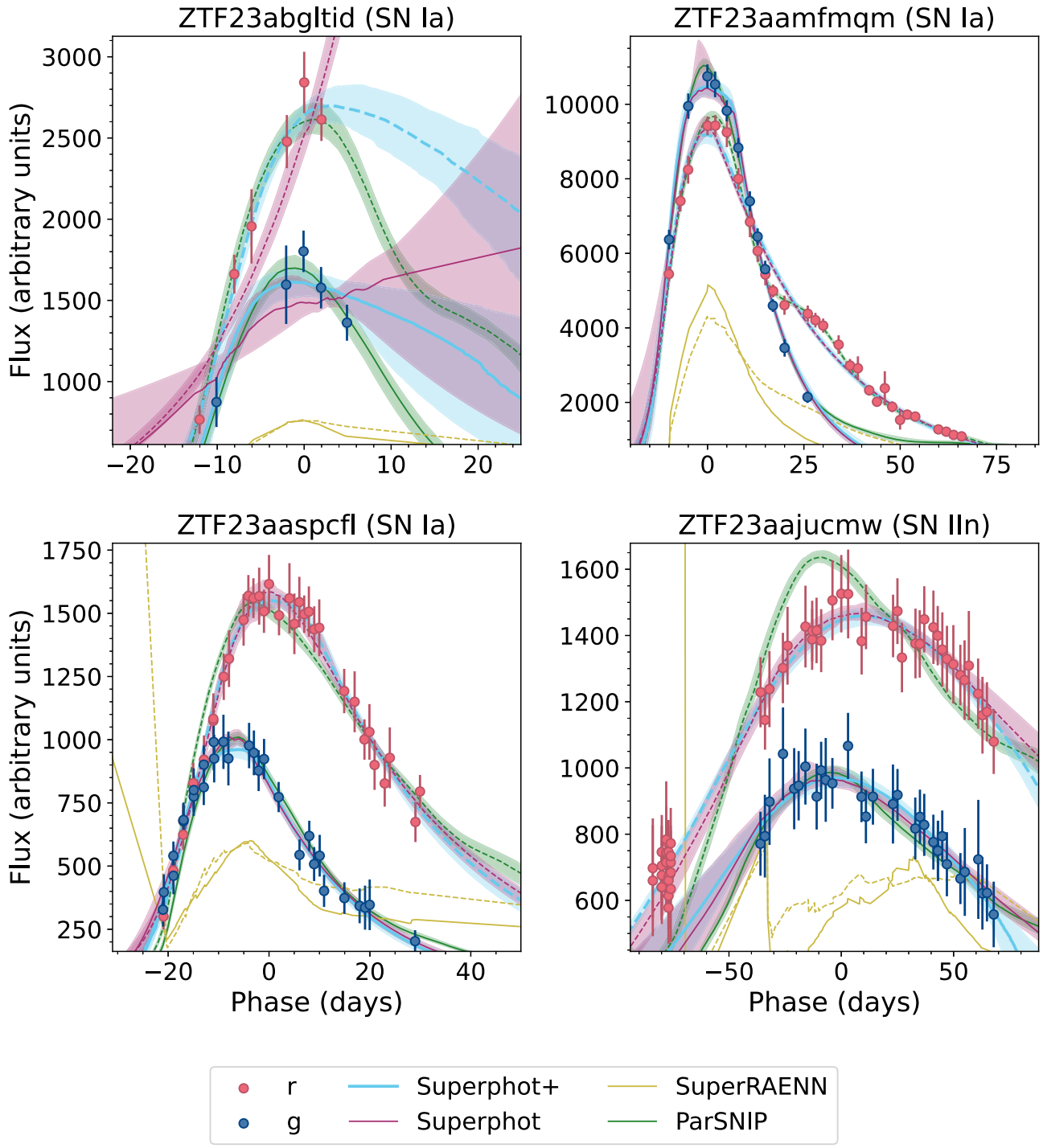


Figure 17. Three SN Ia and one SN IIn light curve, in that order, fitted with each of the four pipelines compared in Table 4. The top left light curve is fit reasonably by ParSNIP and Superphot+, but Superphot suffers from its too broad priors. The top right light curve is fit fairly well by Superphot+ and Superphot, but ParSNIP is the only pipeline that adequately captures the SN Ia's secondary *r*-band bump. All pipelines except SuperRAENN fit the bottom left light curve adequately. The bottom right (SN IIn) light curve has an exceptionally slow rise timescale and suffers from uneven rise sampling; the parametric classifiers Superphot+ and Superphot recreate this light curve best. While SuperRAENN better recreates light curve at training time, it suffers from unstable decoding when fed a uniform time array for plotting.

are shown in Figure 20. The true agreement score for the spectroscopic data set is 0.82 ± 0.02, which is much higher than the expected agreement score. All agreement scores are higher than expected, with the largest improvements among SNe Ia and SNe II. This improved agreement matches expectations, as the features used for classification are very much not independent; the same model is used to generate fit parameters for both pipelines.

To generate the agreement matrix for the photometric data set, we first reallocate Superphot+'s SN IIn predictions just as we did in Section 6.1. This reassigns 262 light curves to other predicted classes, mainly SNe II and SLSNe. The resulting photometric agreement matrix, with an overall agreement score of 0.72 ± 0.00, is shown in Figure 20 (bottom). To get confidence intervals, we feed our photometric data set into each *K*-fold model separately, generating 10 different agreement

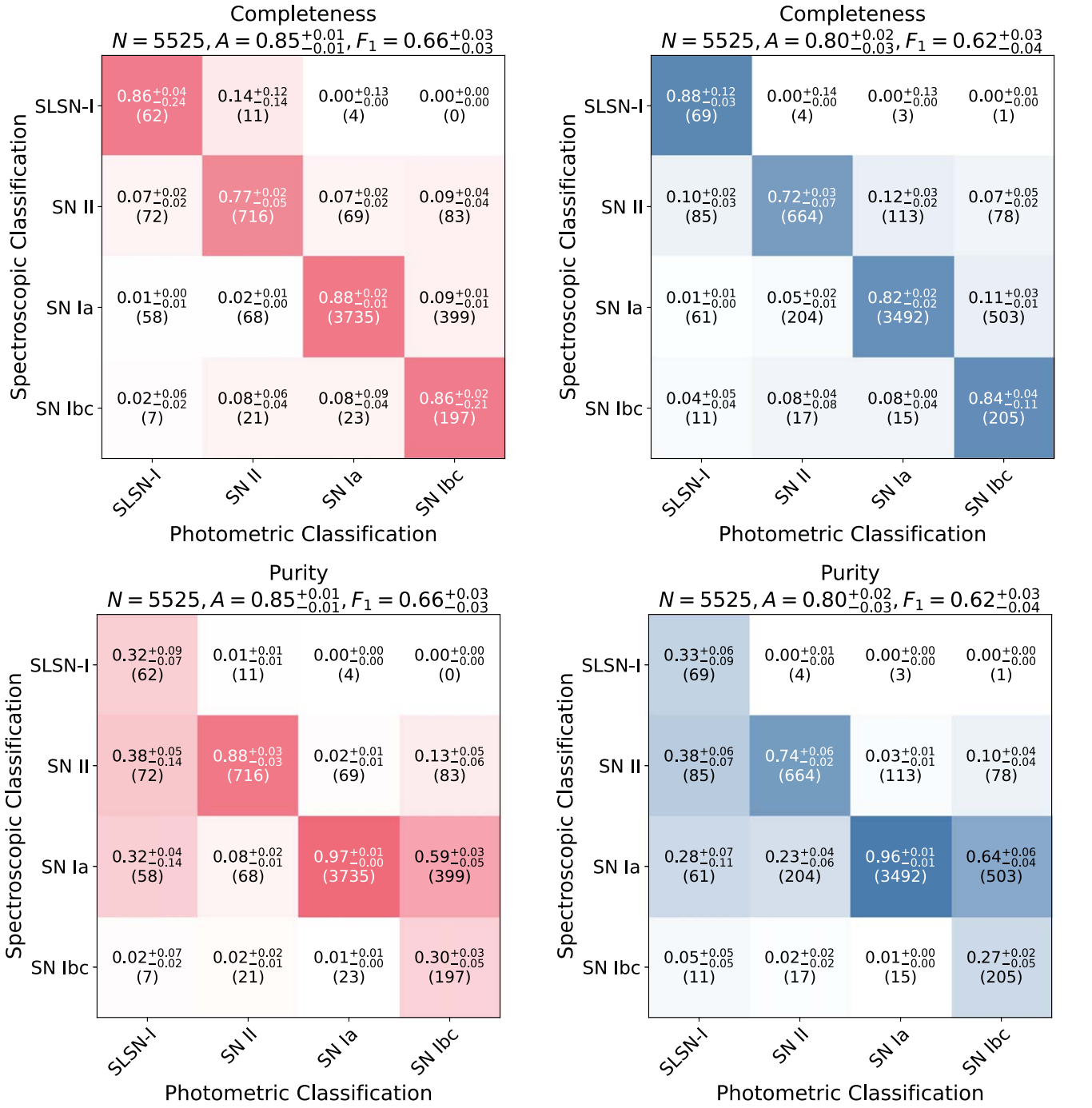


Figure 18. The four-class confusion matrices for Superphot+ (left) vs. ALeRCE-SN (right). The Superphot+ confusion matrices are condensed into four classes by combining the second-highest probability label for objects predicted to be SNe IIn, and excluding samples with SN IIn true labels. This is done because ALeRCE-SN does not include an SN IIn label. Even with this regrouping, the Superphot+ F_1-score is higher than ALeRCE-SN's.

matrices. This aggregate true agreement matrix closely matches the expected agreement matrix, and somewhat mirrors the true agreement matrix for the spectroscopic set, with many ALeRCE-SN SLSNe-I classified as SNe II by Superphot+. However, the photometric agreement matrix magnifies the SLSN versus SN II disagreement we see in the spectroscopic agreement matrix. This could be interpreted as an ALeRCE-SN bias toward SLSN predictions or a Superphot+ bias toward SN II predictions, especially for uncertain light curves. Seeing how Superphot+ does not often assign SN II labels to partial light curves (see Section 5.2), an ALeRCE-SN SLSN bias is more likely. There is similar agreement within samples ALeRCE-SN labels as SNe Ib/c compared to the spectroscopic data set.

