Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **543,** 247–272 (2025) https://doi.org/10.1093/mnras/staf1419
Advance Access publication 2025 August 28

Testing and combining transient spectral classification tools on 4MOST-like blended spectra

A. Milligan [1]★ I. Hook,[1] C. Frohmaier [2] M. Smith,[1] G. Dimitriadis [1] Y.-L. Kim [1,3] K. Maguire [4] A. Möller [5] M. Nicholl [6] S. J. Smartt [6,7] J. Storm [8] M. Sullivan [2] E. Tempel [9] P. Wiseman [2] L. P. Cassarà [10] R. Demarco,[11] A. Fritz[12] and J. Jiang[13,14]

[1]*Department of Physics, Lancaster University, Lancs LA1 4YB, UK*
[2]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[3]*Department of Astronomy & Center for Galaxy Evolution Research, Yonsei University, Seoul 03722, Republic of Korea*
[4]*School of Physics, Trinity College Dublin, The University of Dublin, Dublin 2, Ireland*
[5]*Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Victoria 3122, Australia*
[6]*Astrophysics Research Centre, School of Mathematics and Physics, Queens University Belfast, Belfast BT7 1NN, UK*
[7]*Astrophysics Sub-department, Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK*
[8]*Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany*
[9]*Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere 61602, Estonia*
[10]*INAF – IASF Milano, via Alfonso Corti 12, I-20133 Milano, Italy*
[11]*Institute of Astrophysics, Facultad de Ciencias Exactas, Universidad Andrés Bello, Sede Concepción, Talcahuano, Chile*
[12]*Kuffner Observatory, Johann-Staud-Strasse 10, A-1160 Vienna, Austria*
[13]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[14]*Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*

Accepted 2025 August 20. Received 2025 August 15; in original form 2024 December 23

ABSTRACT

With the 4-metre Multi-Object Spectroscopic Telescope (4MOST) expected to provide an influx of transient spectra when it begins observations in early 2026 we consider the potential for real-time classification of these spectra. We investigate three extant spectroscopic transient classifiers: the Deep Automated Supernova and Host classifier (DASH), Next Generation SuperFit (NGSF), and SuperNova IDentification (SNID), with a focus on comparing the completeness and purity of the transient samples they produce. We manually simulate fibre losses critical for accurately determining host contamination and use the 4MOST Exposure Time Calculator to produce realistic, 4MOST-like, host-galaxy contaminated spectra. We investigate the three classifiers individually and in all possible combinations. We find that a combination of DASH and NGSF can produce a supernova (SN) Ia sample with a purity of 99.9 per cent, while successfully classifying 70 per cent of SNe Ia. However, it struggles to classify non-SN Ia transients. We investigate photometric cuts to transient magnitude and the transient's fraction of total fibre flux, finding that both can be used to improve non-SN Ia transient classification completeness by 8–44 per cent with SNe Ibc benefitting the most and superluminous (SL) SNe the least. Finally, we present an example classification plan for live classification and the predicted purities and completeness across five transient classes: Ia, Ibc, II, SL, and non-SN transients. We find that it is possible to classify 75 per cent of input spectra with >70 per cent purity in all classes except non-SN transients. Precise values can be varied using different classifiers and photometric cuts to suit the needs of a given study.

Key words: instrumentation: spectrographs – techniques: spectroscopic – software: machine learning – software: simulations – transients: supernovae.

1 INTRODUCTION

Since the discovery of the accelerating expansion of the universe a quarter of a century ago (Riess et al. 1998; Perlmutter et al. 1999), significant efforts have been made to investigate the enigmatic properties of dark energy. Many probes into the nature of dark energy exist, including weak lensing and cosmic microwave background measurements (Wittman et al. 2000; Planck Collaboration I 2014). However, one of the most successful at providing strong constraints on cosmological models in the late-time universe is type Ia supernova (SN) cosmology. Understood to be the detonation of white dwarfs around the Chandrasekhar mass limit, SNe Ia detonate at predictable luminosities and as such act as standardizable candles that let us measure the distance to objects over large swathes of cosmic time.

The original discovery of accelerating expansion was performed with a sample of only 42 high-redshift SNe Ia (Riess et al. 1998; Perlmutter et al. 1999). Since then, we have seen a two order of magnitude increase in the number of spectroscopically confirmed SNe Ia. For example, recently the Zwicky Transient Facility have

★ E-mail: a.milligan@lancaster.ac.uk

produced their second data release sample (ZTF DR2) (Rigault et al. 2025) which contains 2677 SN Ia with sufficiently high-quality light curves for use in cosmological fitting. Similarly, the recent Dark Energy Survey (DES) cosmology results (DES Collaboration 2024) use a sample of 1635 SN Ia, derived from their full 5-yr data release.

The earliest samples of transients were separated into two classes: SNe I and SNe II, based on the presence or absence of Hydrogen features in their spectra (Popper 1937; Minkowski 1979). In the years since, these classes have been further subdivided and many new subclasses (Filippenko 1997) and exotic variants have been discovered and suggested, alongside non-SN transients like tidal disruption events (TDEs) and fast blue optical transients (FBOTs, Hills 1975; Drout et al. 2014).

Most optical transients are discovered in photometric surveys. As the number of transients has increased, it has become unfeasible to allocate time for spectroscopic follow-up on each transient individually. Recent photometric classifiers can perform high accuracy classification on transients beyond just classifying them as SN Ia or non-SN Ia (Charnock & Moss 2017; Muthukrishna et al. 2019a; Boone 2019, 2021; Möller & de Boissière 2020; Pimentel, Estévez & Förster 2023; Sheng et al. 2024; Cabrera-Vives et al. 2024; Shah et al. 2025). Additionally, it has been shown that they are capable of classifying transients based on incomplete light curves (Möller & de Boissière 2020; Qu & Sako 2022; Gagliano et al. 2023; Gomez et al. 2023; de Soto et al. 2024). Recent photometric analyses have indicated that SN Ia samples obtained with photometric classifications produce contamination levels that either still allow for robust estimations of cosmological parameters or are even negligible compared to other sources of uncertainty, such as SN Ia astrophysics and how we model the correlation between SN Ia intrinsic properties and host-galaxy properties and how these intrinsic properties evolve with redshift (Jones et al. 2018, 2019; Vincenzi et al. 2024).

While photometric classification is possible, it has several distinct disadvantages. The definitions of SN subclasses are based primarily by spectral features, so spectroscopic classification removes ambiguity, although there are also photometrically defined classifications. For example, SNe IIn are defined spectroscopically by narrow emission lines (Schlegel 1990), while SNe IIP are defined photometrically by a long 'plateau' phase of constant brightness in their light curve (Filippenko 1997). Further, when attempting to constrain cosmology, photometrically classified SN Ia samples often require the addition of spectroscopic information, such as spectroscopically determined host-galaxy redshifts. This is the case in Vincenzi et al. (2024), where 1635 photometrically classified SNe Ia are used for cosmology, the largest single-survey SN Ia sample. Additionally, Vincenzi et al. (2024) use a small sample of spectroscopically classified SNe Ia to constrain the cosmological fitting (see also DES Collaboration 2024). Beyond this, to match the high purities of spectroscopically classified transient samples, photometric classification is usually performed in a binary scheme (SN Ia versus non-SN Ia) or with very broad transient classes (Fraga et al. 2024).

We will, therefore, test the performance of spectroscopic classifiers. Visual classification is made difficult by the overlap of various transient subclasses in parameter space and ambiguity in subclass definitions. This, alongside the increasing number of transients being observed spectroscopically, means that it is increasingly required to automate the process of spectroscopic classification. We seek to investigate the potential to do this with regards to the upcoming 4MOST (4-metre Multi-Object Spectroscopic Telescope) instrument.

The 4MOST (de Jong et al. 2019) is a high-multiplex, fibre-fed spectrographic survey facility in the final stages of assembly before commissioning. It is expected that it will begin taking data in early 2026. There are many varied surveys within the 4MOST consortium, but the survey concerned with transients is the Time Domain Extragalactic Survey (TiDES, Swann et al. 2019; Frohmaier et al. 2025).

With the upcoming Legacy Survey of Space and Time (LSST) being performed from the Vera C. Rubin Observatory, there will be unprecedented numbers of transients discovered photometrically (Ivezić et al. 2019). It is expected that any given pointing of 4MOST will contain a number of live photometric transients and the host galaxies of faded transients, which can then be followed-up with TiDES's allotted fibres. Over a period of 5 yr, TiDES expects to observe 30 000 live transients and perform follow-up on some 200 000 host galaxies (these numbers are dependent on the survey schedules of LSST and 4MOST, both of which are still under development). This approach has already seen success in the Australian Dark Energy Survey (OzDES) performed using the AAOmega spectrograph on the Anglo-Australian Telescope (Saunders et al. 2004; Lidman et al. 2020).

Two of TiDES science goals are to provide live classification of transients accessible to the general scientific community and the classification of a large, pure, cosmological SN Ia sample. As we approach the start of the 4MOST survey in early 2026, uncertainty remains as to how the TiDES transient spectra will be classified and which existing spectroscopic classifiers, if any, are best suited to these two TiDES science goals. Our hope is to provide clarity via the simulation of transient spectra that are as close to what will be observed as possible, including the fact that transient flux observed by a 4MOST fibre will be blended with the flux of its host galaxy. These realistic, blended, simulated 4MOST spectra will allow us to compare the output of various spectroscopic classifiers to known true classifications (see also Kim et al. 2024, which makes use of real spectra in its analysis). Furthermore, we can assess the dependence of classification performance on parameters such as the brightness of the SN and the fraction of host light contaminating the spectrum, and ultimately use this information to outline a plan for the classification of large numbers of TiDES spectra.

There are two main types of automated, spectroscopic classifiers. First, there are template matching programs (for example Duan et al. 2009; Blondin & Tonry 2011; Goldwasser et al. 2022). These, in essence, compare an input spectrum to a bank of transients of known classification. However, there is significant variation in methodology. For example, Howell et al. (2005) bin the input spectrum to match the templates and then calculate a χ^2 value, accounting for contaminant host flux. Blondin & Tonry (2011) instead cross-correlate input and template in redshift, and quantifies the best-fitting template by the height of the cross-correlation peak.

More recent years have seen the rise of the second type: machine-learning methods (for example Harutyunyan et al. 2008; Muthukrishna, Parkinson & Tucker 2019b; Vogl et al. 2020; Fremling et al. 2021; Sharma et al. 2025). In this case, a classifier is provided a training set of templates of known classification and redshift. The classifier 'learns' the features present in various transient classifications and assigns them weights. The presence or not of these learned features is then used to determine a pseudo-probability of an input spectrum belonging to a given classification, which is then used to rank output classifications.

In this paper, we investigate two template-matching classifiers and one machine-learning classifier. More information on the spectroscopic transient classifiers we investigate can be found in Sections 4.1.1–4.1.3. These classifiers were chosen as they are publicly available, widely used and easily obtainable for current and upcoming surveys. Machine-learning algorithms are far faster to

perform classifications once the lengthy training process is complete, but all classifiers as they are used in this work are expected to scale to TiDES.

Hence, this paper is organized as follows. First, in Section 2, we describe the simulations from which we draw our transient and host properties. Also in this section we will discuss some transient templates used in simulating our blended spectra. In Section 3, we will discuss the construction of blended host–transient spectra and the subsequent simulation of 4MOST observations using an Exposure Time Calculator (ETC). Then, in Section 4, we investigate the capabilities of three individual spectroscopic transient classifiers. We go over their function and how they were tested. Their individual performances are presented in Sections 4.3 and 4.5. We investigate the combination of classifiers in Section 5. We first show the results from a simple combination of classifiers and then potential photometric cuts for improving classification in Section 5.1. Finally, in Section 5.2, we present a potential classification pipeline for live classification and SN Ia cosmology. Our conclusions are presented in Section 6.

2 DATA

2.1 Survey simulations

Our objective is to test spectroscopic transient classifiers such that we understand under what conditions they will succeed or fail in correctly determining the transient classes of 4MOST-like spectra. We must simulate a set of spectra that are a good approximation to the real ones observed by the instrument. The specific procedure for the creation of individual spectra is covered more in Section 3, but we first discuss how we obtain a set of realistic properties for transients and their hosts. These properties can then be used to generate each spectrum, which in turn can be used to test each of the pre-existing transient classifiers. The results of these classifications can then be compared to the input spectrum's 'true' properties as a means to quantify the success of a given classifier.

We make use of two pre-existing, sequential simulations to produce a realistic sample of blended host–transient spectra. The first is a simulation of a population of transients and hosts performed in the SUpernova ANAlysis package (SNANA, Kessler et al. 2009). SNANA uses known intrinsic properties of various transient classes in combination with the survey strategy of the LSST survey to generate an LSST-specific transient population (Frohmaier et al. 2025). This simulation produces a population of transient and host objects. From them, we obtain the intrinsic physical properties of host–transient systems. We obtain system redshift, host–transient separation, host r-band magnitude, and transient template information. Throughout this paper, magnitudes are calculated using the LSST r-band filter and are reported in the AB magnitude system (Oke & Gunn 1983). The process of creating simulated spectra is discussed in more detail in Section 3.

The second simulation is a simulation of the 4MOST survey operation of the full 5 yr of observations of the southern sky. Observation targets are taken from the simulated survey input catalogs and their exposure times are computed using the 4MOST ETC. The simulation is carried out with the 4MOST facility simulator (4FS) and makes use of the simulation code SELFIE. More detail about the SELFIE algorithm can be found in Tempel et al. (2020a, b).

This simulation provides further observational properties for each transient. Most importantly, from it we receive a list of all of the transients that were observed. Generally, any transient that is both located within 4MOST's field of view during a visit, and is estimated to require less exposure time than is available during the full visit to meet the TiDES spectral success criterion (average SNR $>$ 3 in 15 Å bins in the wavelength range of 4500–8000 Å, where SNR is the signal-to-noise ratio) will be observed. However, some are not observed due to the limited number of fibres and the demands of other subsurveys.

As the simulations have become more sophisticated, different versions of the input catalogue have been created. Each has had many different simulations of survey operations performed on it. We find that while the individual objects observed may change dramatically between simulations, the bulk properties of the observed transients are consistent. The specific simulation used has little effect on our final results.

The 4MOST observing schedule is currently expected to visit each sky position a small number of times during the 5-yr survey. The survey footprint of 4MOST essentially covers the whole extragalactic sky in the Southern hemisphere. Each visit to a given position will consist of several exposures (most often 2 or 3) of approximately 20 min. The majority of transients ($>$93 per cent) are observed a single time over the course of the survey (Frohmaier et al. 2025).

The r-band magnitude, redshift, and SN flux fraction distributions from the SNANA population simulation of the transients and their hosts from the SNANA population simulation are shown in Fig. 1. The total number of objects in the sample is on the order of 10^5. We see that the sample is heavily biased to $z < 0.6$ and in fact the more distant objects are all superluminous SNe (SLSNe). We also see that, before any correction for fibre sizes, when observing extended objects (see Section 3.3), there is a tendency for host galaxies to have brighter magnitudes than transients.

2.2 Simulated spectra

In addition to realistic physical and observational properties for use in creating simulated 4MOST-like spectra, we require a set of spectral templates of both transients and hosts. The transient templates are drawn from those used in the SNANA population simulations. The included SN classes are Ia, Ib, Ic, II, IIn, IIb, and SLSNe. Most SNe Ia input templates are of the Ia-norm subclass, generated using the Spectral Adaptive Light-curve Template (SALT2) model (Guy et al. 2007), although a small fraction are SNe Iax and SNe Ia 91bg-like (Kessler et al. 2019). Additionally, there are TDEs, and calcium-rich transient (CaRT) objects. These templates are spectral energy distributions (SEDs) intended to simulate realistic photometry. As a result, some of the spectra, especially SLSNe and non-SN transient, are highly smoothed and lacking in spectroscopic features. The full list of template sources is provided in Table 1. Examples of SEDs used in simulated blended spectra are shown in Appendix C.

The galaxy templates from Kinney et al. (1996) are assigned as hosts. The subclasses of galaxy available are elliptical, S0, Sa, Sb, and Sc and a set of starburst templates with a variety of $E(B-V)$ values (see Kinney et al. 1996, Mannucci et al. 2001, for additional information). We scale our galaxy templates using the r-band host magnitudes from the simulation.

For each transient we assign a host-galaxy morphology to match the probability distribution listed in Hakobyan et al. (2012) in their table 5. For Sd and Irregular galaxies for which we have no templates, we assign a random choice between Sb and Sc host spectra (the two most common host morphologies). In cases where Hakobyan et al. (2012) list the host as Morphology A/Morphology B, we choose randomly between A and B. We always assign SLSNe inputs an Sc-type host spectrum since research suggests that SLSNe are found in faint, blue, star-forming galaxies, often with extreme emission lines (Leloudas et al. 2015; Neill et al. 2011). TDEs and CaRTs occupy

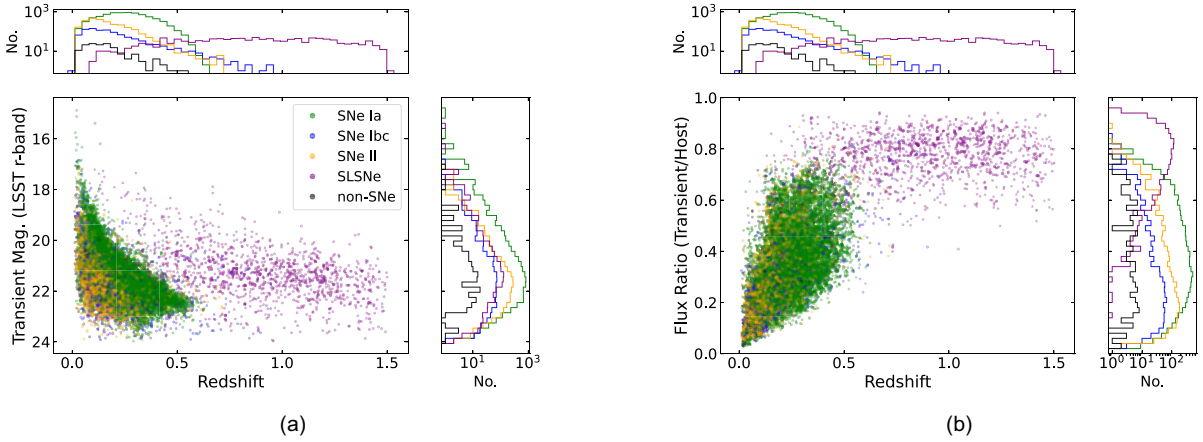


Figure 1. (a) Host galaxy redshift and corresponding transient magnitudes for observed objects in the SELFIE survey simulation. The values are obtained directly from the SNANA population simulation and can be considered the truth values for a given object. The *y*-axis on the attached histograms displays the total number of objects per bin with a logarithmic scale. (b) As in (a), but with the fraction of fibre flux from the transient on the *y*-axis.