7.3. Correcting Class Fractions for Spectroscopic Bias

Like any classifier, Superphot+ has its biases, and we can use those biases from classification of the spectroscopic data set to "correct" our photometric data set's class fractions. This will better inform us on potential systematic differences between the spectroscopic and photometric data sets. We can correct our photometric class fractions by considering the fraction of each

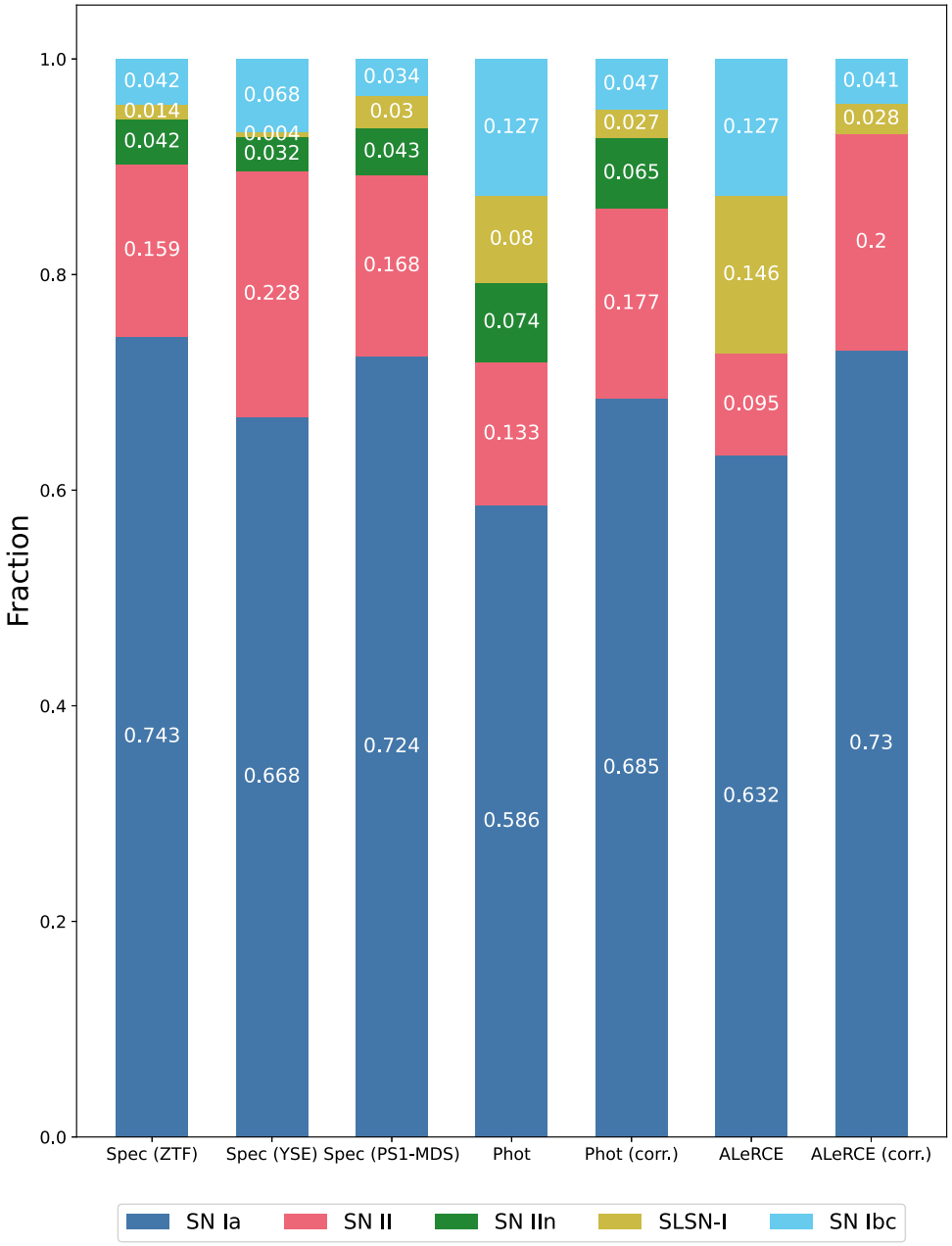


Figure 19. The class fractions of our spectroscopic data set ("Spec ZTF") compared with those from the Young Supernova Data Release 1 data set ("Spec YSE") and the PS1-MDS ("Spec PS1-MDS") set used in Villar et al. (2020). We compare these fractions with those derived from Superphot+'s and ALeRCE-SN's photometric classifications ("Phot" and "ALeRCE," respectively). Additionally, the corrected class fractions (using the confusion matrices from validation) are included for comparison.

predicted class that belongs to other classes (i.e., the per-class purities). For example, the purity matrix from Figure 7 demonstrates that, of the spectroscopic SNe classified as Type IIn, $51\% \pm 9\%$ are true SNe IIn, $28\% \pm 7\%$ are true SNe II, and $15\% \pm 5\%$ are true SNe Ia. We can use these values to adjust the corresponding class fractions in the photometric data set (following Hosseinzadeh et al. 2020; Villar et al. 2020; Aleo et al. 2023):

$$\mathrm{frac}_{i,\mathrm{corr}} = \sum_j \mathrm{frac}_j P_{i,j}. \tag{8}$$

Here, $P_{i,j}$ is the ith row and jth column of our purity matrix, or the fraction of SNe predicted to be class j that are actually from class i. Computing this reallocation for all five classes, we can convert our original class fractions to new, "corrected" class fractions. The new fractions better account for the biases that we know exist in our classifier. If the corrected photometric class fractions still differ from the spectroscopic class fractions, then there are potentially additional biases at play. Either the differing data sets are causing our classifier to behave differently than expected, or the spectroscopic and photometric class fractions intrinsically differ.

The results of these corrections are shown in Figure 19. Correcting Superphot+'s biases increases the SN Ia class fraction to 68.5%, closer to the spectroscopic SN Ia class fraction. We also compute corrections for ALeRCE-SN, which

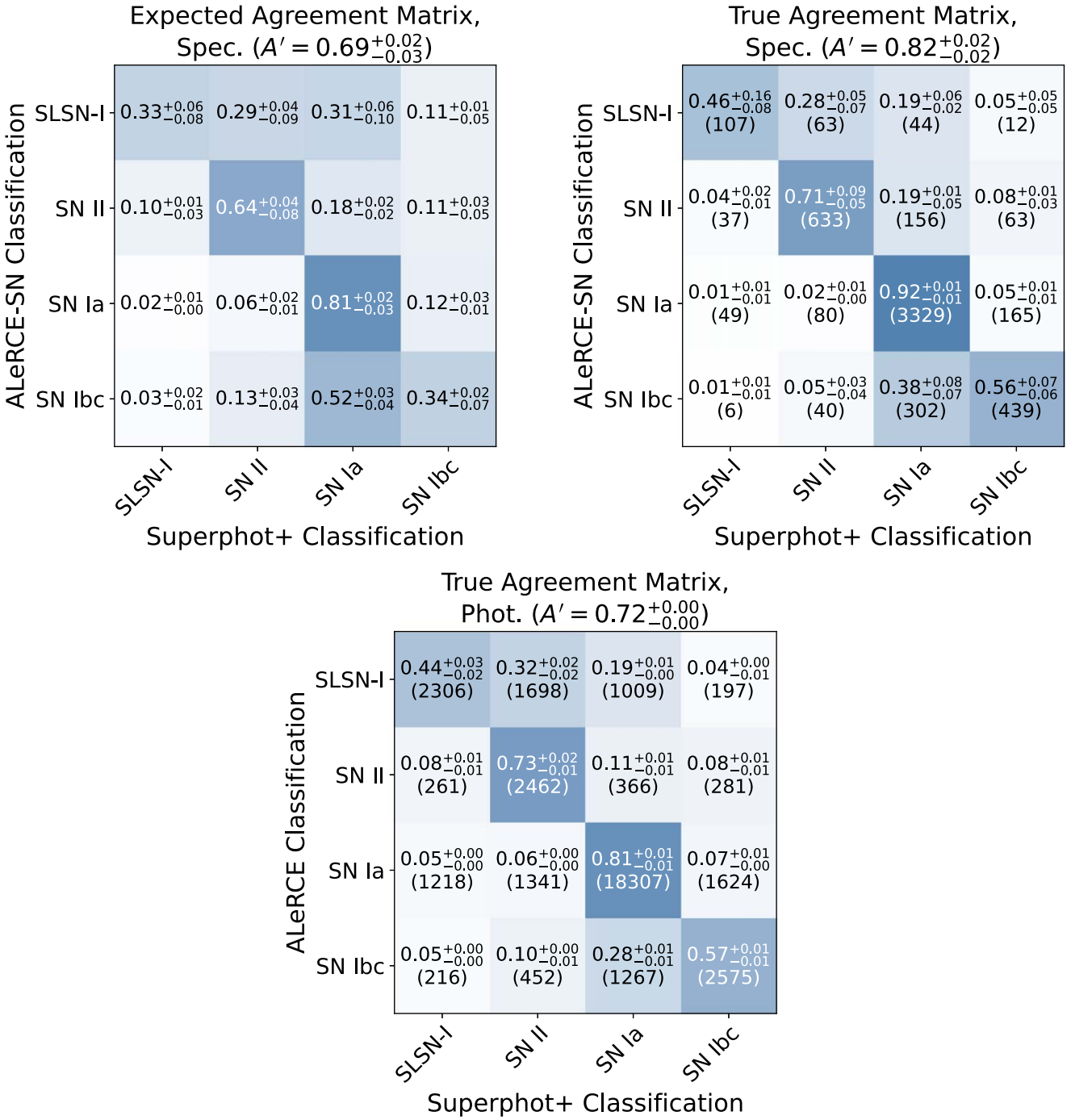


Figure 20. The expected (top left) and true (top right) agreement matrix for the spectroscopic data set between Superphot+ and ALeRCE-SN, as well as the true agreement matrix for the photometric data set (bottom). The classifiers agree more than expected when labeling objects as SNe II, SLSNe-I, and SNe Ib/c. The photometric set's agreement is worse than the spectroscopic set's, which may be due to contamination by nontransients or light-curve differences compared to light curves with spectroscopic classifications. All three matrices suggest Superphot+ is more likely to classify unsure light curves as a common type (SN Ia or SN II), whereas ALeRCE-SN favors SLSN labels, validating Superphot+'s higher purity.

also brings its photometric fractions closer to the spectroscopic class fractions. For both classifiers, SNe II and SLSNe-I are still overrepresented among the corrected photometric fractions, while SNe Ia are underrepresented. The corrected ALeRCE-SN SLSNe-I fraction is still double that of the spectroscopic data set. We note that while correcting for expected biases shifts both of our photometric fractions closer to the spectroscopic sample's, especially for ALeRCE-SN, we cannot use that information to modify individual predictions. Therefore, while ALeRCE-SN's corrected ratios are closer to the spectroscopic data set's ratios (excluding SNe IIn), Superphot+ has less disparate uncorrected fractions among SLSNe-I and SNe II and therefore less biased individual classifications.

8. Conclusions and Discussion

We have presented a novel photometric classifier, Superphot +, that does not require redshift information to assign one of five SN labels. We apply Superphot+ to ZTF light curves, but

Table 4
Comparison to Other Classifiers

Classifier	Multiclass Accuracy	Multiclass F_1	Binary Accuracy	Binary F_1	Citation
Superphot+ (no z)	0.83 ± 0.01	0.61 ± 0.02	0.93 ± 0.01	0.90 ± 0.01	This Work
Superphot+ (with z)	0.88 ± 0.01	0.71 ± 0.02	0.94 ± 0.01	0.92 ± 0.01	This Work
Superphot	0.80 ± 0.01	0.55 ± 0.02	0.90 ± 0.01	0.88 ± 0.02	Hosseinzadeh et al. (2020)
SuperRAENN	0.89 ± 0.01	0.71 ± 0.03	0.94 ± 0.02	0.93 ± 0.02	Villar et al. (2020)
SuperRAENN (only RAENN)	0.83 ± 0.02	0.62 ± 0.04	0.92 ± 0.02	0.89 ± 0.02	Villar et al. (2020)
ParSNIP	0.89 ± 0.02	0.71 ± 0.03	0.96 ± 0.01	0.95 ± 0.01	Boone (2021)
Only z, $M_{\rm peak}$, and A_g	0.77 ± 0.01	0.56 ± 0.02	0.89 ± 0.01	0.86 ± 0.02	This Work

Note. Comparison of Superphot+'s multiclass and SN Ia binary classification metrics (with and without redshifts) with those from training identical LightGBMs on feature sets from other SN pipelines that use redshift information. All classifiers are retrained on identical ZTF data sets. The ALeRCE pipeline does not use redshift information and is therefore not included in this table.

Table 5
Superphot+ Probabilities Assigned to the Photometric Data Set

ZTF Name	IAU Name	Fit Reduced χ^2	ALeRCE-SN Label	p(SN Ia)	p(SN II)	p(SN IIn)	p(SLSN-I)	p(SN Ib/c)
ZTF18aaanaev	2022wkv	0.607	SN Ibc	0.226	0.161	0.104	0.129	**0.380**
ZTF18aabdajx	2018mac	0.914	SN Ia	0.277	0.116	0.116	0.467	0.024
ZTF18aabeszt	...	0.382	SN Ia	0.933	0.026	0.014	0.011	0.016
ZTF18aacsudg	2019pxz	0.876	SN II	0.029	0.241	0.249	0.031	**0.451**
ZTF18aaczmob	2023ecq	0.549	SN Ia	0.434	0.071	0.041	0.012	**0.441**
ZTF18aadrhsi	2018hzz	0.869	SN Ia	0.301	0.158	0.089	0.358	0.094
ZTF18aaexyql	2019tpu	0.596	SN Ibc	0.121	0.082	0.029	0.021	**0.747**
ZTF18aagtwyh	2021oud	0.606	SN Ia	**0.445**	0.257	0.132	0.029	0.136
ZTF18aahsuyl	2021jqj	0.569	SN Ia	**0.934**	0.022	0.014	0.005	0.025
ZTF18aahyute	2021rcu	0.706	SLSN-I	0.029	0.252	**0.691**	0.013	0.015

Note. The first 10 rows of our photometric classification results of the nonspectroscopically classified data set, which are sorted in alphabetical order. The highest probabilities per event are in bold. Superphot+ and ALeRCE agree on 50% of these classifications.

(This table is available in its entirety in machine-readable form in the online article.)

emphasize its easy adaptation to other current and future surveys. Superphot+ uses nested sampling and SVI to rapidly fit light curves in all bands simultaneously, using correlated priors to optimize fits for inconsistently sampled light curves. These fit parameters are then oversampled to correct for class imbalances, and used as training features for a GBM. On the ZTF data set of 6061 spectroscopically classified events, Superphot+ achieves a class-averaged F_1-score of 0.61 ± 0.02. We explore the addition of redshift information within Superphot+, finding better overall performance with $F_1 = 0.71 \pm 0.02$, but worse performance when only considering high-confidence predictions. Superphot+'s performance exceeds or matches previous classifiers that require redshift information, and it outperforms ALeRCE's light-curve classifier. This makes Superphot+, when trained on our ZTF data set, the best-performing publicly available pipeline that does not require redshifts.

We assign new photometric labels to 3558 SN-like ZTF light curves without spectroscopic labels, classifying 58.6% as SNe Ia, 13.3% as SNe II, 7.4% as SNe IIn, 8.0% as SLSNe-I, and 12.7% as SNe Ib/c. Correcting for classifier bias updates these fractions to 68.5% SNe Ia, 17.7% SNe II, 6.5% SNe IIn, 2.7% SLSNe-I, and 4.7% SNe Ib/c. Before corrections, Superphot+ and ALeRCE-SN assign equal predicted SN Ib/c class fractions, and Superphot+ assigns SN IIn and SLSN-I labels approximately as frequently combined as ALeRCE-SN assigns SLSN labels. There is evidence of intrinsic differences within our spectroscopic and photometric data sets, supported by the lower than expected SN Ia prevalence from both classifiers.

Nonparametric light-curve modeling is becoming more prevalent in the literature due to the computational bottleneck of Bayesian fitting to parametric models. However, with the use of SVI and correlated priors, we have demonstrated that Superphot+ is computationally fast enough to keep up with both ZTF and expected LSST alert streams. In addition, parametric fitting benefits from enforcing light-curve structure a priori, improving modeling of very sparsely sampled lightcurves. Therefore, we argue that parametric classifiers are viable for the Rubin era, and we recommend Superphot+ for real-time ZTF and LSST photometric classification.

Superphot+ has been integrated as a filter of the ANTARES alert broker (Saha et al. 2014), where we have incorporated the early-phase variant detailed in Section 5.2. In place of a top-level classifier to isolate SNe, we cross-check star and AGN catalogs, and ignore light curves that were first observed over 1 yr ago. This will undoubtedly lead to classification of non-SNe like AGNs and TDEs, but our existing cuts sufficiently reduce nightly SN counts such that we can visually inspect "strange" events. We save the fit parameters as light-curve properties on ANTARES, enabling further downstream tasks like anomaly detection or real-time inference. We leave this exploration for future work.