Table 1. The relative percentages of each transient class present in our full sample of blended spectra alongside the sources for the spectral templates. Templates can be found in the SNANA public data as part of PLASTICC (Kessler et al. 2019) and ELASTICC (Narayan & ELAsTiCC Team 2023).

Percentage	Class	Source
60.1	SNe Ia	Guy et al. (2007), Hounsell et al. (2018)
0.9	91bg-like	Kessler et al. (2019)
1.1	SNe Iax	Kessler et al. (2019)
1.9	SNe Ib	Vincenzi et al. (2019)
1.4	SNe Ic	Vincenzi et al. (2019)
13.5	SNe II	Vincenzi et al. (2019)
6.5	SNe IIn	Vincenzi et al. (2019)
4.0	SNe IIb	Vincenzi et al. (2019)
9.4	SLSNe	Kessler et al. (2019)
0.7	TDE	Kessler et al. (2019)
0.4	CaRT	Kessler et al. (2019)

such a small percentage of our transients, that we assign them a host type at random. However, we note that there is evidence that TDEs (Wang et al. 2024) and CaRTs (Dong et al. 2022) do show trends in their host galaxy morphologies, but including these in our simulations would have negligible impact in our results.

In order to estimate uncertainties in our results, we split the full sample of transients into samples of 1000 transients. This subsampling is performed randomly, but without resampling (i.e. no transient appears in more than one subsample). For a given parameter, results are obtained by reporting the mean value across all subsamples. The uncertainty on our results are reported in the form of the standard error of the mean.

3 CREATING BLENDED SPECTRA

3.1 The 4MOST Exposure Time Calculator

The 4MOST ETC PYTHON code package[1] allows one to simulate an observation by the 4MOST instrument. For every simulated observation, we must assign a brightness within a specific filter or over a wavelength range. A variety of pre-existing instrument filters are provided.

The code produces a 'raw' or Level 0 (L0) output and a Level 1 (L1) output. Both are in the form of extracted 1D spectra (flux and wavelength for each pixel along the spectrum). The raw output features 4MOST's three spectrograph arms not yet combined and the object flux reported in Analog-to-Digital Units (ADUs). The L1 output is what we use. L1 spectra are generated by being passed through a simulation of the Quality Control 1 (QC1) pipeline and resemble the data products that will be produced by the real instrument. In L1 output, the ADUs of the raw output are converted to a flux observed at the telescope entrance using corrections for the wavelength dependence of the instrument's sensitivity.

The simulation process is shown in Fig. 2. There are still telluric absorption bands present in the L1 output which are added as part of the ETC model. There are five main features with wavelength ranges of 6250–6350, 6860–6940, 7150–7350, 7550–7700, and 8100–8400 Å. These extra features could be misinterpreted by classifiers as being generated by the transient and lead to misclassifications. We account for this by creating a transmission spectrum for each observation. We do this on the assumption that real data will have these features corrected for using 4MOST observations of featureless calibration stars.

We consider the host and transient separately before adding them linearly to form the final spectrum that is input into the ETC for a simulated observation. The magnitudes of both objects are known from the population simulation, but to account for seeing conditions and a finite fibre size on extended galaxies we must adjust these magnitudes. The processes for doing so for SNe and galaxies are shown in detail in Sections 3.2 and 3.3, respectively.

3.2 Transient fibre flux

We assume the transient can be approximated as a point source and that the 4MOST fibre will be placed centrally on the transient. We simulate the fraction of transient flux through a 4MOST fibre using a grid of pixels with a central pixel containing the full transient flux. A Gaussian convolution is then applied to the pixel grid. The standard deviation, σ, of the Gaussian convolution is determined from the full-width half-maximum (FWHM) of the seeing conditions using the expression $\mathrm{FWHM} = 2\sqrt{2\ln 2}\sigma$.

[1]We use V2.3.1 of the PYTHON-based ETC: see QMOSTETC link to documentation.

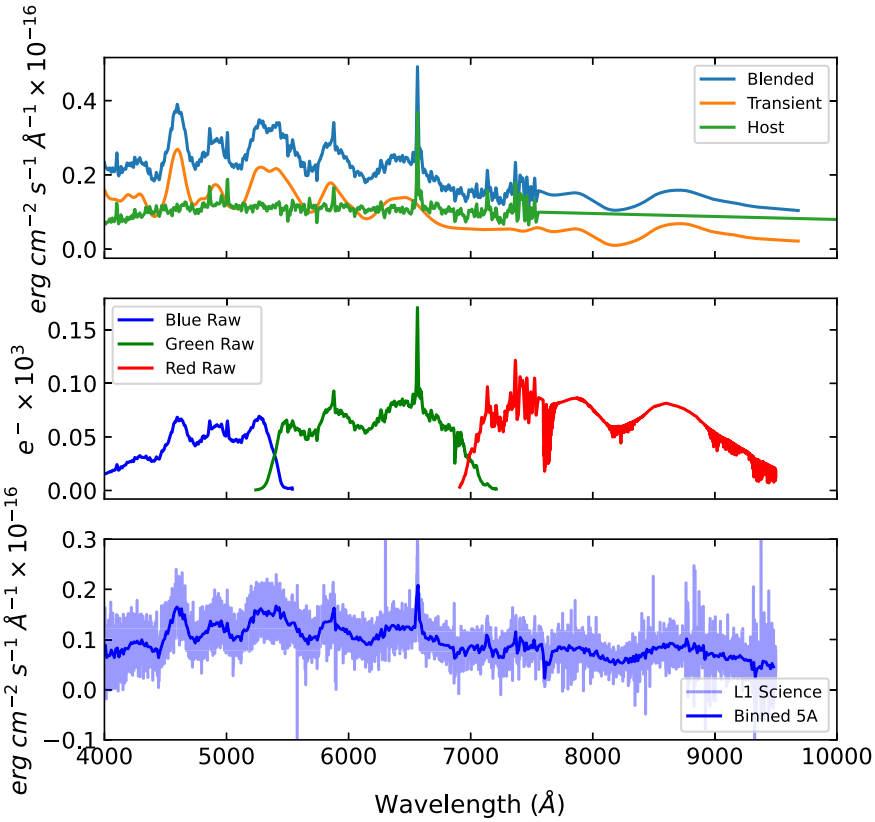


Figure 2. The stages of simulating an observation with the 4MOST ETC code. In this example, a 21st magnitude SN Ia and a 21st magnitude Sc-type host spectra are added linearly. Top panel: template SN, host, and combined spectra. All spectra are deredshifted. The flux is measured in units of erg cm^{-2} s^{-1} Å^{-1} × 10^{-16}. This is the input to the ETC. Middle panel: L0 output of the ETC, showing the extracted spectra from the three spectrograph arms. Flux is presented in units of e^- × 10^3. Lower panel: L1 output of the ETC in which the spectra from the three arms have been joined. The result is flux-calibrated and includes a realization of the noise. This (unbinned) L1 spectrum is what we perform classification on.

The SELFIE simulations do not record seeing conditions for each observation. For our purposes, the seeing conditions are taken to always have a value of 0.8 arcsec, this is similar to the average seeing conditions found at the Paranal Observatory where 4MOST will be located.[2]

Once the Gaussian convolution has been applied, a fibre with a 4MOST fibre diameter of 1.45 arcsec is imposed onto the pixel grid, centred on the SN location. The flux is then summed from the pixels with centres contained within the fibre radius. We find that using a finer pixel grid produces a more accurate value for fibre flux by reducing uncertainty around the fibre edge. This is particularly important in Section 3.3 where the scale of hosts being modelled varies and a balance must be found between accuracy and computation time.

We are assuming a constant value for the seeing, coupled with a constant fibre size, so we see a constant fraction of transient flux down each fibre. The effect is that each transient appears 0.27 mag fainter through the 4MOST fibre. This number does not require a simulation to be determined, as it determined from the integration of a 2D Gaussian out to some radius, but simulations are required for simulating extended hosts of varying size as discussed in Section 3.3.

At seeing < 0.8 arcsec, the fraction of flux down the fibre from both transient and host is increased. Tests show that the increase is larger on average for transients (as they are point sources), so we would expect improved classification in this case. The reverse is true for seeing > 0.8 arcsec and so we would expect worsened classification. Simulations indicated that increasing the seeing value to a uniform 1.2 arcsec had a small, negative effect on transient classification, but ultimately a realistic seeing distribution centred on 0.8 arcsec is expected to have minimal effect on the overall rates of successful transient classification.

3.3 Host fibre flux

The modelling of fibre flux from the transient's host galaxy, an extended object, is more complex. This method involves the dimensionless distance parameter ($d_{\rm DLR}$), first used in Sako et al. (2018), in service of assigning hosts to transients and based on similar methods developed in Sullivan et al. (2006). The $d_{\rm DLR}$ is equal to the ratio of the directional light radius (DLR) of a galaxy and its observed separation from the transient. The DLR is the half-light radius of the galaxy in the direction of the transient. Minimizing the $d_{\rm DLR}$ for galaxies in a crowded field indicates likely hosts for the transient.

The population simulation we draw SNANA-produced physical properties from reports both the $d_{\rm DLR}$ and the host–transient separation. Since we are only concerned with the host's flux in the direction of the transient for the purposes of measuring the flux through a 4MOST fibre, we can consider all galaxies in the simulation to have circular half-light radii equal in radius to their DLRs. It should be noted that the position of the transient is entirely based on the light

[2]From Paranal Observatory website, https://www.eso.org/gen-fac/pubs/astclim/paranal/seeing/?, accessed 2024 January 23.

profile of the galaxy, so that transients are more likely to be placed in brighter regions of their hosts (Vincenzi et al. 2021).

We note that significant work has been performed investigating links between transients and their locations within their host galaxies (see Hakobyan et al. 2016; Aramyan et al. 2016; Galbany et al. 2018, for example). However, since the population simulation preferentially places transients in brighter regions of their host, the resulting spectra may only be biased towards slightly higher levels of contamination from host flux. The effect on our results is negative, and is expected to be negligible.

We model the intensity of the galaxy to be a Sérsic (1963) profile and use a Sérsic index of 0.5 based on values reported in the simulations. While this may not be completely true to life, it represents the case with the most host flux in a blended spectrum and the hardest case to classify. Using a larger Sérsic index causes the average host flux in the fibre to decrease leading to less host contamination. The Sérsic profile is dependent on the value of the constant b_n which in turn is defined by the Sérsic index. A number of approximations for the value exist such as $b_n = 1.9992n - 0.3271$ for $0.5 <= n <= 10$ from Capaccioli (1989) and $b_n = 2n - \frac{1}{3} + 0.009876n$ from the appendices of Prugniel & Simien (1997). We will use the latter, although both produce very similar values for $n = 0.5$.

The intensity profile, in terms of the Sérsic index, n, and b_n, is often expressed as:

$$I(R) = I_e \exp\left\{-b_n\left[\left(\frac{R}{R_e}\right)^{\frac{1}{n}} - 1\right]\right\} \quad (1)$$

where R_e is the effective or half-light radius that encircles half of the total emission of the profile. The effective intensity, I_e, is the intensity at the effective radius.

To obtain the ratio of total galaxy flux to the flux transmitted through the fibre, we need to know the value of the total flux and the effective intensity. The total flux is obtained by integrating the intensity profile in equation (1) which leads to the equation:

$$F_T = 2.8941\pi I_e R_e^2 \quad (2)$$

This gives us the total flux in terms of the effective intensity and the effective radius which is just the DLR (for a more detailed derivation, see Graham & Driver 2005, and references therein). We can find the actual value of the total flux, and thus a value for the effective intensity, from the zero-point magnitude of the AB magnitude system and the total magnitude of the galaxy, m_G, using the equation:

$$F_T = f_0 \times 10^{(m_G/-2.5)} \quad (3)$$

Here, f_0 is the zero point flux of the AB magnitude system. The total host flux, F_T, that appears in our equations, only functions as a scaling factor. We know the true value of m_G from the population simulation. By taking the ratio of total flux to flux in the 4MOST fibre, the value of the total flux cancels out and so it need not be calculated specifically. Once an arbitrary total flux is chosen we can calculate the effective intensity, I_e, using equation (2). We can then use equations (1) and (2) to calculate the ratio between the total flux, the flux down the fibre and thus the host's magnitude as observed by 4MOST down its fibre.

We simulate a host's intensity profile by creating a pixel grid and use the Sérsic profile to determine the average intensity at each pixel. Since we only care about the host's light profile in the direction of the transient, we model each host as a circle with a half-light radius equal to the DLR.

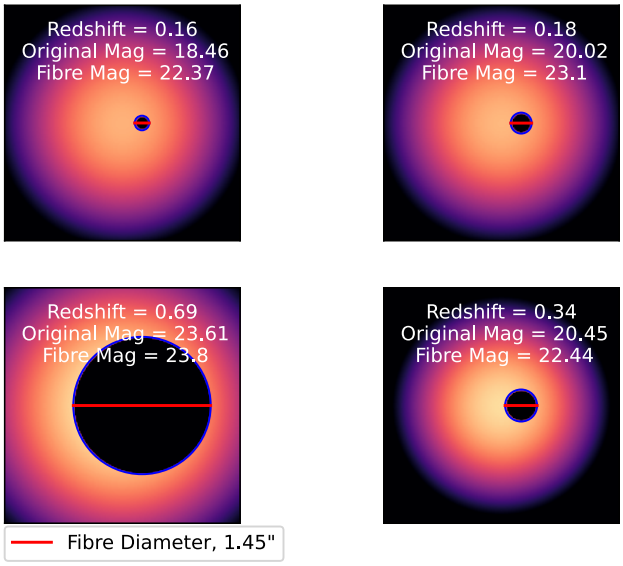


Figure 3. The variation in the host flux through 4MOST fibres. Each panel presents the Sérsic profile of an example host galaxy in our sample simulated on a pixel grid. Superimposed as a circle is the 4MOST fibre of diameter 1.45 arcsec, highlighted by the horizontal line, centred on the transient location. The pixels that contribute to the flux seen by the fibre have their flux set to zero in these images, so that the lost flux can be seen. Redshifts, and host magnitude before and after accounting for fibre losses are provided.

We then apply a Gaussian convolution to the pixel grid to account for atmospheric seeing. We use a 1200 × 1200 pixel grid with each pixel set to 1 per cent of the host–transient separation, a scale where the calculated flux fraction is invariant with small variations in pixel size. The method is identical to that described in Section 3.2. We centre the fibre on the transient location and calculate the fraction of flux in the fibre. Examples of this process are shown in Fig. 3. We see much more significant flux loss than for the SNe.

The 4MOST ETC cannot simultaneously account for both extended and point sources in a simulated observation. This is why we account for fibre losses and seeing effects ourselves, prior to passing the blended spectrum to the ETC. We provide the blended spectrum as being a flat illumination source with brightness measured in magnitudes per square arcsecond to prevent the ETC from re-applying any observational effects like seeing.

As stated in Section 3.2, the effect on the transient magnitude is fairly minimal. Most of the flux from the original point source still falls within the fibre that has a diameter of roughly 2σ relative to the Gaussian convolution. For hosts, their distance, size, and separation from their hosted transient result in significantly more variation in the fraction of the flux that is seen by the fibre (see Fig. 3). This is a critical effect to model. By correcting the host magnitudes for fibre effects, we see an average increase in the host magnitude of about 3.1 mag.

This leads to a reduction in host-galaxy flux contamination in the blended spectra. The distribution of transient fibre flux fractions shown in Fig. 4 demonstrates that we now have more than half of our spectra that are comprised of > 50 per cent transient flux over host. This has significance for spectroscopic classification as will be discussed in Section 5.1.2.

The full process used to create blended spectra as described across Sections 2 and 3 is summarized in Fig. 5.

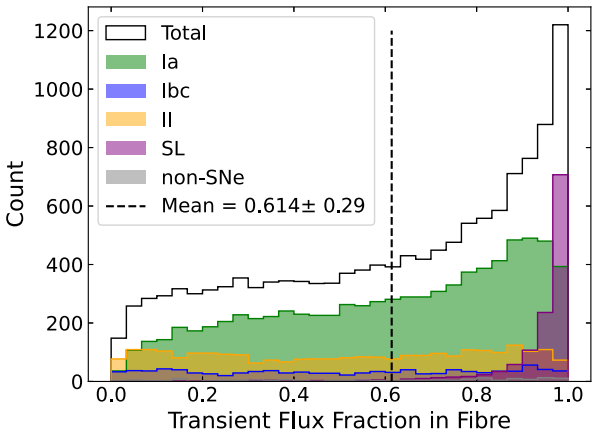


Figure 4. The distribution of transient flux fractions in the fibre. The mean value for all transients is highlighted with the dashed black line. As this accounts for fibre losses in the host galaxy, we see that over half of all of the spectra have more transient flux than host flux through the 4MOST fibre.

4 INDIVIDUAL CLASSIFIERS

4.1 Classifier overviews

4.1.1 DASH

DASH (Deep Automated Supernova and Host classifier) is a deep convolutional neural network. DASH is trained on a set of templates and learns spectral features. Input spectra are broken down into individual features, compared to the features in the training set and then assigned a softmax pseudo-probability to each of its classification bins, named so due to the softmax regression model in the final layer of the deep learning model. The softmax probabilities only are only relative probabilities for one classification bin compared to the others (Muthukrishna et al. 2019b). The highest pseudo-probabilities are then presented in the DASH Generated User Interface (GUI), and a combined softmax probability is produced by summing those of the best output bins until one is reached that either disagrees on transient class or is not in an adjacent phase bin. We discuss our method for converting the softmax probability for individual classification bins into probabilities for SN Ia, Ibc, etc., in Section 4.3. The softmax probability of a classification bin is not necessarily a judgement on the quality of the classification. If every classification bin fits very poorly, then the best fit is not necessarily a good fit (Muthukrishna et al. 2019b).

DASH also calculates an *rlap* cross-correlation value for each output classification bin as an additional flag for classification quality. The *rlap* parameter was originally developed for another transient classifier that we investigate, SNID (SuperNova IDentification). However, we do not make use of it for DASH.

rlap is the product of the correlation scale height ratio, r, and *lap*, an overlap parameter. r is defined as the ratio between the highest normalized cross-correlation peak, h, and the root-mean-square (RMS) error of the antisymmetric component of the cross-correlation product σ_a:

$$r = \frac{h}{\sqrt{2}\sigma_a} \tag{4}$$

lap is the overlap in ln(λ) space between the input and template spectra. A larger *rlap* value indicates more similarities between the input spectrum being classified and the template it is being compared to. Hence, larger *rlap* values indicate a better quality classification. The machine-learning aspect of DASH returns the best-fitting classification bin. Then, *rlap* values are calculated for each spectrum in DASH's training sample in that classification bin. The highest *rlap* produced is returned to the user, with a warning if it less than five. Details on DASH's template set can be found in Muthukrishna et al. (2019b). We do not make use of *rlap* in determining DASH's classification results.

DASH has four modes of operation defined by its ability to fit or not fit transient host galaxies and its ability to use or not use known redshift values. We only make use of the known and unknown redshift modes. In the unknown redshift mode, the redshift is estimated by maximizing *rlap* in redshift space.

Host fitting leads to an increase in the number of output classification bins as each output now has a host class attached to each output. This increase in output bins leads diluted softmax percentages on outputs. However, we note that including a host-fitting step in the classification could remove degeneracy between transient class and redshift. Unfortunately, the host-fitting mode does not function without redshifts provided. For this reason, we do not investigate it.

There are some concerns that must be kept in mind if DASH is to be used as a mechanism to classify transients. For example, while DASH is user-friendly, fast-working, and produces pure samples, it does so somewhat at the cost of user power. Compared to SNIDor NGSF (Next Generation SuperFit, Howell et al. 2005) the user's options are fairly limited. There is no front-end mechanism to pass an error function for weighting the fit or removing wavelength ranges with known contaminant features.

Additionally, and very importantly, the potential SN classes available for classification are somewhat limited. DASH can classify SNe Ia and common CC SNe like Ib/c, II, IIn, and IIP. However, no other classes are included in its training sample and so other classes in the population simulations such as SLSNe, TDEs, and CaRT cannot be classified. They are either 'other' results or contaminants. Some of these transient classes are fairly exotic and rare, but there are many SLSNe in the simulation, and for DASH, they can only act as a source of contaminant classifications.

4.1.2 NGSF

NGSF is a template matching SN classifier. Written in PYTHON, it is based on the Superfit classification package written in IDL (Howell et al. 2005). NGSF requires a set of transient and host templates to compare to the spectrum being classified. We use the updated template set recommended in the source.[3] The input spectrum is sequentially compared to each of these templates while iterating through a variety of redshifts, reddening corrections, and different levels of host contamination for a variety of morphologies. The redshift and reddening arrays that are checked are defined by the user. Each spectrum being fit must be compared to every template at every possible combination of reddening and redshift and for every host galaxy. As a result, the classification time required varies significantly with how fine the redshift sampling is (Goldwasser et al. 2022).

NGSF returns its classification in the form of a χ^2 value for each host, template, redshift, reddening combination. Input spectra are binned to match the templates and then a χ^2 value is obtained using the equation (reproduced from Howell et al. 2005):

$$\chi^2 = \sum \frac{[O(\lambda) - aT(\lambda; z)10^{cA_\lambda} - bG(\lambda; z)]^2}{\sigma(\lambda)^2} \tag{5}$$

[3]From the WISeREP repository, Yaron and Gal-Yam (2012).

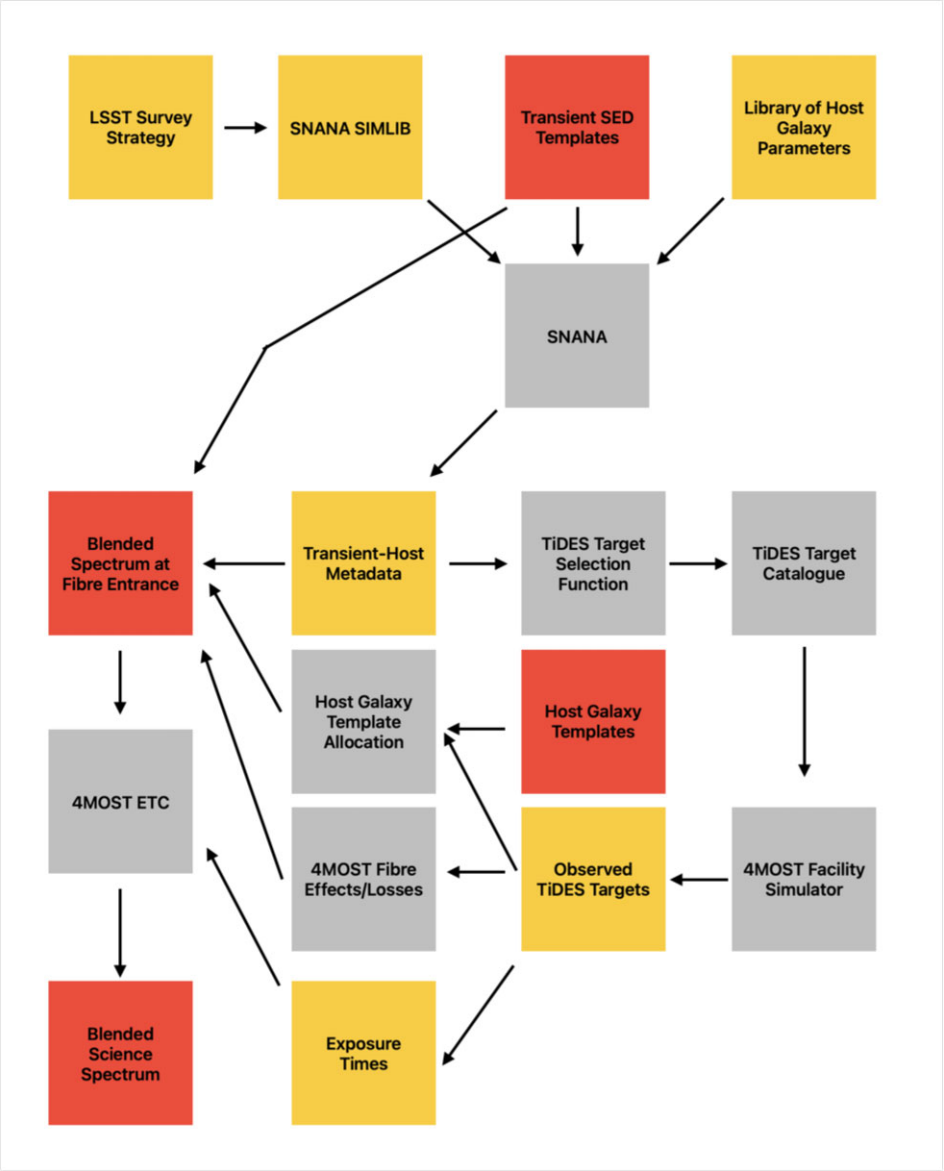


Figure 5. Flowchart showing our simulation pipeline. Adapted from fig. 1 of Frohmaier et al. (2025). Initially an LSST Operation Simulation (OpSim) is converted into a SNANA SIMLIB file. This, alongside a set of transient SEDs and a library of simulated host properties are used as inputs for a SNANA simulation that returns host–transient metadata and light curves. These are input into the TiDES selection function as if operating in real time. This produces a TiDES-specific target catalogue, for which the 4FS generates fibre allocations and exposure times. This gives us a list of observed TiDES targets and their observational properties. Host galaxy templates are assigned to observed transients. The blended spectra have magnitudes and redshifts assigned from the SNANA metadata. Fibre losses are simulated to generate the spectrum at the 4MOST fibre entrance. This spectrum and its assigned exposure time from 4FS are input into the 4MOST ETC which adds realistic noise to the spectrum, producing our final blended science spectrum. Red boxes indicate templates or SEDs, yellow boxes indicate catalogue-level results, and grey indicates a process or algorithm.

where O is the input spectrum, T is the transient template spectrum, G is the host galaxy template spectrum at a given redshift, z, $\sigma(\lambda)$ is the error on the input spectrum, and A_λ is the reddening law. a, b, and c are constants that are varied during the classification process to check the template fit at varying reddening levels and at varying levels of host contamination. NGSF uses the reddening law of Cardelli, Clayton & Mathis (1989). The templates with the lowest $\chi^2_{\rm red}$ is reported as the best template. As NGSF also iterates through different levels of host contamination for each template it returns the estimated galaxy fraction of the best-fitting templates. Since our spectra have known SN and host magnitudes in the fibre, this has potential as another method to judge classification quality.

The throughput in the simulated 4MOST spectra drops below 70 per cent approximately below 4000 Å and above 8000 Å. We chose to limit the NGSF template comparisons to this wavelength range. Since the ETC generates error spectra, we use these for calculating χ^2. In the case where the input spectrum has no attached error spectrum, NGSF has several options for generating error spectra which can be used as weights to calculate a reasonable χ^2 for the input, although these are not the intended methods. It can determine a linear error spectrum or a Savitzky–Golay (SG, Savitzky & Golay 1964) error spectrum.

The SG error spectrum is generated by smoothing the input spectrum with an SG filter and then subtracting the smoothed

spectrum from the original to obtain residuals that are used to construct an error spectrum. The linear error spectrum is constructed using a linear fit to the binned input spectrum. In both cases, this results in the smoothing of narrow features into noise, making both inferior to the use of an included error spectrum.

NGSF has several distinct advantages over DASH, mainly in the form of user control. For example, the ability to set a redshift or reddening constant range with specified values or the capacity to exclude noisy wavelength ranges.

The final, and perhaps most considerable advantage, is NGSF provides easy access to the set of templates it uses. This makes it very easy to update the templates manually to include more examples of existing subclasses or new subclasses altogether. Updates to either require no additional training time, which would be needed to change the templates used by DASH. NGSF's template set contains just over half as many transients as DASH and one-third of the individual spectra, not including galaxy templates.

4.1.3 SNID

SNID is an algorithm for determining the properties of an SN spectrum (Blondin & Tonry 2007). It makes use of cross-correlation techniques and the *rlap* quality parameter to find best-fitting redshifts, phases relative to maximum light, and classes for input templates. *rlap* is discussed in more detail in Section 4.1.1.

We use templates collected from various samples by Kim et al. (2022), where a more complete description can be found. Classifications were performed over the same 4000–8000 Å range as NGSF.

One advantage SNID has is the large variety of built-in transient classes and subclasses available for classification, as well as several morphologies of galaxy, active galactic nucleus (AGN), and a simple notSN classification amongst others that allow SNID to potentially identify non-transient spectra. DASH and NGSF have no capacity to do this. NGSF can easily have new templates added, but DASH would require computationally expensive retraining for the same effect.

Further, addition of more subclasses is very simple. New templates can be added to the SNID repository provided they are in the correct format. Then, the new classifications are added to a simple parameter file. In this paper, we have 30 distinct classifications (a few SLSNe and non-SN classes were added to those that came built-in). However, SNID still seems to perform very poorly when classifying non-SN Ia spectra. This will be discussed further in Section 4.5.

One issue we encounter with SNID is that it occasionally performs a classification wherein none of its templates yield an *rlap* value greater than $rlap_{\rm min}$ and no output is produced. In this case, we assign a best-fitting classification of 'None' which is automatically considered an 'other' classification.

4.2 Classification schema and statistical definitions

With simulated transient spectra realistically blended with host galaxy flux now in hand, we can begin to test spectroscopic transient classifiers. We test the DASH (Muthukrishna et al. 2019b), NGSF (Howell et al. 2005), and SNID (Blondin & Tonry 2011). These classifiers are introduced in Sections 4.1.1, 4.1.2, and 4.1.3, respectively. Our objective is to compare the performance of each classifier on our simulated spectra.

The standards by which we will judge the performance of the classifiers are the purity and completeness of their classifications. Purity and completeness are, for a target transient class, defined as:

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

Here, TP (true positive) are the number of spectra of the target class identified as such. FP (false positive) is the number of non-target class spectra misclassified as the target class. FN (false negative) is the number of target class spectra misclassified out of the target class. TN (true negative) classifications are spectra correctly identified as not being in the target class.

Outside of binary classifications, for a given transient class, the completeness is the fraction of that class that are successfully identified as such. The purity is the fraction of output classifications of that class which are correct. Thus, the rate of contamination in a transient class is 1 – purity for that class.

Throughout Sections 4 and 5 we will, alongside completeness and purity, report the F-score (F_β) for each classifier (Van Rijsbergen 1977) as our figure of merit (FoM). F_β values range between 0 and 1 indicating a poor and a strong classifier, respectively. (F_β) is defined as:

$$\text{F}_\beta = \frac{(1+\beta^2) \times \text{Purity} \times \text{Completeness}}{(\beta^2 \times \text{Purity}) + \text{Completeness}} \tag{8}$$

β is a constant used to preferentially weight the F_β towards completeness or purity. The two main transient objectives of TiDES are the real-time classification of all transients from the TiDES-Live program and the eventual production of an SNe Ia sample for the purpose of fitting cosmology. The number of SNe we expect to obtain from 4MOST–TiDES is orders of magnitude larger than previous surveys such as OzDES (Lidman et al. 2020) or the SuperNova Legacy Survey (Astier et al. 2006). With the large number of spectroscopically observed transients, we believe that purity is a more important factor than classification completeness. This is especially true for the SN Ia sample for cosmology, but even for real-time classification we choose to focus on pure samples.

With this in mind, we generally report the $\beta = 0.5$, $F_{0.5}$, score as our FoM. This assigns greater weight to the classification purity over the F_1-score that weights both metrics equally. To account for multiple classes, each transient class has an individual $F_{0.5}$-score calculated. Then, the average value is obtained by taking the mean, weighted by each class's prevalence in the sample.

Additionally, in Section 5.2, we will make use of the classification accuracy of our classifiers. This is particularly useful for comparison to photometric classifiers, which often use this parameter to quantify success. Accuracy is the fraction of classifications across all classes that are correct. In a binary schema, it is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

We do not aim for any particular purity threshold, but will add a 95 per cent purity line to relevant plots as an arbitrary point of comparison. This purity is similar to that found in SN Ia samples used in cosmology in the literature. For example, Howell et al. (2005) reported an 8 per cent non-SN Ia contamination rate (92 per cent purity) in their final sample of SNe Ia, while Campbell et al. (2013) reported a 3.9 per cent predicted contamination rate (96.1 per cent purity) that has an insignificant effect on their cosmological measurements. In Guy et al. (2010), purity ranges from 100 per cent to 90 per cent are found in various redshift bins up to $z = 1$ and again, they report that the effect on cosmology is minimal compared to other sources of error.

Each classifier returns a list of output classification bins in descending order of the quality metric specific to that classifier. This

Table 2. The SN Ia and non-SN Ia transient subclasses for each classifier. The non-SN Ia transients subclasses included here match the various non-SN Ia input classes listed in Table 1. Any output classifications not included in this table would be considered a misclassification if returned by a classifier.

Classifier	Binary class	5 classes	Corresponding outputs
DASH	SNe Ia	SNe Ia	Ia-norm, Ia-91T, Ia-91bg
.	Non-SN Ia	SNe Ibc	Ib-norm, Ib-pec, Ic-norm, Ic-broad
.	.	SNe II	Ib, IIP, II-pec, IIL, IIn
.	.	SLSNe	–
.	.	Non-SN	–
.	.	Other	Ia-pec, Ia-csm, Ia-02cx
NGSF	SNe Ia	SNe Ia	Ia-norm, Ia 91bg-like, Ia 91T-like, Ia 99aa-like
.	Non-SN Ia	SNe Ibc	Ibn, Ib, Ic, Ic-BL, Ic-pec, IIb
.	.	SNe II	II, II-flash, IIn, IIb-flash
.	.	SLSNe	SLSN-II, SLSN-IIn, SLSN-I, SLSN-Ib, SLSN-IIb
.	.	Non-SN	TDE H, TDE He, TDE H + He, FBOT, ILRT
.	.	Other	Ia 02es-like, Ia-02cx like, Ia-CSM-(ambigious), Ia-pec, Ia-CSM Ia-rapid, Ca-Ia, super-chandra, SN - Imposter, computed
SNID	SNe Ia	SNe Ia	Ia, Ia-norm, Ia-91T,Ia-91bg, Ia-99aa
.	Non-SN Ia	SNe Ibc	Ib, Ib-pec, Ib-norm, Ic, Ic-norm, Ic-pec, Ic-broad, IIb
.	.	SNe II	II, IIL, IIP, II-pec, IIn
.	.	SLSNe	SLSN, SLSN-I, SLSN-Ic, SLSN-IIn
.	.	Non-SN	TDE, Ca-rich, ILRT
.	.	Other	Ia-csm, Ia-pec, Ia-02cx, NotSN, AGN, None LBV, M-star, QSO, C-star, LRN, Gal

is softmax probability (and $rlap$) for DASH, χ^2 for NGSF, and $rlap$ for SNID as mentioned in Sections 4.1.1, 4.1.2, and 4.1.3, respectively. It is not clear if these quality metrics can be used in place of a probability or to what extent they can be compared. Additionally, as each classifier makes use of different templates either for training or matching, it is not necessarily reasonable to compare outputs from each classifier directly.

To determine the best output class for each classifier, we adapt the approach used in Kisley et al. (2023). A blended spectrum is input separately into each classifier. Then, for each classifier, the quality metric for each output classification is used to produce a probability that the input spectrum belongs to each of the output classes in the 5-class schema described in Section 4.2.

For DASH, this is a simple process as it already returns the softmax pseudo-probability for each classification bin. We simply sum the softmax probabilities for the outputs corresponding to each of the five classes and normalize the resulting probabilities by the summed total of all softmax probabilities.

For NGSF, we convert the returned χ^2 values into percentages by evaluating the cumulative density function at that particular χ^2. This is performed using the SCIPY PYTHON library. The resulting relative probabilities for each output are summed by class and normalized by dividing by the total probabilities for all outputs. When redshifts are provided the average number of reported outputs is 9.3. This jumps to over 50 when redshifts are not provided and often numbers of relatively spurious SLSN classifications can overweight that class as an output. To account for this, we only look at up to the 10 best classifications when redshifts are not provided.

For SNID, we are required to make a judgement call as the $rlap$ quality metric it returns is less readily converted to a probability than those of NGSF and DASH. In this case, we obtain the value of $r = rlap \times lap$ and convert it to a probability using the error function $\mathrm{erf}(r)$. For each class, we sum the probabilities for each output in that class and then normalize these into probabilities by dividing by the sum of all output probabilities. We only consider such output classifications that meet the default SNID requirement of $rlap_{\mathrm{min}} = 5$. Because of this, all outputs return probabilities close to unity, meaning that we weight each output nearly equally.