In addition to the ZTF ANTARES filter detailed above, Superphot+ can readily be applied to LSST-like data when the

survey begins in 2025. Indeed, the empirical fitting and subsequent classification described in this work have already been modified for six-band simulated LSST data to be used in the Extended LSST Astronomical Time-series Classification Challenge[13] (ELASTICC). While a simple, ALeRCE-like top-level classifier was designed for this challenge, it has not been optimized to the same extent as Superphot+'s SN classification. The results of the ELASTICC challenge and Superphot +'s application to LSST-like datastreams are left to future work.

Acknowledgments

We thank the anonymous referee for the valuable review and feedback. The Villar Astro Time Lab acknowledges support through the National Science Foundation under AST-2433718, AST-2407922 and AST-2406110, as well as an Adamant Fellowship for Emerging Science Research. This work is additionally supported by NSF under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). K.d.S. and V. A.V. acknowledge support from LSSTC through grant 2023-SFF-LFI-02-Villar. This research received support through Schmidt Sciences. K.d.S. and G.H. thank the LSST-DA Data Science Fellowship Program, which is funded by LSST-DA, the Brinson Foundation, and the Moore Foundation; their participation in the program has benefited this work.

Software: ANTARES (Saha et al. 2014; Narayan et al. 2018; Matheson et al. 2021; ANTARES 2024), ALeRCE (Förster et al. 2021; ALeRCE 2022), Astropy (Astropy Collaboration et al. 2013, 2018, 2022; Robitaille et al. 2023), `dustmaps` (Green 2018; Green et al. 2023), `dynesty` (Speagle 2020; Koposov et al. 2023), `extinction` (Barbary 2017), FLEET (Gomez et al. 2020a, 2020b), `imbalanced-learn` (Lemaître et al. 2017; Imbalanced-learn 2024), `iminuit` (Dembinski et al. 2023), JAX (Bradbury et al. 2018; Frostig et al. 2018), `light-curve` (Malanchev 2021; Malanchev et al. 2021), LightGBM (Ke et al. 2017; Microsoft Corporation 2023), Matplotlib (Hunter 2007; Matplotlib 2023), NumPy (Harris et al. 2020; NumPy Developers 2024), NumPyro (Bingham et al. 2019; NumPyro 2019; Phan et al. 2019), `pandas` (McKinney 2010; The pandas development team 2020), ParSNIP (Boone 2021; Boone & Malanchev 2022), `scikit-learn` (Pedregosa et al. 2011; Grisel & Mueller 2024), Superphot+ (de Soto et al. 2024), Superphot (Hosseinzadeh et al. 2020; Hosseinzadeh 2021), SuperRAENN (Villar et al. 2020; villrv 2020), Pytorch (Paszke et al. 2019), Tensorflow Probability (Dillon et al. 2017).

Appendix A
Light-curve Clipping

Many ZTF light curves suffer from improper template subtraction, leading to excessive constant flux data points long after the light curve has set. To remove these data points, we first calculate the "maximum" absolute slope of the light curve by drawing a line connecting the maximum flux measurement to the last observation of the light curve. If there is a chunk of data points at the end of the light curve with a sufficiently flat slope relative to this maximum slope, then we assume it to be an artifact of template subtraction, and remove it. To accomplish this, we start from the last observation and find the earliest observation such that the absolute slope between these points is less than 0.2 times the value of the "maximum" slope. The 0.2 cutoff was determined empirically by experimenting with various cutoff slopes. We then remove the entire section of data points that follows the observation. This process is done separately for the *g* and *r* bands, as the differences in band amplitudes correspond to differences in maximum slope from the last data point.

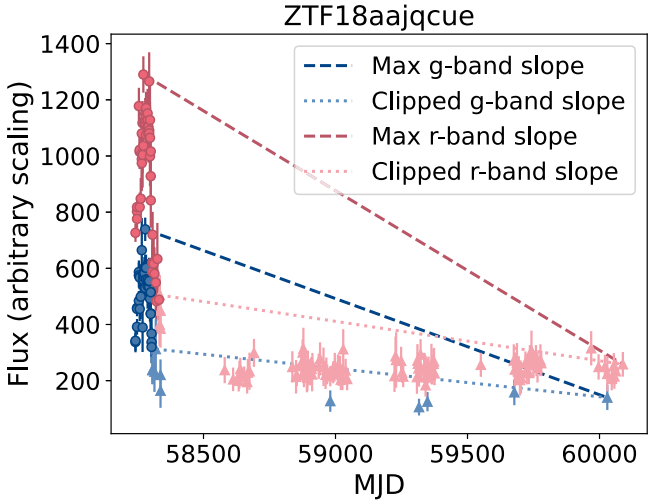


Figure 21. Example ZTF light curve for an SLSN-I with an artificial extended "tail" ($|\text{slope}| \leqslant 0.2|\text{slope}_{\text{max}}|$); this late-time excess flux is due to improper template subtraction. These tails are clipped to maintain fitting integrity and correct improper template subtraction. The maximum and tail slopes are indicated by dashed/dotted lines, respectively, and the clipped data points are represented by triangles.

An example of this method applied to an SLSN-I is shown in Figure 21. Due to improper template image subtraction, a constant nonzero flux is measured in both bands for an additional 1500 days after the SN's fall. As our fit model does not allow for a vertical offset, the incorrectly subtracted tail will lead to poorer fit parameters and artificially high chi-squared values. Our light-curve clipping method successfully removes the tails (triangular data points) by determining that the flux change within the tail is substantially less than the flux change from peak to the last data point.

We do note that one downside of this method is that, in the case of transients with rapid declines, our light-curve clipping strategy may remove a good fraction of postpeak points. For example, the radioactive tails of SN IIP light curves following plateau are sometimes excessively clipped. However, visual inspection shows that enough points are maintained to adequately constrain the fall timescale. Similarly, SNe Ia exhibit secondary "bumps" in the *r* band (e.g., see Kasen 2006 for more on this well-known phenomena), which are sometimes clipped. Finally, there is the potential for excessive clipping following spurious bright data points, but we do not see this in practice among the transients that pass our quality cuts.

Our procedure clips at least one data point from 4324 out of 9526 light curves (45%) in our unpruned spectroscopic data set. The histogram of number of points clipped in each band among those clipped light curves is shown in Figure 22. While most light curves have less than 10 points clipped per band, there is a long tail that extends beyond the figure up to a couple hundred. The light curves with an excessive number of points removed are those showing faulty template subtraction, with many points of constant flux extending out hundreds of days. The

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

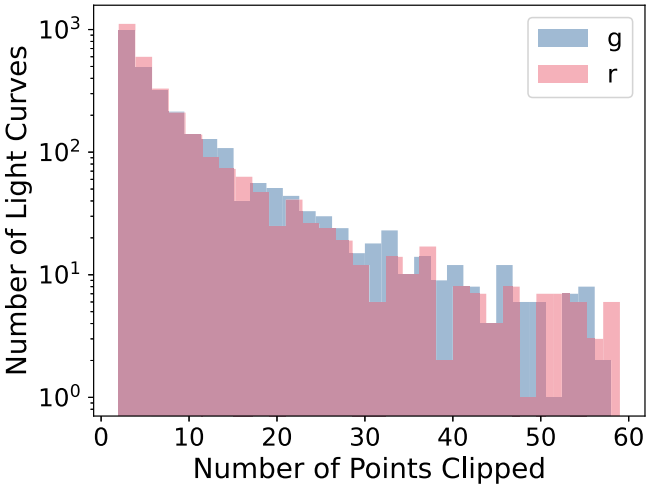


Figure 22. A histogram of the number of points clipped in each band from the light curves in our spectroscopic data set before pruning.

median number of points clipped is six in the g band and five in the r band. Data quality cuts are applied to both data sets after light curves are clipped. Only analyzing spectroscopic light curves that pass our quality cuts yields a near identical number of clipped points distribution, meaning that we are not preferentially removing clipped light curves from the data set. In contrast, only 24% of the photometric data set is clipped, with the median number of clipped points in either bands within this subset being four. This is expected, as light curves without spectroscopic follow-up tend to be more sparsely sampled and therefore have less excess points long after peak to be clipped.

Appendix B
Sampler Selection

While gradient-descent based optimizers (e.g., the Levenberg–Marquardt algorithm) can be used to rapidly obtain best-fit parameters, the resulting fits are often suboptimal, and we cannot properly incorporate the $\sigma_{\rm extra}$ parameter during fitting. In contrast, traditional Metropolis–Hastings Markov Chain Monte Carlo methods better explore the posterior space but take longer to converge and struggle computationally with the piecewise discontinuity at $t = t_0 + \gamma$. Therefore, to optimize the computational efficiency of our pipeline while still maintaining accuracy, we explore a variety of alternate sampling techniques to empirically fit our light curves. These include the following:

1. *Importance nested sampling (Feroz et al. 2019).* Nested sampling uses shrinking ellipsoids to constrain regions of highest posterior density. Importance nested sampling uses information from previous as well as current sampled points to speed up convergence. In our code, we assume unimodality of the posterior space and use a single constraining ellipsoid (Mukherjee et al. 2006). We sample each light curve with 50 live points, 5000 maximum iterations, and a stopping criterion of $\Delta \log Z \leqslant 0.5$. New live points are sampled with random walks from previous points (Skilling 2006), which is more effective than uniform or slice sampling for our problem's dimensionality.
2. *No U-Turn Sampler (Hoffman & Gelman 2011).* No U-Turn Sampler (NUTS) is a Hamiltonian Monte Carlo (Brooks et al. 2011) sampler, which creates an analog between finding posterior density maxima and minimizing the potential energy of a Hamiltonian system. This technique allows efficient and sometimes large steps across the parameter space. It makes no unimodality assumptions. The original Superphot (Hosseinzadeh et al. 2020) uses PyMC3's implementation of NUTS (Salvatier et al. 2016).
3. *Stochastic variational inference (Hoffman et al. 2012).* SVI approximates the posterior by assuming an approximate posterior shape a priori, which in our case is a multidimensional Gaussian. Our loss function is the negative of the evidence lower bound, which, when maximized, minimizes the disparity between our approximate and the true posterior distributions. Approximating the posteriors in this way significantly speeds up convergence compared to the other two methods, but SVI also tends to underestimate the variance of the marginal distributions.
4. *MIGRAD (Fletcher 1970).* MIGRAD is a very fast variable metric method, which iteratively approximates the negative log likelihood distribution as a multidimensional quadratic function. It performs very well near the true solution, but is prone to get stuck in local minima.
5. *CERES (Agarwal et al. 2023).* CERES, like `scipy.curve\inferior fit`, uses the Levenberg–Marquardt algorithm (a combination of the Gauss–Newton method and gradient descent) to iteratively minimize the least-squares error of each fit. While the fastest of the algorithms, it is very prone to getting stuck in local minima or high-loss regions. This algorithm does not incorporate any $\sigma_{\rm extra}$ parameters, or prior distribution information (except for the parameter limits, and means for the initial guess). We include it here to demonstrate the shortcomings of gradient-descent based optimizers for our problem.

Our nested sampling script is implemented using the Python package `dynesty` (Speagle 2020), whereas the NUTS and SVI algorithms are implemented using the `numpyro` package (Bingham et al. 2019; Phan et al. 2019). The latter uses the JAX (Frostig et al. 2018) backend to speed up differentiation and fuse numerical functions, improving runtime drastically. MIGRAD is called through `iminuit` (Dembinski et al. 2023), which is a Python interface to the Minuit2 C++ library (James & Roos 1975). Minuit2 switches to a simplex algorithm (Nelder & Mead 1965) if MIGRAD cannot return a successful fit. The CERES fits are run through the `light-curve` (Malanchev et al. 2021) package's `VillarFit` subroutine.

All techniques listed except MIGRAD and CERES are built on a Bayesian framework, in which the likelihood $p(\boldsymbol{f}|\boldsymbol{\theta})$, prior $\pi(\boldsymbol{\theta})$ and posterior $p(\boldsymbol{\theta}|\boldsymbol{f})$ probabilities are related following Bayes theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{f}) \propto p(\boldsymbol{f}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{B1}$$

where $\boldsymbol{\theta}$ is the set of 14 model parameters described above, and $\boldsymbol{f}$ is the set of observed fluxes. We model the likelihood of each observation $p(f_i|\boldsymbol{\theta})$ as its probability if drawn from the multidimensional Gaussian with mean $F(t_i, \boldsymbol{\theta})$ and variance $\sigma_i^2 + \sigma_{\rm extra}^2$. Here, $F(t_i, \boldsymbol{\theta})$ is the model flux calculated for parameters $\boldsymbol{\theta}$ at time t_i.

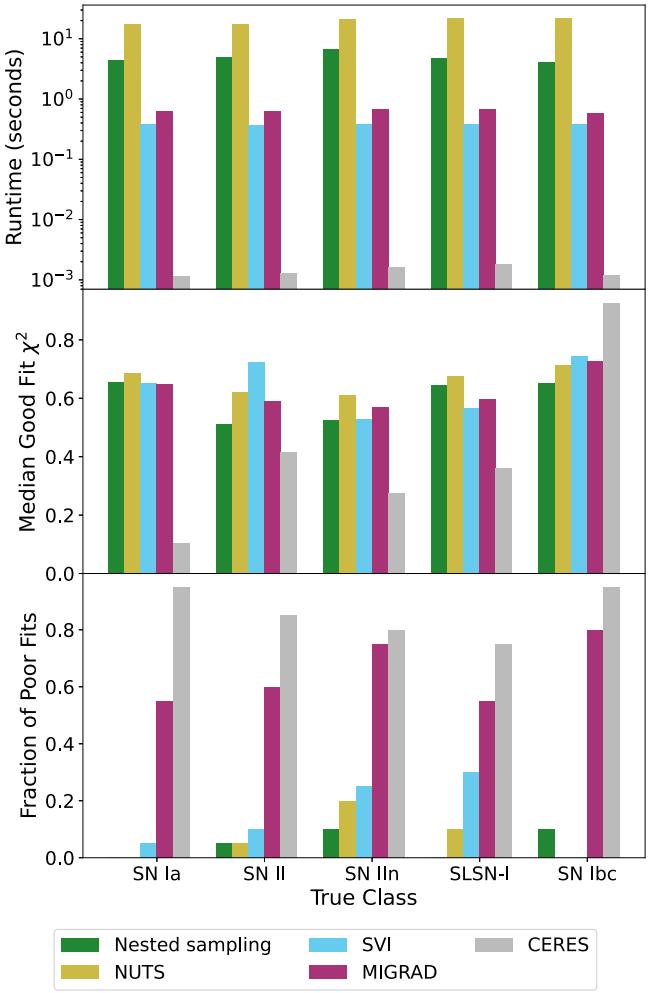


Figure 23. Comparison of the mean fitting runtime, median reduced χ^2 value, and fraction of $\chi^2_{\rm red} > 1.2$ fits from all available Superphot+ samplers, averaged across 20 light curves of each true spectroscopic class. Nested sampling and NUTS are the slowest but most accurate options, with almost no poor fits. SVI is much faster and still correctly fits most light curves. MIGRAD and CERES struggle with the enforced joint constraints inherent during fitting, so fail to fit >50% of the light curves.

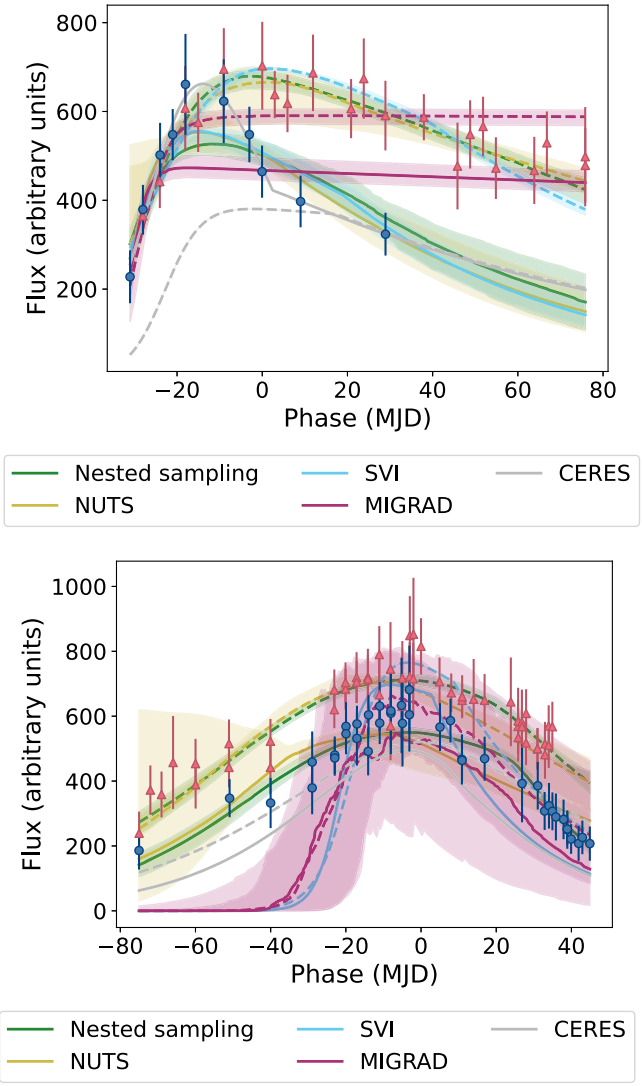


Figure 24. Median model fluxes (solid line) along with the 16th and 84th percentiles (shaded region) for each sampler in Superphot+, for two light curves in our data set. CERES and MIGRAD fail at fitting both light curves, whereas SVI only fits the top light curve correctly. Only NUTS and nested sampling fit both light curves well.