Following these procedures provides us, for each spectrum for each classifier, the probability that the input is an SN Ia, Ibc, II, an SLSN, a non-SN transient or a non-transient ('other') spectrum. This standardization of method allows for easy comparison of classification ability between the three classifiers.

We distinguish between SNe Ia that are 'cosmologically useful' and SNe Ia that are not. Ia-norm are counted as cosmologically useful, as are 91T-like SNe Ia. The latter are overbright, hot SN Ia and are usually included in cosmological samples (Ginolin et al. 2025). SNe Ia 91bg-like standardization for cosmology is debated (see Graur 2024, and references therein). Here, we consider them alongside Ia-norm inputs and output classifications. Any output that is not an SN Ia subclass is considered a non-Ia output.

To account for output classes for which we have no input spectra, we create the 'other' classification bin. This is a catch-all for automatic misclassifications from peculiar SN Ia subclasses (Ia-csm, Iax, etc.) or non-transient classes like 'Gal', 'm-star', 'None', etc. The list of 'other' classification outputs for each classifier are also included in Table 2. For the purposes of calculating completeness, classifications that end up in the 'other' class are considered FNs.

Some examples of successful and unsuccessful classifications are shown in Appendix B.

4.3 Binary classification results

In this section, we will be considering a binary classification. SNe will either be classified as an SN Ia or non-SN Ia. This is far fewer classes than each classifier has the potential to output, and we recognize that combining multiple output classes into a single, non-Ia class is not the same as requiring that a classifier chooses between two classes. We will also be tracking non-SN Ia transients that are misclassified as Ia contaminants.

Throughout this section, classification will be performed with known redshifts, simulating the case where a transient has a spectro-

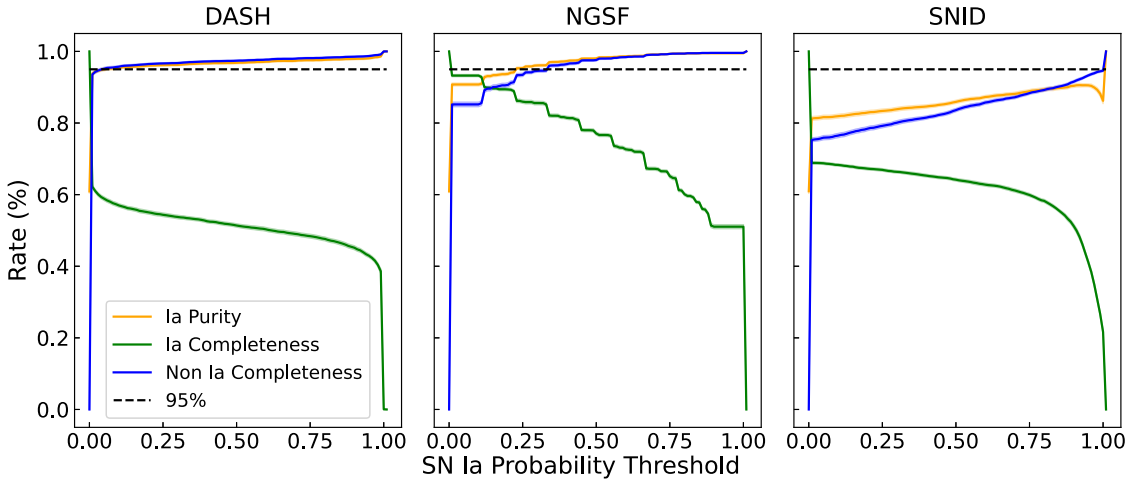


Figure 6. SN Ia completeness (green), SN Ia purity (orange), and non-SN Ia completeness (blue) as a function of SN Ia probability threshold for each classifier. Input spectra are considered an SN Ia output if the returned SN Ia probability is greater than a given threshold, regardless of whether a different class is more probable. A rate of 95 per cent is marked by the dashed black line as an arbitrary point of comparison.

scopic redshift determined from its host galaxy or its own emission features. In practice, this means that we provide the classifier with the true redshift from the simulation as a known redshift. Results for classification with unknown redshifts, or just photometric priors are shown in Section 4.5 and throughout Section 5.

We run the classifiers in non-interactive mode to mimic an automated classification plan for very large numbers of spectra. We note that this is not the way these classifiers were intended to run. Classifiers occasionally maximize their output metric with an incorrect classification, despite correct classifications being the second – or third – best result. For example, this can occur where two output class are similarly favoured (say SN Ib and Ic) or where a completely spurious output classification is found due to redshift inaccuracy (a high-z SLSN classed as a low-z SNe Ia). By using all reported classifications from a classifier and converting to a probability for each of our output classes, we avoid this issue.

Our method of converting classifier outputs into probabilities returns the probability that a transient belongs to the SN Ia, Ibc, II, SL, or non-SN transient classes defined in Table 2. In this section, we consider only the SN Ia probability and a binary SN Ia–non SN Ia classification schema. If the SN Ia probability exceeds an arbitrary threshold then that classifier will report it as an SN Ia, regardless of the probabilities of the other four classes. In Section 4.5, where we consider the full 5-class schema, we will swap to having the classifiers report each transient as whichever of the five classes has the greatest probability.

In Fig. 6, we investigate the SN Ia completeness, purity and non-SN Ia completeness for each classifier as a function of an SN Ia probability threshold. We can see that it is not immediately clear if a probability threshold should be applied for any of the classifiers. DASH's SN Ia completeness, purity and non-SN Ia completeness remain almost constant for most SN Ia probability thresholds. Only at very low thresholds do we report purities under 95 per cent and only at very high thresholds do we see a large loss in SN Ia completeness. One could reasonably assign 0.5 as the required SN Ia probability to be considered an SN Ia.

Similarly SNID could reasonably have an SN Ia probability threshold set anywhere between 0.5 and 0.8. Below this, we see significant losses to SN Ia purity, and above this, we see the same sudden loss in SN Ia completeness as displayed by DASH.

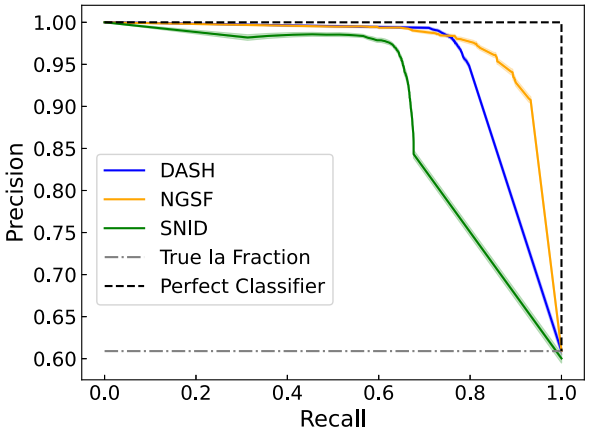


Figure 7. Purity–completeness (precision–recall) curves for each of DASH, NGSF, and SNID in the case of binary SN Ia–non-SN Ia classification. A theoretical, perfect, binary classifier is presented by the black dashed line. The closer a classifier's curve matches the perfect classifier, the better that classifier is performing. The grey dashed line indicates the fraction of input spectra that are SN Ia, which is the minimum possible purity obtained when the SN Ia probability threshold is set to zero.

NGSF is the only classifier to show a different trend. Here, the SN Ia purity and non-SN Ia completeness quickly rise to unity. Meanwhile the SN Ia completeness starts at unity for no probability threshold, before steadily dropping as the threshold is made more stringent. A case could be made to perform NGSF classification with an SN Ia probability threshold of anywhere from 10–25 per cent. Above this and the only change is a loss in SN Ia completeness.

In Fig. 7, we present purity–completeness (also known as precision–recall) curves for all three classifiers. A theoretically perfect classifier is shown as a point of comparison. A perfect classifier will return perfect purity at all levels of completeness as determined by varying the SN Ia probability threshold used to calculate each parameter. The only exception is the case where the threshold is set to zero. In this case, the completeness is 100 per cent by definition, while the purity drops to match the fraction of the total input sample that are actually SN Ia, which is approximately 60 per cent in this case.

We can see from Fig. 7 that NGSF performs closest to the theoretically perfect classifier. NGSF is followed by DASH and SNID in that order. The uncertainty for each classifier, indicated by the transparent shaded regions around each curve, indicates an uncertainty on the order of 0.5 per cent. This suggests that the classification results are stable across random samples of the full transient population. In other words, based on Fig. 7, we would expect that NGSF outperforms DASH and SNID across all of our subsamples under this binary classification schema. However, Fig. 7 gives very little information about the non-Ia transients. For example, NGSF could classify all SNe Ib as SLSNe, and in this schema, this would constitute perfect classification.

We do not report the numerical results for binary classification as the SN Ia classification is unchanged and allowing any non-Ia input to be 'successfully' classified as any non-Ia output significantly inflates the non-SN Ia classification completeness and purity.

4.4 Redshift priors

Using the SN Ia probability as a threshold gives a good indicator of the completeness and purities, we can expect for each classifier and, also, allows use to construct purity–completeness curves that indicate that NGSF is the best-performing classifier in our binary schema. However, in this section, we will proceed assuming that the output classification with the highest probability for each classifier is that classifier's output. This is partially to remove our need to assign arbitrary and distinct probability thresholds to each classifier and because it is the only method that is applicable for non-binary classification schemes. This avoids the situation where the SN Ia probability exceeds the threshold while being less than the probability that the transient belongs to a different class.

We test each classifier both with and without redshift priors. Using redshift priors means that for each input spectrum we provide the classifiers with the true transient redshift as found in the input population simulation. In the case of using unknown redshifts, we give no redshift information to DASH and SNID. NGSF is instructed to check redshifts between $0 < z < 1.5$ with a sampling of $\Delta z = 0.05$.

Perhaps one of the most likely scenarios during the operation of TiDES–4MOST is the case where we will not have a spectroscopic redshift, but will have a photometric redshift estimate. We would like to be able to investigate classifier performance in this scenario.

The minimum science requirement for LSST–DESC as reported in The LSST Dark Energy Science Collaboration (2018) is that the RMS scatter between photometric redshifts and true redshifts should not exceed $0.03(1 + z)$. Graham et al. (2018) and Mitra et al. (2023) investigate LSST photometric redshifts instead assuming $0.02(1 + z)$ as the RMS error between photometric and spectroscopic redshifts. We will proceed using the 2 per cent uncertainty.

For NGSF and SNID, we are able to simulate the use of photometric redshift priors. We randomly generate a photometric redshift ($z_{\rm phot}$) from a Gaussian distribution centred on the true redshift and with width equal to 2 per cent of $1 + z$. Then, we have each classifier attempt an 'unknown' redshift classification over the truncated redshift range defined by a 2 per cent uncertainty in $1 + z_{\rm phot}$.

Unfortunately, DASH does not natively have the option to attempt classification over a custom redshift range. The only way for DASH to simulate photometric redshift priors is to have each classifier fit the randomly generated $z_{\rm phot}$ as a known redshift, which would prohibit a direct comparison to NGSF and SNID. We found that this fitting of a 'known', but slightly incorrect, redshift resulted in poorer performance than providing no redshift at all.

Because of this, we do not report on the classification potential of photometric redshifts throughout the paper. However, for completeness, we do report the results from NGSF and SNID using them in the unknown redshift mode over a custom redshift range as described previously and making use of the 5-class classification schema as used in Section 4.5. These results are found in Table 3 alongside the known and unknown redshift classification results. Additionally, when discussing combined classifiers in Section 5, we report the SN Ia completeness and purity for the combined NGSF–SNID classifier using photo-z priors.

4.5 5-class classification

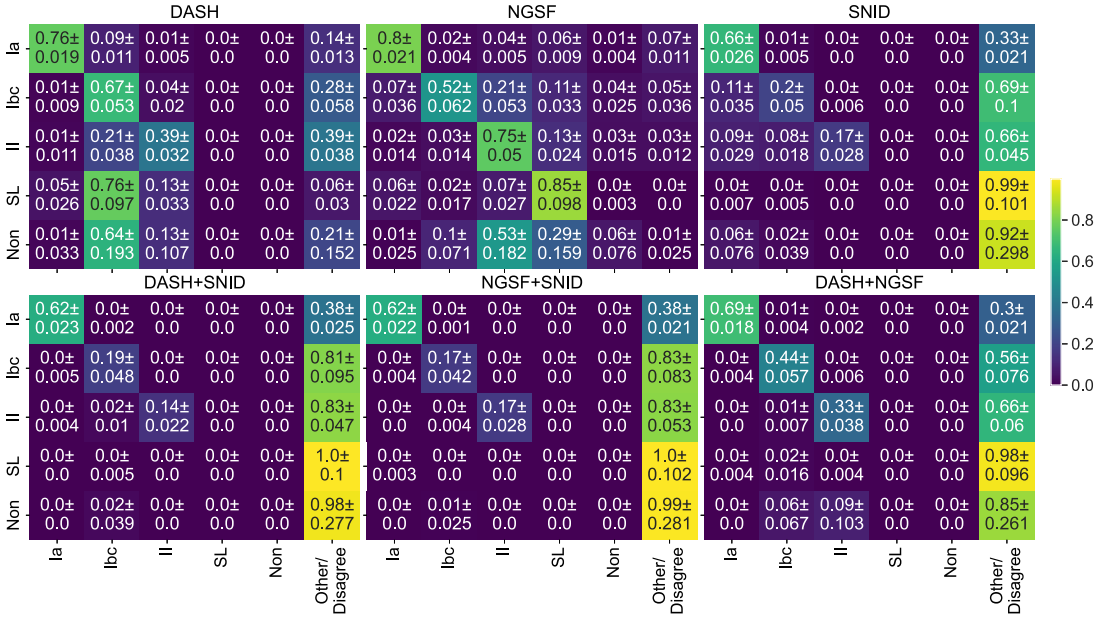
In this section, we make use of a classification system that includes five transient classes: SNe Ia, SNe Ibc, SNe II, SLSNe, and non-SN transients, following the work of Kim et al. (2024). The breakdown of classifier output subclasses that correspond to each of these inputs is indicated in Table 2.

Table 3 shows the blended spectra being classified with the non-SN Ia transient output bin divided into SNe Ibc, SNe II, SLSNe, and non-SN transients.

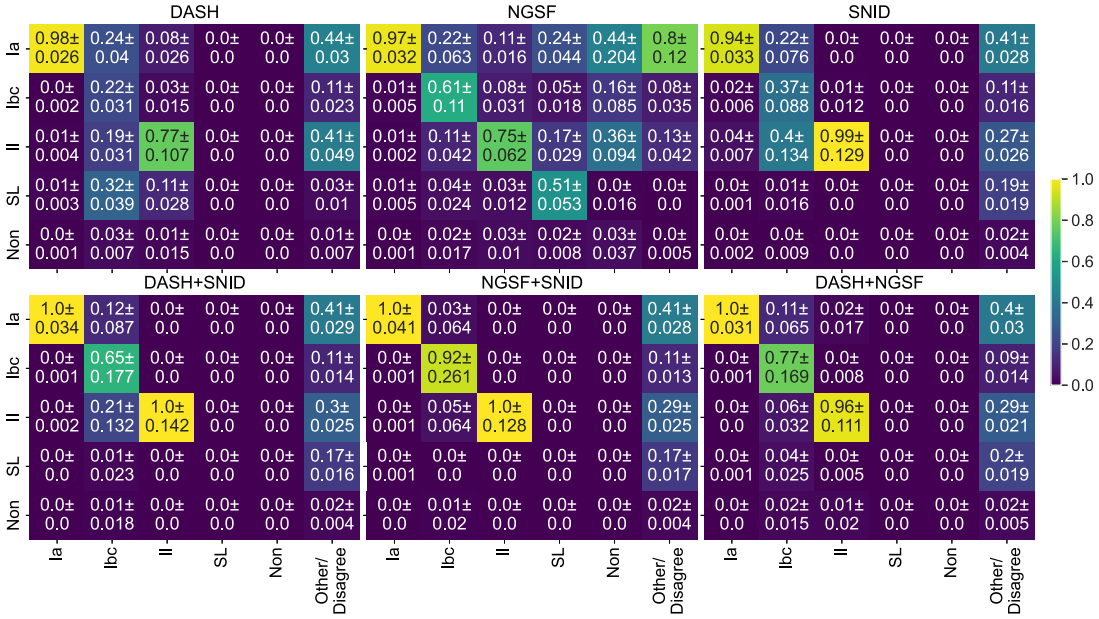
The 5-class schema allows us to see finer detail about each classifier's ability to classify CC SNe and non-SN transients. This is particularly relevant for judging a classifier's ability to perform live TiDES classification across a range of different transient classes. $F_{0.5}$-scores reported throughout this section are the population size-weighted average of the $F_{0.5}$-scores of the five individual classes.

Table 3. The completeness for classifying SNe Ia, SNe Ibc, SNe II, SLSNe, and non-SN transients. Also presented are the SN Ia purity and the $F_{0.5}$-score for each classifier. The highest value in each column is highlighted in bold. Classification with photometric priors for NGSF and SNID are provided alongside known and unknown redshift classification. $F_{0.5}$-score is calculated based on the average scores of all five transient classes reported, weighted by their population size.

Classifier	Ia completeness	Ibc completeness	II completeness	SL completeness	Non-SN completeness	Ia purity	$F_{0.5}$-score
DASH, known z	0.760 ± 0.004	0.68 ± 0.01	0.39 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{0.981 \pm 0.002}$	0.711 ± 0.003
DASH, unknown z	0.516 ± 0.004	$\mathbf{0.69 \pm 0.02}$	0.32 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.968 ± 0.003	0.639 ± 0.003
DASH, photo z	–	–	–	–	–	–	
NGSF, known z	$\mathbf{0.798 \pm 0.005}$	0.52 ± 0.02	$\mathbf{0.753 \pm 0.006}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.05 \pm 0.02}$	0.971 ± 0.002	$\mathbf{0.814 \pm 0.004}$
NGSF, unknown z	0.560 ± 0.006	0.39 ± 0.02	0.35 ± 0.01	0.25 ± 0.01	0.02 ± 0.01	0.917 ± 0.003	0.627 ± 0.005
NGSF, photo-z	0.551 ± 0.006	0.48 ± 0.01	0.563 ± 0.008	0.85 ± 0.01	0.03 ± 0.01	0.935 ± 0.002	0.699 ± 0.003
SNID, known z	0.661 ± 0.006	0.20 ± 0.01	0.174 ± 0.007	0.0 ± 0.0	0.0 ± 0.0	0.929 ± 0.004	0.649 ± 0.003
SNID, unknown z	0.661 ± 0.006	0.15 ± 0.01	0.167 ± 0.006	0.0 ± 0.0	0.0 ± 0.0	0.835 ± 0.004	0.585 ± 0.005
SNID, photo-z	0.644 ± 0.004	0.11 ± 0.01	0.083 ± 0.005	0.0 ± 0.0	0.0 ± 0.0	0.850 ± 0.005	0.552 ± 0.007



(a) Normalised by Row



(b) Normalised by Column

Figure 8. Confusion matrices showing the results for the three individual classifiers and all three combinations of two of the three classifiers working simultaneously. Confusion matrices are normalized by (a) row, indicating completeness in each class and (b) column, indicating the purity of each class. The 'other' output classification is reserved for output classifications with no corresponding input class and, in the case of the combined classifiers, an input spectrum that causes the two classifiers to disagree on the output class. Classification was performed with redshift priors provided in all cases. High completeness and purity samples would be indicated by high concentration along the matrix diagonal. Horizontal scatter indicates loss of completeness, and vertical scatter indicates loss of purity.