Bayesian techniques tend to be slower than gradient-based methods as they attempt to map out the entire posterior probability space rather than just find a nearby minimum within the space. However, they do tend to better handle problems with joint parameter constraints, which applies to our piecewise model. To ensure the piecewise transition happens after a (potentially infinitesimally small) plateau region, we enforce that the derivative of the first piecewise portion be $\leqslant 0$ at $t - t_0 = \gamma$. This works out to

$$e^{\frac{-\gamma}{\tau_{\rm rise}}} \times \left(\frac{1}{\beta} - \tau_{\rm rise} - \gamma\right) \leqslant \tau_{\rm rise}. \tag{B2}$$

We also ensure that the derivative of the second piecewise portion is more negative than that of the first portion at $t - t_0 = \gamma$, to properly model a postplateau drop off. This criterion simplifies to

$$\gamma \leqslant \frac{1 - \beta\tau_{\rm fall}}{\beta}. \tag{B3}$$

This simultaneously ensures that the model flux is always positive. We enforce these two criteria for both bands independently, within every sampler except CERES.

To compare the efficiency and accuracy of each sampler's fits, we average both the mean runtime and median reduced chi-squared across 20 objects of each spectroscopic class. We use the median instead of mean reduced chi-squared value because some techniques treat each dimension independently, and sometimes a probable selection of parameters independently leads to a poor model fit jointly; using the median better handles these outlier $\chi^2_{\rm red}$ values. Additionally, if the median $\chi^2_{\rm red}$ is greater than 1.2, we consider it a "poor" fit, as we further justify in Section 3.1. A comparison of the median reduced chi-squared values (among $\chi^2_{\rm red} < 1.2$ fits) and mean fitting time, as well as the fraction of poor fits, averaged across each true spectroscopic class is shown in Figure 23. Additionally, we show how each sampler fits example light curves in Figure 24. We see that there is a clear trade-off between runtime and fitting accuracy, with nested sampling and NUTS taking the

longest yet yielding almost no poor fits. Nested sampling is about an order of magnitude faster than NUTS. Nested sampling, NUTS, and SVI have the lowest fraction of poor fits, which is expected from their Bayesian nature. Both SVI and MIGRAD run *at least* an order of magnitude faster than nested sampling. SVI is limited by a one-time compilation so will be faster per fit after more consecutive fits. MIGRAD can potentially be sped up by an order of magnitude after compiling the likelihood function with numba. MIGRAD successfully fits about half the light curves, and struggles with handling the joint constraints between our model parameters. CERES seems to fail when fitting most of our light curves (returning the initial guess of the prior means), which could be due to it drawing directly from fit parameter space rather than log space, like our other samplers do.

Ultimately, we decide that, because time is not a limiting factor for classifier training and adding new photometric labels to archival data, we use nested sampling for training and inference from archival data sets, such as the one analyzed in this paper. However, nested sampling proves too slow for consistent real-time fitting and classification, especially in anticipation of the LSST alert stream. Within the publicly available ANTARES filter, we use SVI as the default fitting method but switch to nested sampling on a per-light-curve basis if the SVI fit is poor. We leave all five samplers as available options in the Superphot+ codebase.

Appendix C
Bayesian Oversampling to Balance Classifier Training

To balance training across our five classes, we explore using only the median fit parameters in conjunction with (1) multiplying the loss contribution from each light curve by a class weight, or (2) using synthetic minority oversampling techniques (SMOTE; Chawla et al. 2002; similar to Hosseinzadeh et al. 2020; Villar et al. 2020) to generate a balanced training set. Alternatively, we try a Bayesian oversampling approach where we draw multiple parameter sets from the posterior distribution of each minority-class light curve, and treat those draws as independent classifier inputs during training. The number of fits used per light curve depends on the relative training set abundance of that light-curve classification, with the classifier using more parameter sets per light curve for less prevalent classes. This allows us to input an equal number of light curves from each class into our classifier, while also better accounting for the fit uncertainty within each light curve. The distributions of samples resulting from both SMOTE and our Bayesian oversampling technique are compared in Figure 25, with the latter yielding improved classification accuracy and more robust performance metrics across classes and K-folds. We end up using 22,730 oversampled feature sets per class across our training and validation sets, corresponding to five sets of fit parameters per SN Ia light curve and `round` $(22{,}730/n)$ sets per light curve for each less prevalent class, with n being the events per class in our training (or validation) set. We find that drawing more samples per SN class did not improve classifier performance, while drawing fewer samples hurts SN Ia classification. These fits are drawn from the light-curve posterior space with replacement, so inputting the same set of model parameters into our classifier multiple times is possible among the less prevalent classes. Because of this, we oversample *after* dividing the SNe between the training, validation, and test sets, so there is no possibility of repeat fits or fits from the same light curve appearing in both a training and test set.

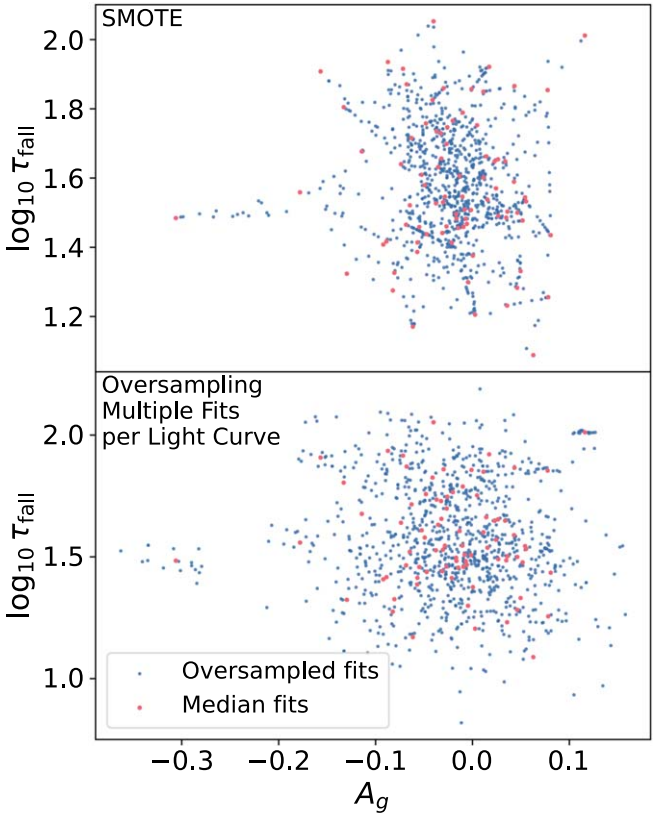


Figure 25. Comparison of oversampling techniques to balance our imbalanced spectroscopic data set. Here, we show only SLSNe-I, which is our smallest class in the sample, for visual clarity. SMOTE (top) draws samples from hyperplanes connecting pairs of observed SNe. This leads to a low diversity of samples when connecting outliers. In contrast, sampling directly from the set of equally weighted fits (bottom) for each light curve creates smoother oversampling.

Appendix D
Classifier Architecture Selection

Here, we explore two architecture options for SN classification. First, we train a multilayer perceptron (MLP), which is a simple neural network with fully connected layers of "neurons"; each neuron applies a nonlinear activation function to a linear combination of the input values. For simplicity, we only use dense layers in the MLP, with a constant number of neurons per hidden layer. Additionally, we include a 50% dropout rate for each hidden layer, and a 20% dropout rate on the input layer, as suggested by Hinton et al. (2012). Here, dropout refers to randomly masking out neuron values during training, which prevents a small subset of nodes from being the main determinant for discerning specific classes. As a result, nodes more evenly contribute to discerning important features within the latent space. The dropout applied to the input layer prevents the network from relying too heavily on specific fit parameters; instead, it more evenly weighs information from multiple regions of the light curve. This is important for accurately classifying partial light curves. To optimize the network architecture (number of hidden layers and neurons per layer), we perform a grid search over one to five hidden layers and 4–256 neurons per layer in intervals of 2^n, and calculate the validation set's class-averaged F_1-score for each combination.

We find that three hidden layers with 128 neurons each maximize the combined validation F_1-score across all K-folds. Across 10 folds, the MLP's F_1-score spans 0.59 ± 0.03 across an 80% confidence interval. For context, this is slightly lower than Superphot's reported F_1-score, which uses redshift information.

Next, we train a GBM using LightGBM (Ke et al. 2017). We attempt to optimize LightGBM's hyperparameters using a grid search, and find negligible performance improvements across different hyperparameter configurations. The results in this paper are obtained using the DART boosting strategy, `goss` sampling strategy, a max tree depth of 5, max number of leaves of 20, a regularization of 5, and 250 estimators. The LightGBM trains an order of magnitude faster than the MLP, and yields an F_1-score 80% confidence interval of 0.61 ± 0.02. While the median F_1-score is higher than that of the MLP, the two confidence intervals overlap (i.e., the worst-performing K-folds for the GBM report a lower F_1-score than the best-performing K-folds for the MLP). Because the K-folds were selected independently for the MLP and GBM, we thus argue that the lower MLP performance can potentially be attributed to the way the training set was randomly divided among folds for each architecture. Even so, we finalize the LightGBM model as the default classifier, as it is 2 orders of magnitude faster to train compared to the MLP.

Appendix E
Detailed Partial Light-curve Performance

Here, we delve deeper into Superphot+'s classification of partial light curves when trained on all features versus only early-phase features. Figure 26 shows how the completeness, purity, and F_1-score evolve for each classifier for each spectroscopic class as the light curves are truncated at different phases. The early-phase classifier especially excels for light curves truncated near peak, with much higher phase $= 0$ values for SN IIn purity (0.50 ± 0.25 versus 0.03 ± 0.02), SN Ia completeness (0.79 ± 0.09 versus 0.45 ± 0.13), and SN II completeness (0.31 ± 0.11 versus 0.15 ± 0.08) compared to the full-phase classifier. In contrast, the full-phase classifier yields significantly higher SN II completeness (0.41 ± 0.13 versus 0.72 ± 0.05) at late phases.

SLSN-I classification for both variants is stable from very early phases, as is SN IIn classification without postpeak features. This is expected, as both classes can be set apart by their slower rise evolution, which is constrained weeks before peak brightness. Both variants assign SN Ia and SN Ib/c labels with better completeness and purity near peak, where the peak color is fit more precisely; from Figure 3, we infer SNe Ib/c are most distinguishable by redder colors. The classification accuracy for SNe II (and SNe IIn with postpeak features) only increases weeks after peak, as classification relies on constraint of their characteristic plateaus.

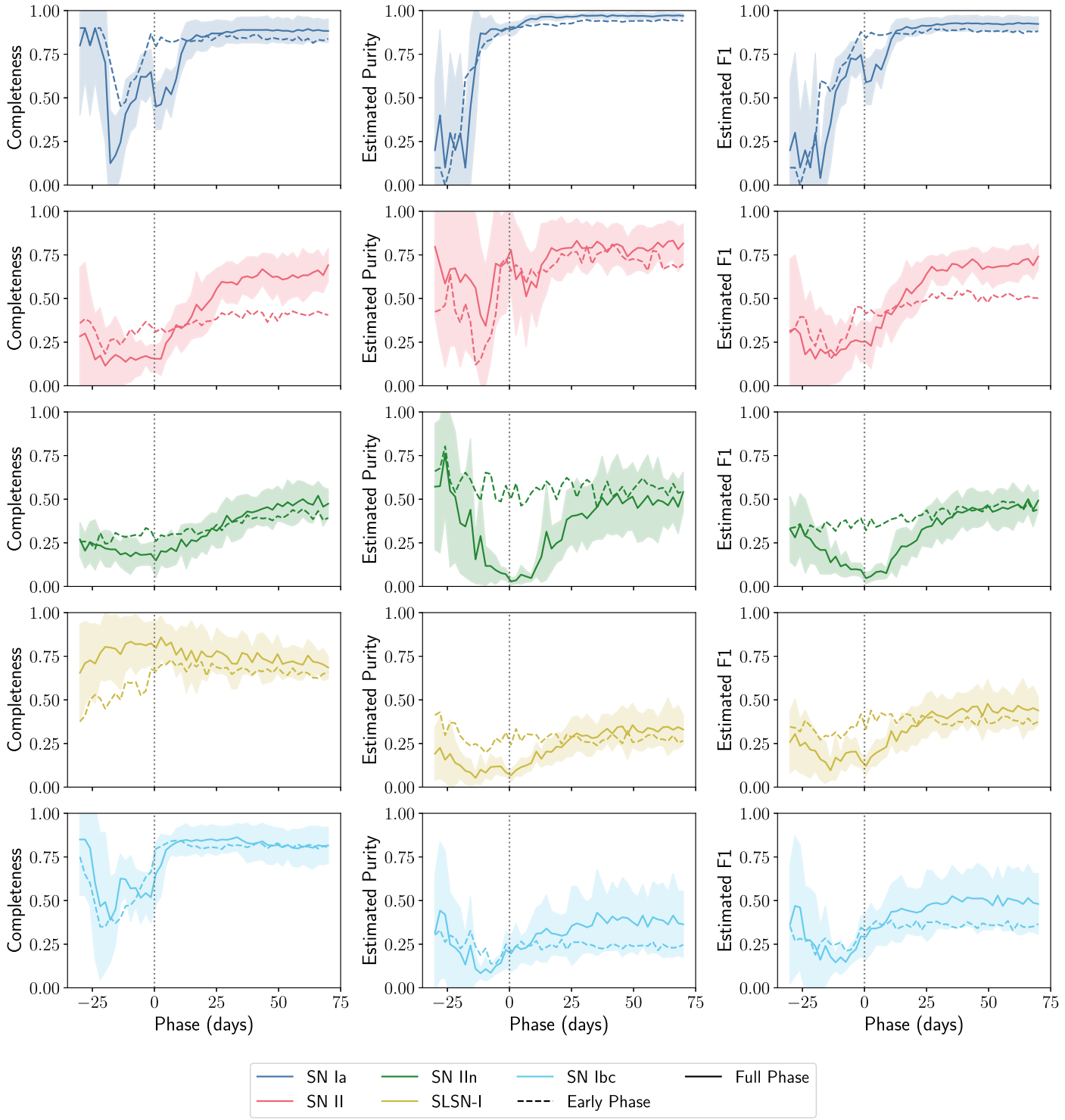


Figure 26. Classification completeness, purity, and F_1-score as a function of light-curve phase, for each spectroscopic class. The dashed line corresponds to the early-phase classifier, whereas the full-phase classifier is represented by the solid line and associated 1σ K-fold uncertainty margins. While SNe Ib/c classification accuracy improves significantly a few days before peak magnitude, light curves are only consistently labeled as SNe II or SNe IIn $\simeq$15 days after peak, when the postpeak behavior can be adequately measured.

ORCID iDs

Kaylee M. de Soto https://orcid.org/0000-0002-9886-2834
V. Ashley Villar https://orcid.org/0000-0002-5814-4061
Edo Berger https://orcid.org/0000-0002-9392-9681
Sebastian Gomez https://orcid.org/0000-0001-6395-6702
Griffin Hosseinzadeh https://orcid.org/0000-0002-0832-2974
Doug Branton https://orcid.org/0009-0009-7822-7110
Sandro Campos https://orcid.org/0009-0007-9870-9032
Melissa DeLucchi https://orcid.org/0000-0002-1074-2900
Jeremy Kubica https://orcid.org/0009-0009-2281-7031
Olivia Lynn https://orcid.org/0000-0001-5028-146X
Konstantin Malanchev https://orcid.org/0000-0001-7179-7406
Alex I. Malz https://orcid.org/0000-0002-8676-1622