The results from Table 3 are presented as confusion matrices for the case with known redshifts in Fig. 8(a).

As mentioned in Section 4.4, we take the output classification with the highest probability for each input spectrum as the output class, or best class. We impose no additional limit on the best class's probability beyond it being the highest probability. Across all classifiers we see small uncertainties (1–2 per cent) on purity and completeness, indicating that the classification rates are stable.

In every case, the classifier's training sets are dominated by SNe Ia. This may lead to DASH overweighting features learned from SNe Ia templates, resulting in an increased likelihood that an SN Ia classification bins will be amongst DASH's top classification. Similarly, SNID and NGSF, when the input does not match well with any of their templates, and lacking a redshift to help discount templates, are most likely to find SN Ia templates as the best-matching templates as SNe Ia are the majority of their template banks. Across

all classifiers, there is potential for SNe Ia to be the best matches in the absence of any good matches.

More detailed discussion on how input SN Ia spectra are being classified by DASH, SNID, and NGSF can be found in Appendix A, in Fig. A1. Similarly, more detailed discussion on the origin of contaminant classifications for each classifier can be found in Fig. A2.

4.5.1 *DASH results*

We see that our DASH results, both with and without redshift priors, have very impressive SN Ia purities well over 95 per cent. However, the SN Ia completeness, while fairly good with redshift priors, falls to just above 50 per cent without. This is the largest drop in performance upon the removal of redshift information, alongside NGSF's loss of SN Ia completeness.

It becomes apparent that DASH is reasonably successful at classifying SNe Ibc when redshift priors are provided, but is far less successful at classifying type II SNe. Unlike what we see in its SN Ia completeness, when redshift priors are removed, there is not much change in performance for Type II SNe. The Ibc classification completeness actually improves slightly, while the Type II classification completeness decreases, but by far less than that of the SNe Ia. It cannot be stated strongly enough that DASH natively lacks all capacity to classify SLSNe and the various non-SN transients. Indeed, in Section 5, all combinations of classifiers that include DASH are incapable of successfully classifying any SLSNe or non-SN input spectra.

Additionally, there is significant classification of input spectra into peculiar-Ia subclasses, often SN Ia-csm. This is particularly prevalent in transient spectra with Sc-type host galaxies, which make up a large fraction of our SN Ia hosts (Hakobyan et al. 2012), likely due to emission lines present in the host template. The narrow emission lines from the host are misinterpreted as circumstellar medium (CSM) interaction, leading to a Ia-csm classification.

Strangely, only DASH's outputs exhibit this trend. Where 40 per cent of Sc-type hosts produce a Ia-csm classification in DASH, less than 1 per cent do in both NGSF and SNID. Fortunately, this has no effect on classification purities in any class as SN Ia-csm is considered peculiar and outputs of Ia-csm are not included in final samples. However, it does have a significant effect on completeness.

4.5.2 *NGSF results*

NGSF and DASH classify SNe Ia very similarly when redshift priors are provided. The difference in completeness for SN Ia (79.8.3 per cent versus 76.0 per cent) is slightly in favour of NGSF, the purity of the resulting SN Ia samples are almost identical, within 2 percentage points of each other. When removing redshift priors we see a loss of performance across Ia classification for both classifiers. The SN Ia classification completeness difference is similarly sized as in the case where redshifts are known, with NGSF reporting 5 per cent higher completeness. However, while DASH reports only very slightly reduced (by less than a single percentage point) SN Ia purity, NGSF's corresponding rate drops by around 5 percentage points when redshift information is not provided.

In the 5-class scheme, the finer non-SN Ia output classes leads to mixed classification results for NGSF. The Ibc completeness is fair at just over 50 per cent with redshift priors. The non-SN transient completeness is very poor, well under 10 per cent with and without redshift priors (see Appendix C), although NGSF is the only classifier that gets any of these input spectra correct. NGSF produces particularly impressive completeness in SN II and SLSN classifications when redshift priors are provided, but also reports drops in completeness of around 50 percentage points when redshift priors are not provided. This is still much better than SNID, which classifies no input SLSNe correctly, and DASH which, as mentioned previously, cannot classify them.

With redshift information, NGSF is the strongest classifier in terms of classification completeness. Only DASH exceeds it in SNe Ibc completeness. Without redshifts, the balance between NGSF and DASHis far closer due to NGSF's far larger loss of performance.

Indeed, when considering only the $F_{0.5}$-scores, NGSF is now clearly the best-performing classifier when redshifts are known. This is by a large margin, at least 0.1 larger than that of DASH or SNID. With unknown redshifts all three classifiers have $F_{0.5}$-scores between 0.58 and 0.64. Here, DASH's score is heavily influenced by its superior SNe Ia purity, which is heavily weighted in our weighted $F_{0.5}$-score.

As would be expected, if a slightly incorrect photometric redshift (see Section 4.4) with a small range of redshift values about it to consider is provided, performance improves compared to receiving no redshift at all. The $F_{0.5}$-score for NGSF with photo-zs fall between that produced by known (spectroscopic) and unknown redshifts.

4.5.3 *SNID results*

SNID has a much lower SN Ia completeness than DASH and NGSF when given redshift priors, and with unknown redshifts we see a significant drop in performance in the SN Ia purity metric. However, without redshift priors we do see it outperform DASH and NGSF in regards the SN Ia completeness. In fact, its SN Ia completeness is nearly invariant under a lack of redshift information. However, while the SN Ia completeness is maintained, this must be balanced against the significant drop in SN Ia purity, which leads SNID to a poorer $F_{0.5}$-score than DASH or NGSF without redshift information.

SNID produces poor classification completenesses in all non-SN Ia transient subclasses in the 5-class schema. With or without redshift information, it only achieves SN Ibc and II completenesses between 10 per cent and 20 per cent. Like DASH, it classifies no SLSN or non-SN transient correctly, but while DASH is incapable of outputting such classifications, SNID instead fails to do so. A large number of our blended spectra are classified as 'Gal' (a galaxy template) by SNID, leading to an 'other' output. It appears that galaxy contamination may be a limiting factor. Indeed, NGSF is trained to classify host and transient simultaneously which may explain its superior performance.

When photometric classification is possible, the results are the opposite of that seen with NGSF. For all transient classes with classification completeness greater than zero without redshift information, the completeness is lower with photometric priors. SNID's SN Ia purity does improve with photometric redshifts relative to a lack of redshift information, but the final $F_{0.5}$-score is still lower. SLSNe are well classified by NGSF, as photo-zs force the classification into the superluminous regime, yet this does not appear to occur in SNID.

It should be noted that SNID was intended to have significant human oversight in classification, so relatively poor results under complete automation are not unexpected. Additionally, while SNID's $F_{0.5}$-score is lower than the other two classifiers, its F_1- or F_2-scores are not. As SNID maintains SN Ia completeness when redshifts are unknown, and so F_β-scores that are weighted to more heavily favour completeness ($\beta > 1$) lead to SNID matching NGSF's performance and exceeding DASH's when redshifts are unknown.

Table 4. The SN Ia completeness and purity for all possible combinations of two or three classifiers. Successful classification requires an SN Ia output from all involved classifiers. For the combined NGSF–SNID classifier, we also report the same results assuming the presence of photometric priors. The highest value in each column is highlighted in bold.

Classifiers	Redshift	Ia completeness	Ia purity	$F_{0.5}$-score
DASH and NGSF	Known z	**0.689 ± 0.005**	0.9995 ± 0.0003	**0.757 ± 0.004**
NGSF and SNID	.	0.621 ± 0.006	0.9994 ± 0.0003	0.687 ± 0.005
DASH and SNID	.	0.623 ± 0.006	0.9984 ± 0.0004	0.674 ± 0.006
All	.	0.590 ± 0.006	**1.0 ± 0.0**	0.669 ± 0.005
DASH and NGSF	Unknown z	0.367 ± 0.004	0.997 ± 0.001	0.566 ± 0.006
NGSF and SNID	.	0.424 ± 0.006	0.976 ± 0.004	0.566 ± 0.006
DASH and SNID	.	**0.456 ± 0.006**	0.991 ± 0.001	**0.589 ± 0.006**
All	.	0.324 ± 0.005	**0.998 ± 0.001**	0.510 ± 0.006
NGSF and SNID	Photo-z	0.427 ± 0.007	0.990 ± 0.001	0.553 ± 0.004

5 USING MULTIPLE CLASSIFIERS AT ONCE

For both live classification of transients and when creating SN Ia samples for cosmology, it is critical to limit contamination in the output sample. For live classification, this is important for all SN classes. For cosmology, it only matters that the SN Ia sample is of high purity, even to the detriment of the SN Ia completeness. This is particularly true given the very large number of transients that 4MOST is expected to observe. Table 3 shows that individual classifiers struggle to limit contamination in the output SN Ia sample and are poor classifiers of even broad non-Ia SN classes. The obvious question is: what is the result of combining the classifications from different classifiers for each transient?

We first investigate the effect of classifying spectra with all combinations of two out of the three classifiers. In these cases, if both classifiers are not in agreement on the output classification, then the result defaults to an 'other' output regardless of the quality of either classification. Any output classifications from individual classifiers that do not match any of our potential output classes (Ia-pec, non-transients, etc.) are also discarded as 'other' outputs.

Fig. 8 shows that when using known redshifts, requiring two classifiers to agree has the effect of reducing the overall completeness for all five original output classes and a large increase in the number of 'other' outputs compared to the individual classifier results. However, we also see a large increase in the purity of SNe Ia, SNe II and, to a lesser extent, SNe Ibc. This can be seen by high concentrations along the confusion matrix diagonals.

The extreme case for a combined classifier is to use all of DASH, NGSF, and SNID simultaneously. The results for SNe Ia are shown in Table 4. With the combination of all three classifiers, we now classify around 60 per cent of all SNe Ia when redshifts priors are provided, but get very few successful classifications for any other input class. The sample of classified SNe Ia produced by this combined classification is completely pure.

Without redshifts we report reduced success. While SN Ia purities remain very high, the non-SN Ia completenesses remain around 10 per cent or less and the SN Ia completeness is nearly halved to 33 per cent. This is very low compared to other combined and individual classifiers. It remains to be determined where exactly the optimum balance lies between pure and large SN Ia samples for the purposes of cosmology. Regardless, combined classification has the promising ability to improve SNe Ia, II and, to a lesser extent, SNe Ibc purity.

Using all three classifiers, 87 per cent of SNe II are misclassified as 'other' or SNe Ibc. However, in this case the purity of output SN II sample is very high. In fact, by using a combined classifier consisting only of DASH and NGSF, we retrieve some of the classification completeness, classifying just under a third of SNe II successfully to produce a sample that is 96.4 per cent pure. Similarly, one can obtain a 77 per cent pure sample of SNe Ibc, although this can be improved to 92 per cent at the cost of only one-third of the completeness (44 per cent to just 17 per cent) if DASH–SNID is used instead.

Due to DASH's presence in this combined classifier, the classification completenesses of SLSNe and non-SN transients are zero. Indeed this can also be seen in Fig. 8, in both double classifier combinations including DASH, which cannot output SLSN classifications without retraining with a different template set that contains SLSN spectra.

The poor classification completeness shown in Fig. 8(a) and Table 4 suggests that the use of combined classifiers alone is not particularly appropriate for live transient classification. However, it does indicate the potential for very pure SN Ia and SN II samples, although the latter sample has very low classification completeness. As a result, combined classifiers could still form an important part of a live classification plan.

A combined classifier could be used as a first classification step to remove this high purity SN Ia sample prior to additional, later classification steps. Depending on the classifier used, this can also be done for the very pure (but low completeness) SN II sample produced. When spectroscopic redshifts are known, DASH–NGSF is an obvious choice due to its high $F_{0.5}$-score. Without redshifts it should be noted that a DASH–SNID classifier returns the best $F_{0.5}$-score. The marginally reduced purity is compensated by the higher completeness. However, unlike the case of known redshifts where DASH–NGSF is clearly the best-performing classifier, when redshifts are not known all three double classifiers have similar $F_{0.5}$-scores. Both with spectroscopic redshifts and unknown redshifts, when using all three classifiers, the reduction in completeness is more significant than the negligible improvement in purity compared to classifying with DASH–NGSF only. We investigate the potential for a second stage of classification in Section 5.1.

We conclude that that the best-performing classifier is DASH–NGSF. When redshifts are known, the SN Ia and SN II completenesses is 10 percentage points higher or more than using all three classifiers. This amounts to the addition of hundreds of transients into the final sample at the cost of doubling an already negligible non-SN Ia contamination. In the case where redshifts are not known this logic holds true, but with a combination of DASH and SNID. As shown in Table 3, NGSF is particularly affected by a lack of redshift information. However, without redshift priors, all three double classifiers perform similarly with regard to $F_{0.5}$-scores.

5.1 Potential photometric cuts

Individually, we see mixed results from the classifiers. Depending on the classifier and redshift information used, completeness can change by up to 50 per cent and SN Ia purities by as much as 15 per cent. From a cosmology perspective, we obtain both high-purity and reasonably high completeness in SN Ia classification from DASH and NGSF, but only when redshift information is known, and it is yet unclear to what extent prior redshift information will be available for TiDES transients.

From a live classification perspective, there appears to be no single classifier from which we can expect a reasonable classification completeness across the SN Ibc, II, SL, and non-SN classes. More importantly, the result of these low completenesses is that misclassified transients must be contributing to lowering the purity of some other class.

To this point, we have attempted classification on every transient that has received any exposure time in the survey simulation. We will now investigate two obvious sources of 'other' classification to see if applying cuts to the sample prior to classification will improve results. In Section 5.1.2, we investigate making cuts on the fraction of fibre flux deriving from the transient (as opposed to its host galaxy), and in Section 5.1.1, we investigate cuts based on the brightness of the transient. Both of these quantities should be reasonably obtainable from the same LSST photometry that TiDES will use to flag potential transient targets.

In both cases, photometric cuts are performed based on the LSST r-band magnitude at the time of simulated 4MOST observation. The transients in the simulation are binned in phase every five days and so there may be a discrepancy between of a few days between the simulated observation and the date of the reported magnitude. In reality, transients added to the 4MOST observing queue, for which we know the triggering magnitude from LSST, will only remain in the 4MOST observing queue for four days (Frohmaier et al. 2025) before needing refreshed with fresh photometry. So a discrepancy of several days between last known magnitude and 4MOST observation is realistic. We expect transient alert packets from LSST to be sufficient to perform the following photometric cuts.

5.1.1 Apparent transient magnitude

The most obvious sample cut that can be introduced from photometric information is a cut on transient magnitude. In this section, we investigate the potential for applying a cut to our transient sample based on the r-band magnitude of the transient.

Fig. 9 presents the completeness and purity of SN Ia classification for all three classifiers as a function of transient r-band magnitude. It also proposes two potential values for a transient magnitude cut to our sample. These values, 21.8 and 22.5 mag, are derived in Frohmaier et al. (2025) as the magnitudes that correspond to transient spectral SNRs of 5 and 3, respectively, where spectral SNR is calculated as the average in 15 Å bins between 3500 and 8000 Å. Indeed Frohmaier et al. (2025) report the SNR = 5 threshold as the conservative minimum to meet TiDES's spectral success criteria, with the SNR = 3 limit a more optimistic estimate based on the work of Balland et al. (2009). Here, we find that these SNR cuts of 5 and 3 correspond roughly to the SN Ia completeness falling below 80 per cent and the purity falling 95 per cent, respectively.

As NGSF produced the best individual $F_{0.5}$, in Table 5 we present classification results from NGSF, but now with the effects of cutting transients fainter than 21.8 and 22.5 mag. This does remove nearly half of the transients from the final sample for the stricter 21.8 mag cut. However, we generally see significant improvements across SN Ia completeness, SN Ia purity and $F_{0.5}$-score as stricter magnitude cuts are employed.

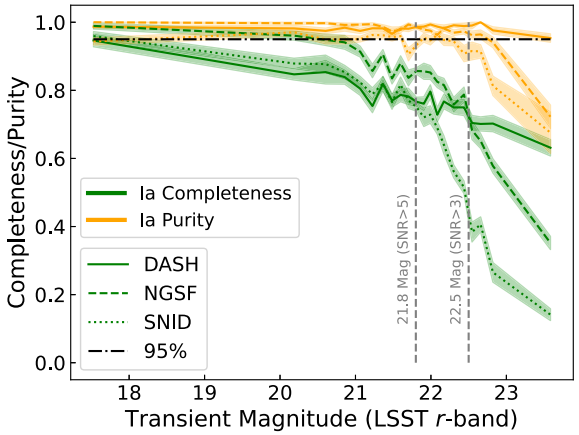


Figure 9. The SN Ia purity (orange, upper lines) and completeness (green, lower lines) as report by DASH, NGSF, and SNID as a function of the true transient magnitude for the SNe Ia in all of our subsamples. The SNe Ia are in non-linear magnitude bins of ∼30 transients, with each plotted point at its bin's centre. The shaded areas indicate the standard error on the mean of completeness and purity in each bin. 95 per cent purity is marked by a black dashed line. Two potential transient magnitude cuts are marked by grey dashed lines at 21.8 and 22.5 mag. We find that these limits roughly correspond to completeness dropping below 80 per cent and purity falling below 95 per cent, respectively.

DASH and SNID, while not shown, also follow this trend. NGSF outperforms SNID across all metrics both with and without redshift priors. However, without redshifts DASH does produce $F_{0.5}$-score about 0.01 larger than NGSF, mainly the result of DASH maintaining a high Sn Ia purity which is very heavily weighted in the $F_{0.5}$-score. However, NGSF, with spectroscopic redshifts, produces $F_{0.5}$-scores around 0.1 larger than DASH or SNID.