References

Agarwal, S., Mierle, K. & The Ceres Solver Team, 2023 Ceres Solver, v2.2, https://github.com/ceres-solver/ceres-solver
Aleo, P. D., Malanchev, K., Sharief, S., et al. 2023, ApJS, 266, 9
ALeRCE, 2022 ALeRCE client, https://alerce.readthedocs.io/en/latest/
Anderson, J. P., González-Gaitán, S., Hamuy, M., et al. 2014, ApJ, 786, 67
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Brooks, S., Gelman, A., Jones, G. L., & Meng, X. 2011, Handbook of Markov Chain Monte Carlo (1st ed.; New York: Chapman and Hall)
Barbary, K. 2017, Extinction v0.3.0, Zenodo, doi:10.5281/zenodo.804967
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Bingham, E., Chen, J. P., Jankowiak, M., et al. 2019, Journal of Machine Learning Research, 20, 1, http://jmlr.org/papers/v20/18-403.html
Blondin, S., Matheson, T., Kirshner, R. P., et al. 2012, AJ, 143, 126
Boone, K. 2019, AJ, 158, 257
Boone, K. 2021, AJ, 162, 275
Boone, K., & Malanchev, K. 2022, kboone/parsnip: v1.3.1, Zenodo, doi:10.5281/zenodo.6980374
Bradbury, J., Frostig, R., Hawkins, P., et al., 2018 JAX: Composable Transformations of Python+NumPy Programs, v0.3.13, http://github.com/google/jax
Bradley, A. P. 1997, PatRe, 30, 1145
Branch, D., Fisher, A., & Nugent, P. 1993, AJ, 106, 2383
Carrasco-Davis, R., Reyes, E., Valenzuela, C., et al. 2021, AJ, 162, 231
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321
Claeys, J. S. W., de Mink, S. E., Pols, O. R., Eldridge, J. J., & Baes, M. 2011, A&A, 528, A131
Conti, P. S., & Niemelä, V. S. 1976, ApJL, 209, L37
Davis, J., & Goadrich, M. 2006, in Proc. 23rd Int. Conf. Machine Learning (New York: ACM), 233
De Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, Msngr, 175, 3
de Soto, K., Villar, A., Berger, E., et al. 2024, Superphot+: Realtime Fitting and Classification of Supernova Light Curves v0.0.7, Zenodo, doi:10.5281/zenodo.12519870
Dembinski, H., Ongmongkolkul, P., Deil, C., et al. 2023, scikit-hep/iminuit v2.24.0, Zenodo, doi:10.5281/zenodo.8249703
Dillon, J. V., Langmore, I., Tran, D., et al. 2017, arXiv:1711.10604
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, OJAp, 2, 10
Filippenko, A. V. 1997, ARA&A, 35, 309
Filippenko, A. V. 2005, in ASP Conf. Ser. 332, The Fate of the Most Massive Stars, ed. R. Humphreys & K. Stanek (San Francisco, CA: ASP), 34
Fitzpatrick, E. L., & Massa, D. 2007, ApJ, 663, 320
Fletcher, R. 1970, CompJ, 13, 317
Foley, R. J., Challis, P. J., Chornock, R., et al. 2013, ApJ, 767, 57
Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al. 2021, AJ, 161, 242
Fremling, C., Miller, A. A., Sharma, Y., et al. 2020, ApJ, 895, 32
Frostig, R., Johnson, M., & Leary, C. 2018, in Conf. on Machine Learning and Systems, https://mlsys.org/Conferences/doc/2018/146.pdf
Gagliano, A., Contardo, G., Foreman-Mackey, D., Malz, A. I., & Aleo, P. D. 2023, ApJ, 954, 6
Gagliano, A., Narayan, G., Engel, A., & Kind, M. C. 2021, ApJ, 908, 170
Gal-Yam, A. 2012, Sci, 337, 927
Gal-Yam, A. 2019, ARA&A, 57, 305
Gal-Yam, A. 2021, AAS Meeting Abstracts, 53, 423.05
Gomez, S., Berger, E., Blanchard, P. K., et al. 2020a, ApJ, 904, 74
Gomez, S., Berger, E., Blanchard, P. K., et al. 2020b, FLEET Finding Luminous and Exotic Extragalactic Transients v1.0.0, Zenodo, doi:10.5281/zenodo.4013965
Gomez, S., Berger, E., Blanchard, P. K., et al. 2023a, ApJ, 949, 114
Gomez, S., Villar, V. A., Berger, E., et al. 2023b, ApJ, 949, 113
Graham, M. L., Connolly, A. J., Ivezić, Z., et al. 2018, AJ, 155, 1
Green, G. 2018, JOSS, 3, 695
Green, G., Edenhofer, G., Krughoff, S., et al. 2023, gregreen/dustmaps: v1.0.11, Zenodo, doi:10.5281/zenodo.8210973
Grisel, O., Mueller, A., Lars, et al. 2024, scikit-learn/scikit-learn: Scikit-learn v1.4.0, Zenodo, doi:10.5281/zenodo.10532824
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv:1207.0580
Hoffman, M., Blei, D. M., Wang, C., & Paisley, J. 2012, arXiv:1206.7051
Hoffman, M. D., & Gelman, A. 2011, arXiv:1111.4246
Hosseinzadeh, G., Berger, E., Metzger, B. D., et al. 2022, ApJ, 933, 14
Hosseinzadeh, G., & Dauphin, F. 2021, Superphot v1.2.0, Zenodo, doi:10.5281/zenodo.5520623
Hosseinzadeh, G., Dauphin, F., Villar, V. A., et al. 2020, ApJ, 905, 93
Hsu, B., Blanchard, P. K., Berger, E., & Gomez, S. 2024, ApJ, 961, 169
Hunter, J. D. 2007, CSE, 9, 90
Lemaitre, G., Nogueira, F., Aridas, C. K., Victor, D. & Imbalanced-learn team, 2024 Imbalanced-learn Documentation, https://imbalanced-learn.org/stable/
Ivezić, Z. & LSST Science Collaboration 2018, The LSST System Science Requirements Document LPM-17, Rubin Observatory, https://docushare.lsst.org/docushare/dsweb/Get/LPM-17
James, F., & Roos, M. 1975, CoPhC, 10, 343
Jones, D. O., Scolnic, D. M., Riess, A. G., et al. 2017, ApJ, 843, 6
Kangas, T., Yan, L., Schulze, S., et al. 2022, MNRAS, 516, 1193
Kasen, D. 2006, ApJ, 649, 939
Kasen, D., Woosley, S. E., & Heger, A. 2011, ApJ, 734, 102
Ke, G., Meng, Q., Finley, T., et al. 2017, in Advances in Neural Information Processing Systems 30, ed. I. Guyon et al. (San Francisco, CA: Curran Associates, Inc.) https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
Kessler, R., Narayan, G., Avelino, A., et al. 2019, PASP, 131, 094501
Kisley, M., Qin, Y.-J., Zabludoff, A., Barnard, K., & Ko, C.-L. 2023, ApJ, 942, 29
Koposov, S., Speagle, J., Barbary, K., et al. 2023, joshspeagle/dynesty: v2.1.3, Zenodo, doi:10.5281/zenodo.8408702
Kornilov, M. V., Semenikhin, T. A., & Pruzhinskaya, M. V. 2023, MNRAS, 526, 1822
Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, Journal of Machine Learning Research, 18, http://jmlr.org/papers/v18/16-365.html
Leoni, M., Ishida, E. E. O., Peloton, J., & Möller, A. 2022, A&A, 663, A13
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Malanchev, K., 2021 Light-Curve Feature Extraction Library for python, light-curve-python, https://github.com/light-curve/light-curve-python
Malanchev, K. L., Pruzhinskaya, M. V., Korolev, V. S., et al. 2021, MNRAS, 502, 5147
Masci, F. J., Laher, R. R., Rusholme, B., et al. 2018, PASP, 131, 018003
Masci, F. J., Laher, R. R., Rusholme, B., et al. 2020, The ZTF Science Data System (ZSDS) Explanatory Supplement, Pipelines, Definitions, Data Products, Access & Usage 5.0, Zwicky Transient Facility, https://web.ipac.caltech.edu/staff/fmasci/ztf/ztf_pipelines_deliverables.pdf
Matheson, T., Stubens, C., Wolf, N., et al. 2021, AJ, 161, 107
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman, 56
Metzger, B. D., Vurm, I., Hascoët, R., & Beloborodov, A. M. 2013, MNRAS, 437, 703
Microsoft Corporation, 2023 LightGBM v4.1.0, https://lightgbm.readthedocs.io/en/v4.1.0/
Mukherjee, P., Parkinson, D., & Liddle, A. R. 2006, ApJL, 638, L51
Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002
Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, ApJS, 236, 9
Nelder, J. A., & Mead, R. 1965, CompJ, 7, 308
Nicholl, M., Guillochon, J., & Berger, E. 2017, ApJ, 850, 55
NSF NOIRLab ANTARES Team, 2024 ANTARES, https://nsf-noirlab.gitlab.io/csdc/antares/antares/
NumPy Developers, 2024 NumPy 1.26.3 release notes, https://numpy.org/devdocs/release/1.26.3-notes.html
NumPyro 2019, NumPyro 0.12.1 documentation, Uber Technologies, Inc., https://num.pyro.ai/en/0.12.1/
Nyholm, A., Sollerman, J., Tartaglia, L., et al. 2020, A&A, 637, A73
Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach et al. (Red Hook, NY: Curran Associates, Inc.), 8024, https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Perley, D. A., Fremling, C., Sollerman, J., et al. 2020, ApJ, 904, 35
Pessi, P. J., Anderson, J. P., Folatelli, G., et al. 2023, MNRAS, 523, 5315
Phan, D., Pradhan, N., & Jankowiak, M. 2019, arXiv:1912.11554
Ridley, E. J., Nicholl, M., Ward, C. A., et al. 2024, MNRAS, 531, 1905

Robitaille, T., Tollerud, E., Aldcroft, T., et al. 2023, astropy/astropy: v5.3.1, Zenodo, doi:10.5281/zenodo.8136839
Rose, B. M., Baltay, C., Hounsell, R., et al. 2021, arXiv:2111.03081
Rubin, A., Gal-Yam, A., Cia, A. D., et al. 2016, ApJ, 820, 33
Russeil, E., Malanchev, K. L., Aleo, P. D., et al. 2024, A&A, 683, A251
Saha, A., Matheson, T., Snodgrass, R., et al. 2014, Proc. SPIE, 9149, 914908
Saito, T., & Rehmsmeier, M. 2015, PLoSO, 10, 1
Salvatier, J., Wieckiâ, T. V., & Fonnesbeck, C., 2016 PyMC3: Python Probabilistic Programming Framework, Astrophysics Source Code Library, ascl:1610.016
Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141
Sanders, N. E., Soderberg, A. M., Gezari, S., et al. 2015, ApJ, 799, 208
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Sheng, X., Nicholl, M., Smith, K. W., et al. 2024, MNRAS, 531, 2474
Skilling, J. 2006, BayAn, 1, 833
Smith, N. 2014, ARA&A, 52, 487
Speagle, J. S. 2020, MNRAS, 493, 3132
Stein, R., Reusch, S., Franckowiak, A., et al. 2023, MNRAS, 521, 5046
The Matplotlib Development Team 2023, Matplotlib: Visualization with Python v3.8.2, Zenodo, doi:10.5281/zenodo.10150955
The pandas development team 2020, pandas-dev/pandas: Pandas v2.0.3, Zenodo, doi:10.5281/zenodo.8092754pandas-dev/pandas: v2.0.3
Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018, PASP, 130, 064505
Tyson, J. A. 2002, Proc. SPIE, 4836, 10
Villar, V. A., Berger, E., Miller, G., et al. 2019, ApJ, 884, 83
Villar, V. A., Hosseinzadeh, G., Berger, E., et al. 2020, ApJ, 905, 94
villrv 2020, villrv/SuperRAENN: v1.0 Release, Zenodo, doi:10.5281/zenodo.3968715
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261