Cutting on r-band magnitude results in a significant reduction in sample size, so this would not be appropriate by itself for automatic classification. However, it could serve as a useful step in a pipeline for broad classification.

In Section 5 we found that, while combined classifiers are very good at creating high purity, low completeness SN Ia samples, they are poor classifiers of non-SN Ia classes. This makes them ineffective for TiDES live transient classifications. We also found in Sections 4.3 and 4.5, that the individual classifiers produce mediocre completeness and purity in most transient classes when operating on every transient observed in the 4MOST survey simulation. However, for TiDES transients brighter than $r = 21.8$ mag, NGSF appears to be a good choice for automated live classification.

However, this comes with several caveats. First, there will be significant performance loss when redshift information cannot be provided. Second, this only applies with relatively broad transient classes. For example, NGSF often classifies Ib-norm inputs as SN Ic subclasses. Just under 50 per cent of Ibc classification are SNe Ib classified as SNe Ic and vice versa. Finally, and perhaps most importantly, while the SNe Ia purity is high, the purity of the other classification bins can be far lower. For example the SN II purity is 77 per cent, and the Ibc purity is just 70 per cent (see Table 6).

From the point of view of the potential cosmology sample of SNe Ia obtained in Section 5, cutting transients from our sample based on their apparent magnitudes has less impact on the purity than the completeness. All three classifiers see between 0–4 per cent

Table 5. Ia classification results and 5-class weighted $F_{0.5}$-score for NGSF. We report the results with r-band magnitude cuts of 21.8 and 22.5 mag, as well as with no cuts. Completeness and $F_{0.5}$-score are calculated with the sample size after the cut is applied, but we note that mean Ia sample is reduced in size to 55 per cent and 83 per cent by magnitude cuts at 21.8 and 22.5 mag, respectively. The highest values in each column are highlighted in bold.

Redshift prior	r-band cut	Ia completeness	Ia purity	$F_{0.5}$-score
Known z	21.8	**0.882 ± 0.005**	**0.987 ± 0.002**	**0.876 ± 0.003**
.	22.5	0.837 ± 0.005	0.981 ± 0.002	0.842 ± 0.003
.	None	0.798 ± 0.005	0.971 ± 0.002	0.814 ± 0.004
Unknown z	21.8	**0.606 ± 0.006**	**0.936 ± 0.005**	**0.668 ± 0.006**
.	22.5	0.585 ± 0.006	0.933 ± 0.005	0.655 ± 0.005
.	None	0.560 ± 0.006	0.917 ± 0.005	0.627 ± 0.005

Table 6. The completeness and purity of each of our classes in the 5-class scheme under photometric cuts. The magnitude cut requires SNe r-band magnitude <21.8 and reduces the sample size to 61.7 per cent. The flux fraction cut requires that transient flux fraction >0.3 and reduces the sample size to 80.3 per cent. Using both reduces the sample size to 52.2 per cent. Completeness and F_1-score are the based on the transients in the classified sample, so objects removed by the photometric cuts do not contribute. Only NGSF is shown, having been identified as the most promising candidate for live classification. Bold values indicate the highest percentage in that row.

Metric	No cut	Mag. cut	Flux frac. cut	Both
Ia comp.	0.798 ± 0.005	0.882 ± 0.005	0.888 ± 0.004	**0.952 ± 0.003**
Ia purity	0.971 ± 0.002	0.987 ± 0.002	0.973 ± 0.002	**0.988 ± 0.002**
Ibc comp.	0.52 ± 0.02	0.65 ± 0.02	0.64 ± 0.02	**0.75 ± 0.02**
Ibc purity	0.61 ± 0.02	0.70 ± 0.02	0.72 ± 0.02	**0.84 ± 0.02**
II comp.	0.753 ± 0.006	0.836 ± 0.009	0.78 ± 0.01	**0.88 ± 0.01**
II purity	0.748 ± 0.009	0.767 ± 0.007	0.847 ± 0.007	**0.860 ± 0.007**
SL comp.	0.85 ± 0.01	**0.913 ± 0.007**	0.845 ± 0.006	**0.913 ± 0.007**
SL purity	0.51 ± 0.01	0.75 ± 0.01	0.62 ± 0.01	**0.84 ± 0.02**
Non-SN comp.	0.05 ± 0.02	**0.07 ± 0.02**	0.04 ± 0.02	0.04 ± 0.02
Non-SN purity	0.04 ± 0.01	0.05 ± 0.02	**0.15 ± 0.08**	0.13 ± 0.08
$F_{0.5}$-score	0.814 ± 0.004	0.876 ± 0.003	0.866 ± 0.003	**0.920 ± 0.002**

improvement. Compared to the needs of live classification, it is less clear if this small improvement in purity compensates for the significant fraction of the sample discarded before classification. In fact, the DASH–NGSF combined classification produces a higher SN Ia purity and classifies a greater number of SNe Ia in total (since the completeness of the 21.8 mag cut NGSF classification is around 50 per cent when cut transients are accounted for).

5.1.2 Transient flux fraction

After transient magnitude, the second obvious source of classification error in our sample comes from high levels of host galaxy flux in our spectra. In this section, we discuss the effectiveness of DASH, NGSF, and SNID as a function of transient flux fraction (contrast), where the transient flux fraction is the fraction of the flux in a 4MOST fibre that originates from the transient. We report the potential to improve classification results by introducing a sample cut in transient flux fraction-redshift space. We investigate using our 5-class classification schema as in previous sections.

Generally, the trends in classification rates against the transient flux fraction are as one would expect. As the transient flux fraction increases (the spectrum's host contamination is reduced), we see improvements in the SN Ia completeness and purity. The shape of these plots is very similar to those produced by transient magnitude binning in Fig. 9. The purity tends to approach 95 per cent at transient flux fractions of 40–50 per cent if it is not already above that in the most contaminated bin. Fig. 10 indicates that all three classifiers have similar slopes in their purity with different initial values. Although not shown in the figure, the same trend was found without redshift priors, albeit with slightly smaller values for DASH and much smaller values for NGSF and SNID.

We look at our results in flux fraction-redshift space in Fig. 11. At high redshift only, transients that have bright absolute magnitudes, especially transients in the SLSN class, will be observed. So transient flux fraction is likely to be high as we are biased to intrinsically brighter transients while host brightness remains constant. However, we also expect the spectral features of our transients to be shifted outside of 4MOST's wavelength range, making them harder to classify. Indeed the *rlap* classification quality parameter employed by DASH and SNID depends directly on the wavelength overlap between the input spectrum and matching template. We hope to find regions of this parameter space without contaminants or fewer misclassifications, where we could assign positive results a greater degree of certainty.

A few obvious points of interest are the trend to greater transient flux fractions with increasing redshift and the incidence of unsuccessful classifications of SNe Ia (orange histograms) beginning to drop off as the transient flux fraction surpasses around 40 per cent. The SN Ia count histograms are fairly uniform for the three classifiers in the relative distributions of the successful and unsuccessful SN Ia classifications, but we see variation in the width of the successful classification histogram. In particular, there are obvious differences in the number of misclassified SNe Ia between the classifiers.

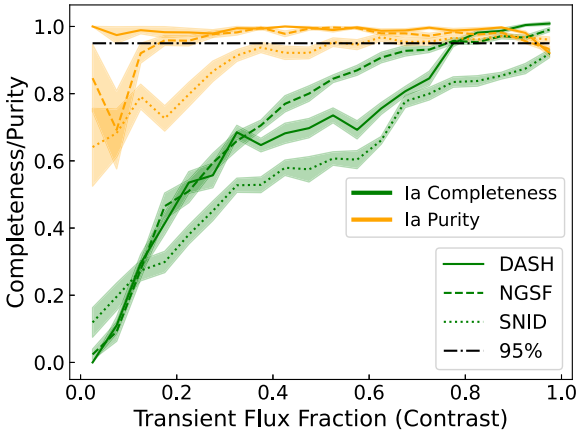


Figure 10. The SN Ia completeness (green, lower lines) and purity (orange, upper lines) as a function of the fraction of the total flux in the spectrum that originates from the transient. The SNe Ia in each of our subsamples are in 20 linear bins between transient fibre flux fractions (contrast) of 0 and 1. Redshift is known in all cases. Uncertainty in indicated by the shaded regions. Shaded regions are defined by the standard error on the mean in each bin between our random subsamples. All three classifiers produce similar trends in SN Ia completeness and purity. In every case, the classification completeness and purity improve as the transient flux fraction increases.

Also concerning are the clusters of SLSNe at high redshift ($z > 0.6$) that are classified as SNe Ia in all three classifiers, although most prevalently in DASH and NGSF. These SLSNe are being fit overwhelmingly as SNe Ia-91bg. This does lead to a potential mechanism for increasing purity. As can be seen in Fig. 11, the successful SN Ia classifications (and indeed instances of SNe Ia in general) drop off quite sharply after $z = 0.60$. Each classifier has contaminants beyond this redshift that could be dismissed out of hand if accurate spectroscopic redshifts for host galaxies are known, or if photometric redshifts indicate it is likely that $z > 0.60$.

For now, with the precise extent to which TiDES will have host redshift information, we do not implement such a cut. However, we make note of it and strongly encourage such a cut's usage in the cases where redshifts are known.

An obvious location for a cut on the transient flux fraction is the point at which the good SN Ia classifications begin to dominate over misclassifications. This occurs at a transient flux fraction of roughly 0.2 for DASH, 0.2 for NGSF, and 0.3 for SNID, we generalize this to a cut at a flux fraction of 0.3.

A second tempting cut is on very large transient flux fractions, greater than 0.9. In DASH and NGSF, there are clusters of very bright, high flux fraction, SLSNe being falsely classified as SN Ia. However, we choose not to pursue this cut, simply because removing SNe in these bins would also remove the regions with the highest density of correct classifications.

In Table 6, we present the results of our 5-class classification schema for NGSF as we employ a variety of different photometric cuts to the input sample. We see that using only a cut for transient flux fractions greater than 0.3 returns similar classification results across most transient classes to the 21.8 transient magnitude cut employed in Section 5.1.1. The Ia classification performance is nearly identical, with the other classes best performances spread fairly evenly. Using both cuts results in even better performance, indeed it produces the

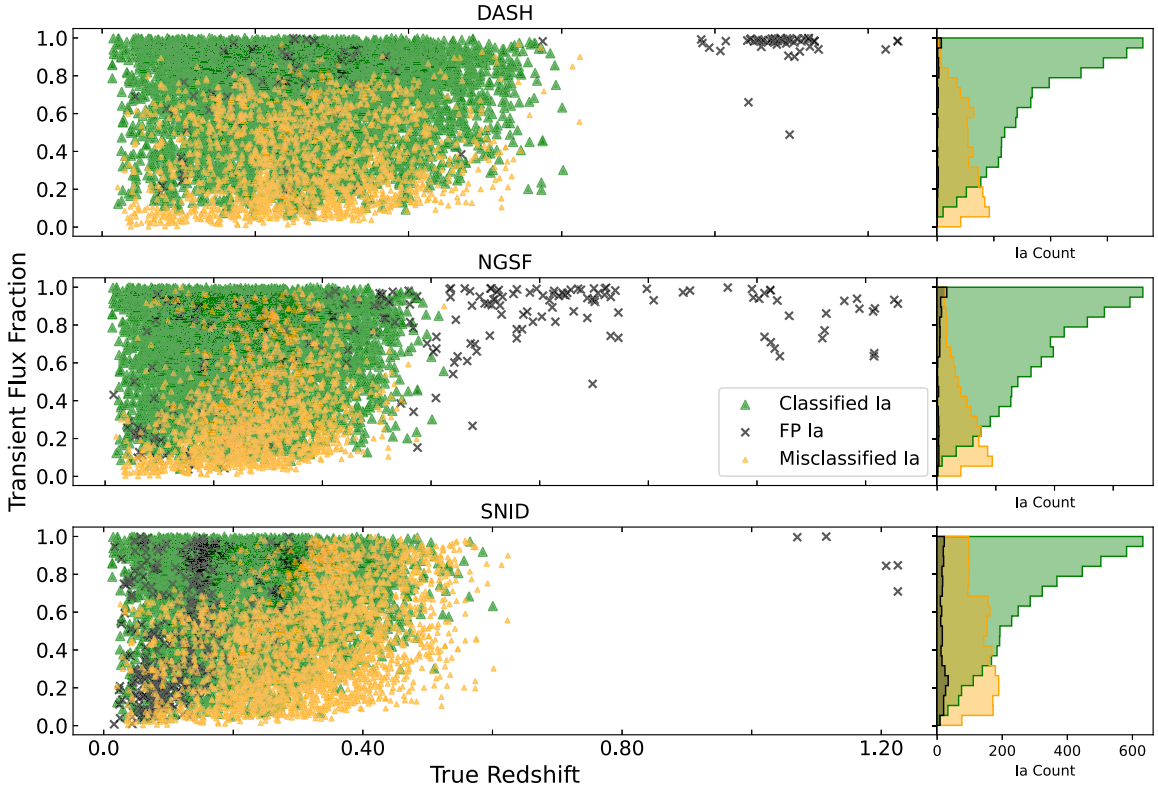


Figure 11. The classification results in the binary schema with known redshifts for all three classifiers in transient flux fraction-redshift space. Large green and small orange triangles indicate good SN Ia classifications and failed SN Ia classifications, respectively. The black crosses indicate SN Ia false positives (that is a non-SN Ia classified as an SN Ia.) The histograms show the corresponding counts with the same colour scheme. There are regions of the parameter space for each classifier where false positive SN Ia classifications cluster, often at high redshifts ($z > 0.6$). We also see similar distributions for successful and unsuccessful SN Ia classifications.

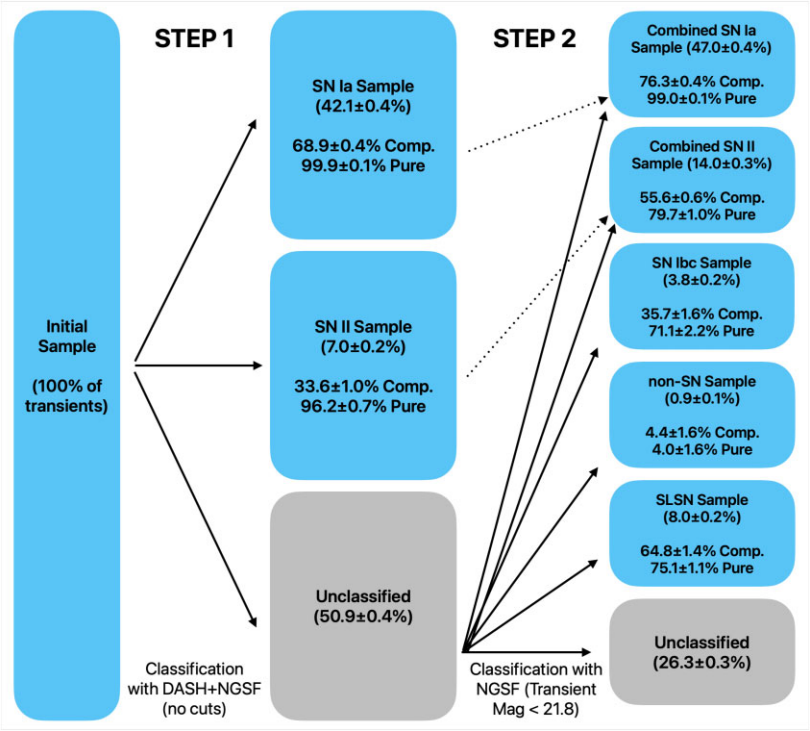


Figure 12. An example of a classification pipeline that could be employed by TiDES for the purpose of live classification of transients. The output samples of from each step in the classification pipeline are provided with their completeness and purities labelled. The samples of SNe Ia and SNe II provided after the second classification step represent the combination of the transients from the first classification and those from the second. Percentages of the total sample size are listed in brackets for each classes final sample. Classifications are performed with redshift information.

largest $F_{0.5}$-score, followed by the magnitude cut and then the flux fraction cut. However, these performance benefits must be weighed against the large fractions of the sample removed from consideration and thus not reflected in the $F_{0.5}$-score.

We conclude cautiously that the best photometric cut for live classification is likely to be transient transient magnitude $r > 21.8$, the middle ground between improved performance and reduction in sample size. Although arguments can be made for the flux fraction cut or both. In all three cases, the non-SN transient completeness and purities are very poor. This is the result of low numbers (or a complete absence) of templates in the template banks/training samples and, additionally, the fact that non-SN input spectra are just smooth-blue continua (see Appendix C).

5.2 An example classification plan

In this section, we propose just one possible scheme that could be employed by TiDES for live classification of transients. The pipeline is illustrated in Fig. 12 and assumes redshift information is provided for all classifications. The pipeline consists of two separate classifications of the sample of transients. First, the full sample is classified by the combined DASH–NGSF classifier recommended in Section 5. This produces very pure samples of SNe Ia and SNe II although, particularly for the latter, the completeness is low. The SNe Ia sample produced by this first classification step has 99.9 per cent purity and should be appropriate for use in cosmology.

From the sample of spectra not classified by the combined classifier, we now take only those with a transient magnitude brighter than 21.8 mag as discussed in Section 5.1.1. These bright objects are then reclassified with just NGSF. This produces reasonably pure and complete samples of SNe Ibc and SLSNe. It also classifies a few additional SNe Ia and SNe II which can be combined with the existing samples to increase their completeness at the cost of their purities. The only class with poor results is the non-SN transients. Here, we only classify 4 per cent correctly and over 95 per cent of the resulting sample is contamination from other classes. This is an issue with NGSF's template bank and the absence of such spectra from DASH's training set. When considered in full, the classification pipeline leaves just over a quarter of transients unclassified.

This is a reasonably successful classification. It outperforms any individual spectroscopic classifier that we have tested in this work in regards to purity. This classification scheme obtains a very pure SNe Ia sample for cosmology in addition to producing classification completeness and purities in non-SN Ia classes that are suitable for live transient classification. See Fig. 12 for the completeness and purity of each class after each step of the classification pipeline.

We note that this classification pipeline has a higher $F_{0.5}$-score than NGSF. However, the choice of β in equation (8) allows for greater importance to be placed on completeness rather than purity. The F-scores for several values of β across several classification schemes are presented in Table 7. We can see that while NGSF individually performs best in F_1- and F_2-scores, when the score is weighted to favour completeness ($\beta > 1$, the various versions of the classification pipeline presented in this section have the highest score when $\beta = 0.5$ and purity is weighted more heavily. In fact, at even lower values of $\beta \leq 0.1$, the combined DASH–NGSF classifier would have the best score. As a result, it is hard to objectively state the superior classifier, it will depend on the objectives of a particular study.

Fortunately, there is significant room for fine-tuning to specific science cases. For example, replacing the cut on transient magnitude to the cut on transient flux fraction as discussed in Section 5.1.2, the pipeline will produce samples with higher completeness at the

Table 7. The $F_{0.5}$-, F_1-, and F_2-scores of several classifiers mentioned throughout this paper. Each choice of β indicates a different priority in the classifier. Smaller β-values increasingly weight the F-score towards good purity results, while increasingly large values instead weight in favour of completeness. β-values of 0.5 and 2 and used by convention. The largest value(s) in each column are in bold.

Classifier	$F_{0.5}$	F_1	F_2
Pipeline: Mag. cut	**0.830 ± 0.002**	0.757 ± 0.002	0.698 ± 0.003
Pipeline: Flux frac. cut	**0.831 ± 0.003**	0.773 ± 0.003	0.726 ± 0.003
DASH–NGSF only	0.757 ± 0.004	0.645 ± 0.005	0.566 ± 0.004
NGSF only	0.814 ± 0.004	**0.786 ± 0.004**	**0.765 ± 0.004**

cost of purity. Additionally, the percentage of unclassified objects drops to just 18 per cent. In this case, the SLSN purity drops to around 65 per cent, but this is compensated by an completeness of over 80 per cent.

Additional cuts from photometric information can be added to either stage of the pipeline to increase purity at the cost of completeness. Different cuts than those discussed here can be used, which will affect each class differently, allowing for parties interested in specific SNe classes to be specific in their classification.

The final advantage of such a classification model is that it is versatile and easily communicated to the community. By providing only the class from the 5-class output probabilities from each classifier, the r-band magnitude of the transient and host near time of observation, and the redshift of the system, it would be possible for members of the community to adjust the transient sample selected to suit their specific science requirements by varying classifiers or probability thresholds.

5.2.1 *Comparison to photometric classification results*

In this subsection, we compare three recent photometric classification papers surrounding a recent photometric classifier and its use with the DES (Möller et al. 2022).

Möller & de Boissière (2020) present the photometric transient classifier SUPERNNOVA classifying simulated light curves with spectroscopic redshift information and incomplete light curve information. Additionally, Möller et al. (2022, 2024) present SUPERNNOVA classification results on real light curves with and without host redshifts, respectively.

Specifically, Möller et al. (2024) present the binary classification of DES 5-yr data release SNe without any redshift information provided as a prior. When the light curves of transients being fit without redshifts are trimmed to only include photometry up to peak brightness, SUPERNNOVA produces a binary accuracy, a Ia completeness, and a Ia purity of 90.46 per cent, 92.49 per cent, and 91.93 per cent, respectively. By comparison, if operated as a binary classifier without redshift, our classification plan from Section 5.2 produces a binary accuracy, a Ia completeness and a Ia purity of 85.6 ± 0.4 per cent, 44.5 ± 0.6 per cent, and 94.4 ± 0.3 per cent. Additionally, we can consider only the high-confidence SN Ia sample produced by the combined NGSF–DASH classifier to improve the SN Ia purity to 99.5 ± 0.1 per cent at the cost of reducing completeness to just 36.4 ± 0.6 per cent.

As seen in Table 3, NGSF has significant performance loss when redshift information is not provided. As such, the binary accuracy, SN Ia completeness and purity can be improved to 91.4 ± 0.4 per cent, 55.6 ± 0.6 per cent, and 95.3 ± 0.3 per cent by replacing the DASH–NGSF classification step with an equivalent DASH–SNID classification. However, this does come at the cost of worse performance in the 5-class mode of operation.

Möller et al. (2022) also apply SUPERNNOVA to the photometric sample produced by the DES 5-yr data release. This produces a cosmologically useful sample of 1484 SNe Ia with spectroscopic redshifts. The predicted completeness and purity of the sample are 98.51 per cent and 97.73 per cent, respectively. Again, we consider both the high-confidence SN Ia sample and the larger, less confident, SN Ia sample produced by our classification pipeline. Now with redshift priors, the less confident sample has an completeness of 76.3 ± 0.4 per cent and purity of 99.0 ± 0.1 per cent. We can sacrifice some completeness to improve purity and use the high confidence SN Ia sample produced by the combined DASH–NGSF classifier. This increases purity to >99.9 per cent with completeness just under 70 per cent. Our classification plan produces an SNe Ia sample with a percentage contamination that is more than a factor of 10 lower, at the cost of lower completeness and accuracy, than SUPERNNOVA. This is true whether redshift information is available or not.

While most photometric classifiers function purely in a binary (SN Ia versus non-SN Ia) schema and with complete light curves, in Möller & de Boissière (2020), SUPERNNOVA reports results using ternary and seven-way classification schema, similar to our 5-class schema.

SUPERNNOVA reports an accuracy of 77.8 per cent for its ternary schema (SNe Ia, Ibc, and II) and 64.2 per cent for the seven-way classification schema (SNe Ia, IIP, IIn, IIL1, IIL2, Ib, and Ic). In each case, these are the accuracies expected from light curves consisting, on average, of 2.4 distinct nights of multicolour observations up to 2 d before peak brightness. These percentages improve to 81.5 per cent and 69.8 per cent for an average of 3.1 distinct nights of multicolour observations up to 2 d after peak brightness. All classifications also make use of spectroscopic redshifts.

For comparison our example pipeline, in the 5-class schema (SNe Ia, Ibc, II, SL, and non-SNe), produces a comparable classification accuracy of 90.1 ± 0.2 per cent. Additionally, if we consider only SNe Ia, Ibc, and II to mimic the ternary schema, we obtain an accuracy of 93.2 ± 0.3 per cent. In both cases, we do not consider unclassified spectra in our calculation of the accuracy. In the ternary scheme, non-SN transient and SLSN outputs are considered unclassified.

From Frohmaier et al. (2025), the requirements to flag a transient for spectroscopic follow-up are three $griz$ detections in two distinct nights, with the added requirement that at least one of these detections be brighter than 22.5 mag. We also assume spectroscopic redshifts are available. Our use of spectroscopy produces a roughly 15 per cent improvement on the accuracies from photometry with similarly incomplete light curves.

6 CONCLUSIONS

In this paper, we set out to determine whether the classification of transients discovered by 4MOST–TiDES can be automated using one or more spectroscopic transient classifiers. We want to know which classifier(s) are the best from a live classification and cosmological point of view. To do this, we simulated realistic blended spectra using pre-existing simulations and the 4MOST ETC and classified them using DASH, NGSF, and SNID. We place a focus on classification purity due to the large sample sizes produced by TiDES, and employ the $F_{0.5}$-score as our purity-weighted FoM.

The classification performances of DASH, NGSF, and SNID are weaker than those reported in their original papers. This is the result of different quality data and fainter SNe, alongside significant host contamination. We find that, individually, NGSF produces the best $F_{0.5}$-score for known redshift classifications, although its performance loss is across all transient classes large if redshift information cannot be provided. None of the individual classifiers were robust enough to recommend their use for automated classification.

We find that the purities in SNe Ia can be greatly improved by using several classifiers at once and requiring an agreement between them on each classification. This is costly for transient completeness, but with the benefit of having vastly reduced contamination in the output sample. We get good results from a combination of DASH and NGSF, with SNe Ia completeness of 69.4 ± 0.5 per cent and purity of 99.94 ± 0.03. Purity can be marginally improved by including SNID in the combined classifier, but at the cost of a much reduced completeness.

This allows for the automation of SNe Ia classification and the production of good cosmology samples. However, it alone does not lead to a solution for general automated classification for TiDES. The combined DASH–NGSF classifier struggles to classify SNe Ibc and SNe II with a high completeness, although what it does classify is quite pure. It is incapable of classifying SLSNe and non-SN transients, as DASH, by default, has not been trained to classify them.

We investigated a variety of photometric cuts that could be applied to our data to improve the resulting transient classifications for individual classifiers. We found that only classifying transients with r-band magnitudes brighter than 21.8 could significantly improve classification purity across all transient classes, but at the cost of classification completeness. Similar results can be obtained by only classifying objects for which SNe flux comprises more than 30 per cent of the flux within the observing 4MOST fibre.

We present an example classification plan in Section 5.2. We emphasize that such a classification pipeline is easily fine-tuned to specific science cases and conclude it is viable for live automated classification and these modifications require only the classifier outputs and some photometric information to be performed. The specific classification pipeline present in this paper outperforms the $F_{0.5}$-scores of all combinations of one, two, or three classifiers. In Table 7, we indicate how one might choose a different classifier than our pipeline depending on whether the purity of the sample or the completeness is considered most important for particular research goals.

We have demonstrated the capacity of an example classification pipeline to produce a very high purity SN Ia sample at the cost of completeness, and a sample with far higher completeness with lower purity. A future step in this work will be to optimize the classification scheme via end-to-end cosmological simulations, in order to show which sample best constrains the cosmology and which combination of classifiers and photometric cuts minimize the uncertainty on derived cosmological parameters.

Importantly, it is currently unclear to what extent 4MOST–TiDES will be able to obtain redshift information from host galaxies to be used in transient classification. The change in completeness and purities is significant between known and unknown redshifts and represents perhaps the largest uncertainty in the results of this paper. Work is currently underway investigating how consistently a redshift can be derived from features in blended host–transient spectra. Even in the case that live spectroscopic redshifts cannot be obtained from hosts, we are optimistic that it will be possible to obtain some host redshifts from legacy surveys such as the Dark Energy Spectroscopic Instrument (DESI) survey (Dey et al. 2019) and Sloan Digital Sky Survey (SDSS) (York et al. 2000; Wolf et al. 2016; Almeida et al. 2023). Host photo-zs also present a promising middle-ground between spectroscopic and unknown redshifts.

Finally, it is likely to be possible to bolster the spectroscopically confirmed transient samples with photometrically classified transients once full light-curve data are produced by LSST.

ACKNOWLEDGEMENTS

AM gratefully acknowledges support from a Science and Technology Funding Council (STFC) PhD studentship and the Faculty of Science and Technology at Lancaster University. IH gratefully acknowledges support from the Leverhulme Trust [International Fellowship IF-2023-027] and the Science and Technologies Facilities Council [grants ST/V000713/1 and ST/Y001230/1]. Y-LK has received funding from the Science and Technology Facilities Council [grant no. ST/V000713/1] and was supported by the Lee Wonchul Fellowship, funded through the BK21 Fostering Outstanding Universities for Research (FOUR) Program (grant no. 4120200513819) and the National Research Foundation of Korea to the Center for Galaxy Evolution Research (RS-2022-NR070872 and RS-2022-NR070525). AM is supported by the ARC Discovery Early Career Researcher Award (DECRA) project no. DE230100055. ET was supported by the Estonian Ministry of Education and Research (grant TK202), Estonian Research Council grant (PRG1006), and the European Union's Horizon Europe research and innovation programme (EXCOSM, grant no. 101159513). KM is funded by Horizon Europe ERC grant no. 101125877. PW acknowledges support from the Science and Technology Facilities Council (STFC) grants ST/R000506/1 and ST/Z510269/1. RD gratefully acknowledges support by the Agencia Nacionel de Investigación y Desarrollo (ANID) BASAL project FB210003. MN is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 948381) and by UK Space Agency grant no. ST/Y000692/1. The authors thank the anonymous reviewer for comprehensive comments that led to a much improved paper. The authors thank the anonymous referee for their useful comments.

DATA AVAILABILITY

The set of SNID templates used throughout can be made available on request. Additionally, the full set of blended spectra used throughout are to be made available through a public repository on Lancaster University's PURE research information system.

REFERENCES

Almeida A. et al., 2023, ApJS, 267, 44
Aramyan L. S. et al., 2016, MNRAS, 459, 3130
Astier P. et al., 2006, A&A, 447, 31
Balland C. et al., 2009, A&A, 507, 85
Blondin S., Tonry J. L., 2007, ApJ, 666, 1024
Blondin S., Tonry J. L., 2011, Astrophysics Source Code Library, record ascl:1107.001
Boone K., 2019, AJ, 158, 257
Boone K., 2021, AJ, 162, 275
Cabrera-Vives G. et al., 2024, A&A, 689, A289
Campbell H. et al., 2013, ApJ, 763, 88
Capaccioli M., ed., 1989, The World of Galaxies. Springer, NY
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Charnock T., Moss A., 2017, ApJ, 837, L28
de Jong R. S. et al., 2019, The Messenger, 175, 3
de Soto K. M. et al., 2024, ApJ, 974, 169

DES Collaboration, 2024, ApJ, 973, L14
Dey A. et al., 2019, AJ, 157, 168
Dong Y. et al., 2022, ApJ, 927, 199
Drout M. R. et al., 2014, ApJ, 794, 23
Duan F.-Q., Liu R., Guo P., Zhou M.-Q., Wu F.-C., 2009, Res. Astron. Astrophys., 9, 341
Filippenko A. V., 1997, ARA&A, 35, 309
Fraga B. M. O. et al., 2024, A&A, 692, A208
Fremling C. et al., 2021, ApJ, 917, L2
Frohmaier C. et al., 2025, preprint (arXiv:2501.16311)
Gagliano A., Contardo G., Foreman-Mackey D., Malz A., Aleo P., 2023, in American Astronomical Society Meeting Abstracts, p. 103.02
Galbany L. et al., 2018, ApJ, 855, 107
Ginolin M. et al., 2025, A&A, 695, A140
Goldwasser S., Yaron O., Sass A., Irani I., Gal-Yam A., Howell D. A., 2022, Transient Name Server AstroNote, 191, 1
Gomez S., Villar V. A., Berger E., Gezari S., van Velzen S., Nicholl M., Blanchard P. K., Alexander K. D., 2023, ApJ, 949, 113
Graham A. W., Driver S. P., 2005, Publ. Astron. Soc. Aust., 22, 118
Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2018, AJ, 155, 1
Graur O., 2024, MNRAS, 530, 4950
Guy J. et al., 2007, A&A, 466, 11
Guy J. et al., 2010, A&A, 523, A7
Hakobyan A. A., Adibekyan V. Z., Aramyan L. S., Petrosian A. R., Gomes J. M., Mamon G. A., Kunth D., Turatto M., 2012, A&A, 544, A81
Hakobyan A. A. et al., 2016, MNRAS, 456, 2848
Harutyunyan A. H. et al., 2008, A&A, 488, 383
Hills J. G., 1975, Nature, 254, 295
Hounsell R. et al., 2018, ApJ, 867, 23
Howell D. A. et al., 2005, ApJ, 634, 1190
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jones D. O. et al., 2018, ApJ, 857, 51
Jones D. O. et al., 2019, ApJ, 881, 19
Kessler R. et al., 2009, PASP, 121, 1028
Kessler R. et al., 2019, PASP, 131, 094501
Kim Y. L. et al., 2022, PASP, 134, 024505
Kim Y.-L. et al., 2024, PASP, 136, 114501
Kinney A. L., Calzetti D., Bohlin R. C., McQuade K., Storchi-Bergmann T., Schmitt H. R., 1996, ApJ, 467, 38
Kisley M., Qin Y.-J., Zabludoff A., Barnard K., Ko C.-L., 2023, ApJ, 942, 29
Leloudas G. et al., 2015, MNRAS, 449, 917
Lidman C. et al., 2020, MNRAS, 496, 19
Mannucci F., Basile F., Poggianti B. M., Cimatti A., Daddi E., Pozzetti L., Vanzi L., 2001, MNRAS, 326, 745
Minkowski R., 1979, A Source Book in Astronomy and Astrophysics, 1900-1975. Harvard Univ. Press, Cambridge, Massachusetts and London
Mitra A., Kessler R., More S., Hlozek R., LSST Dark Energy Science Collaboration, 2023, ApJ, 944, 212
Möller A., de Boissière T., 2020, MNRAS, 491, 4277
Möller A. et al., 2022, MNRAS, 514, 5159
Möller A. et al., 2024, MNRAS, 533, 2073
Muthukrishna D., Narayan G., Mandel K. S., Biswas R., Hložek R., 2019a, PASP, 131, 118002
Muthukrishna D., Parkinson D., Tucker B. E., 2019b, ApJ, 885, 85
Narayan G., ELAsTiCC Team, 2023, in American Astronomical Society Meeting Abstracts, p. 117.01
Neill J. D. et al., 2011, ApJ, 727, 15
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Perlmutter S. et al., 1999, ApJ, 517, 565
Pimentel Ó., Estévez P. A., Förster F., 2023, AJ, 165, 18
Planck Collaboration I, 2014, A&A, 571, A1
Popper D. M., 1937, PASP, 49, 283
Prugniel P., Simien F., 1997, A&A, 321, 111
Qu H., Sako M., 2022, AJ, 163, 57
Riess A. G. et al., 1998, AJ, 116, 1009
Rigault M. et al., 2025, A&A, 694, A1
Sako M. et al., 2018, PASP, 130, 064002
Saunders W. et al., 2004, in Moorwood A. F. M., Iye M., eds, Proc. SPIE Conf. Ser. Vol. 5492, Ground-based Instrumentation for Astronomy. SPIE, Bellingham, p. 389
Savitzky A., Golay M. J. E., 1964, Anal. Chem., 36, 1627
Schlegel E. M., 1990, MNRAS, 244, 269
Sérsic J. L., 1963, Bol. Asoc. Argentina de Astron. La Plata Argentina, 6, 41
Shah V. G., Gagliano A., Malanchev K., Narayan G., The LSST Dark Energy Science Collaboration, 2025, preprint (arXiv:2501.01496)
Sharma Y. et al., 2025, PASP, 137, 034507
Sheng X. et al., 2024, MNRAS, 531, 2474
Sullivan M. et al., 2006, ApJ, 648, 868
Swann E. et al., 2019, The Messenger, 175, 58
Tempel E. et al., 2020a, MNRAS, 497, 4626
Tempel E. et al., 2020b, A&A, 635, A101
The LSST Dark Energy Science Collaboration, 2018, preprint (arXiv:1809.01669)
Van Rijsbergen C. J., 1977, J. Doc., 33, 106
Vincenzi M., Sullivan M., Firth R. E., Gutiérrez C. P., Frohmaier C., Smith M., Angus C., Nichol R. C., 2019, MNRAS, 489, 5802
Vincenzi M. et al., 2021, MNRAS, 505, 2819
Vincenzi M. et al., 2024, ApJ, 975, 86
Vogl C., Kerzendorf W. E., Sim S. A., Noebauer U. M., Lietzau S., Hillebrandt W., 2020, A&A, 633, A88
Wang M., Ma Y., Wu Q., Jiang N., 2024, ApJ, 960, 69
Wittman D. M., Tyson J. A., Kirkman D., Dell'Antonio I., Bernstein G., 2000, Nature, 405, 143
Wolf R. C. et al., 2016, ApJ, 821, 115
Yaron O., Gal-Yam A., 2012, PASP, 124, 668
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: SNE IA FITS AND CONTAMINANT ORIGINS

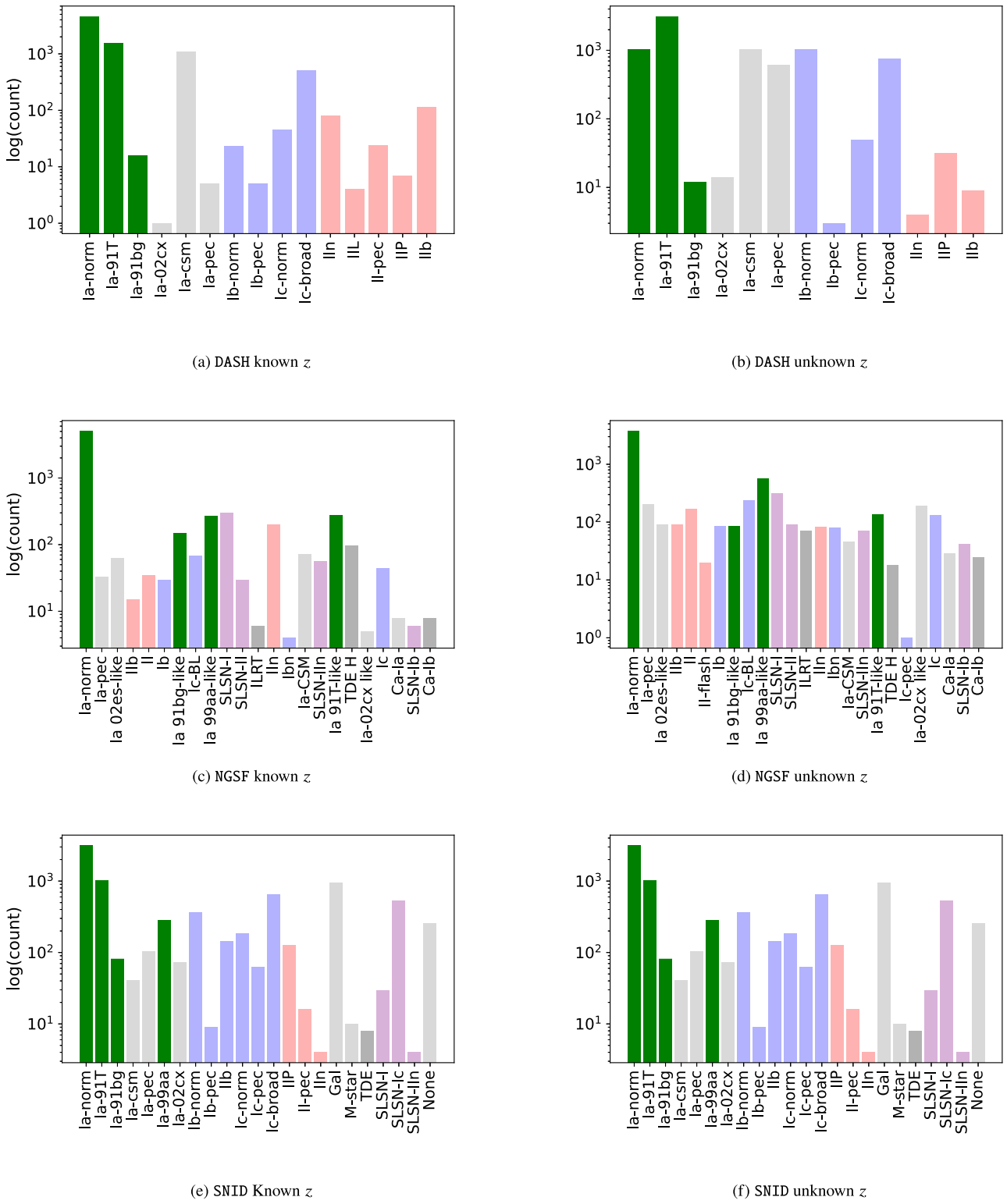
Fig. A1 shows how the SN Ia input spectra are being fit by each classifier, based on the subclass of the best-fitting template. In each case, the green bars indicate the good SN Ia classification bins. Non-SN Ia bars of various colours indicate all of the misclassifications. In all three classifiers, we investigate we see the same effects of moving from using redshift priors to not.

There is a shift in successfully classified SNe Ia from the Ia-norm class into other SN Ia and SN Ia-pec subclasses. Additionally, the number of SNe Ia incorrectly classified as non-SNe Ia can be seen in the non-green bars universally increasing in height. Both of these effects serve to diminish the SN Ia classification rate without redshifts.

Of note are the tendency of DASH to classify transients as SN Ia-csm. This seems to be the result of narrow galaxy emission lines from Sc host templates masquerading as the narrow lines of an ejecta-csm interaction. The inclusion of SN Ia-csm as an acceptable SN Ia class for DASH does improve the SN Ia classification rate, but at the cost of contamination rates exceeding 15 per cent. A similar effect occurs with SNID, except that it does seem to prefer to correctly identify them as galaxies with a 'Gal' output.

Fig. A2 shows the origin of the contaminant results for each classifier. We can see immediately that DASH suffers as a result of having no ability to classify SLSNe, as they make up the largest fraction of contaminants when redshift priors are known.

When redshift information is removed, DASH loses classification performance for all transient classes in both completeness and purity. The fractional decrease in the number of SN Ia and contaminant classification is almost exactly the same, and this results in the purity remaining high (see Table 3). The input template classes that produce contaminants is entirely different when redshift priors are removed, now being almost entirely from SNe II. For SLSNe, forcing the



(a) DASH known z

(b) DASH unknown z

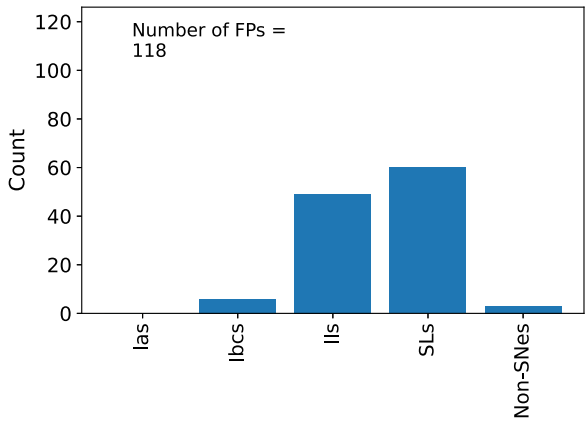
(c) NGSF known z

(d) NGSF unknown z

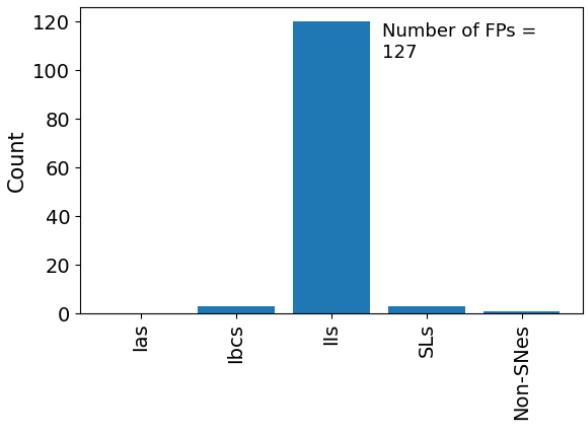
(e) SNID Known z

(f) SNID unknown z

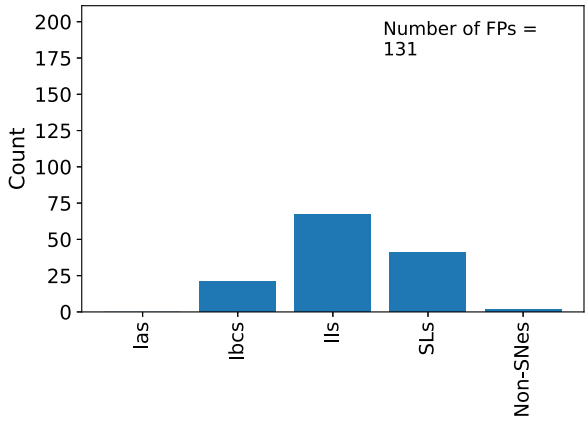
Figure A1. Graphical representation of how SN Ia input spectra are being classified by each classifier with (left column) and without (right column) redshift priors. The subclass of the best-fitting templates is assumed as the subclass of the output. Each histogram lists only the subclasses with at least one output classification. SN Ia subclasses are green, Ibc are blue, II are red, SLSNe are purple, non-SNe are black, and 'other' classes (Ia-pec, non-transients) are grey (see also the subclass names on the x-axis). The shift from Ia-norm to other SNe Ia subclasses when redshift priors are removed can be seen.



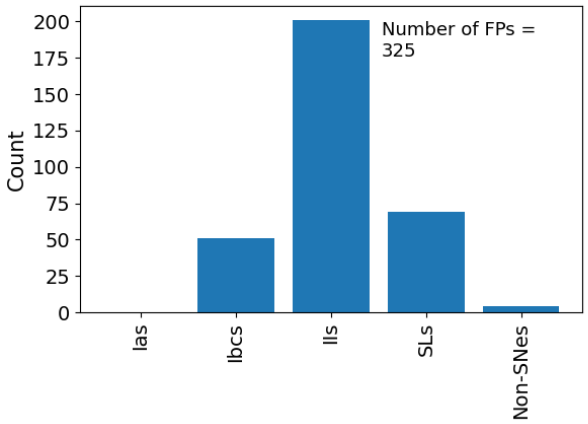
(a) DASH known z



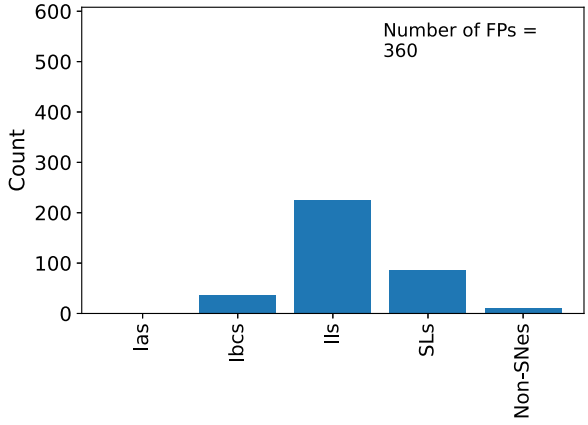
(b) DASH unknown z



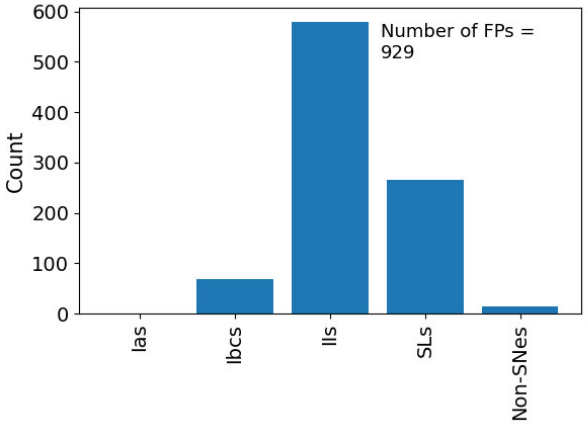
(c) NGSF known z



(d) NGSF unknown z



(e) SNID known z



(f) SNID unknown z

Figure A2. The distribution of true classifications for objects classified as Ia above the quality threshold to qualify as contaminant results. Input classes are those from the 5-class classification schema. The number of contaminants for each classifier-redshift prior combination are listed on each subplot. The number of FPs increases significantly without redshift priors for NGSF and SNID. SLSNe are often over-represented as FPs.

classification to high redshifts by using priors resulted in many contaminant Ia classifications. When redshift priors are removed, SLSNe are instead misclassified as other non-SN Ia transients or as SNe Ia-pec. This is a good change from the point of view of SN Ia sample purity.

While we see the contaminant numbers produced by DASH maintained when removing redshift knowledge, NGSF and SNID both produce double or more contaminant SN Ia classifications. NGSF and DASH both classify predominantly SNe II as contaminant SNe Ia when redshift priors are removed, a significant change from the ratio of classes that produce contaminants with redshift priors. SNID's distribution of contaminants remains almost identical between regimes, although again SNe II are the largest contributor.

Type II SNe are the largest non-SN Ia component of the sample and as expected always dominate the contaminant distribution. In fact, in nearly all cases, the relative number of contaminants originating from the different input non-SN Ia classes at least vaguely mimics their relative abundance in the full sample, slightly shifted by each classifier's ability to classify different classes. Only Fig. A2(b) bucks this trend, producing a large overabundance of SN II contaminant classifications.

APPENDIX B: EXAMPLE CLASSIFICATION

In this appendix we provide some individual classifications as context. We focus on several of the most common types of classification and misclassification. All presented classifications are from NGSF as it is the most prevalent in our suggested classification plan in Section 5.2.

Fig. B1 shows four attempted classifications with NGSF. Fig. B1(a) shows a successful SN Ia classification. We find that noisy spectra, where the transient is faint, or spectra with significant host contamination are often hard to classify as would be expected. This is shown in Fig. B1(b). We also see an overabundance of misclassifications from spectra with the Sc host template. These are often the result of the classifier misinterpreting the strong galaxy emission as narrow features from the transient. This leads to a classifications of SN Ia-csm and other narrow emission transient subclasses like Ibn, IIn, etc. This is shown in Fig. B1(c). False positive SN Ia classifications can arise from many effects. Shown in Fig. B1(d), we have a low host contamination SN Ib being misinterpreted as a Ia-norm with significant host contamination. This suggests that there is degeneracy between SN subclass and host contamination levels.

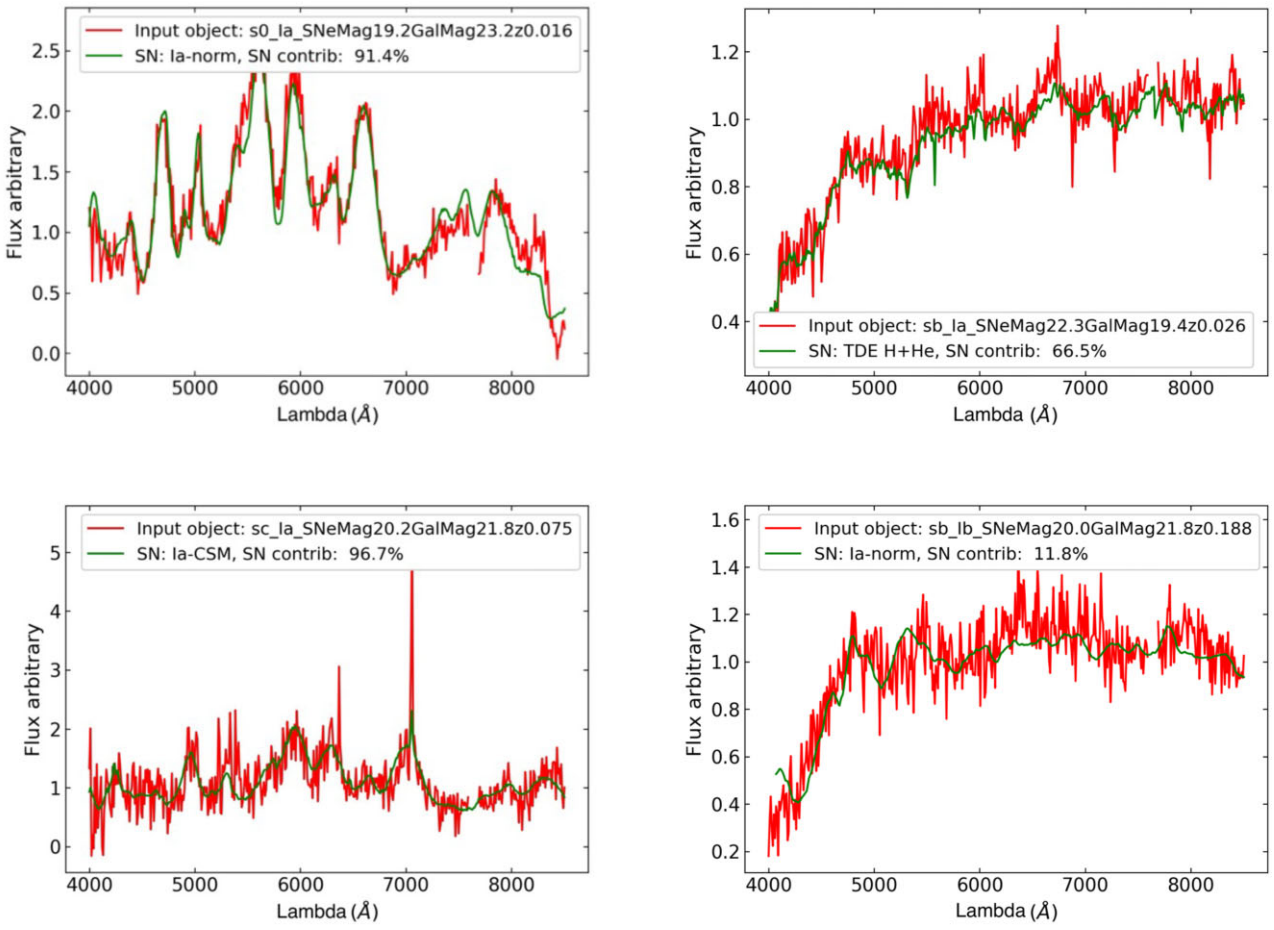


Figure B1. Four individual classification results from NGSF. Top left: a good classification of a bright, low contamination SN Ia. Top right: a misclassification of a highly contaminated SN Ia. Bottom left: a misclassification of a bright SN Ia due to narrow galaxy features from its Sc host. Bottom right: an example of an SN Ia false positive where a low contamination SN Ib is misinterpreted as an SN Ia with high contamination. In each case, the input is plotted in red (noisy) with relevant information in the legend. The best-fitting template spectrum is plotted in green and the best-fitting transient class is provided in the legend. The host galaxy fraction of NGSF's best-fitting template is included in the legend with the best fit.

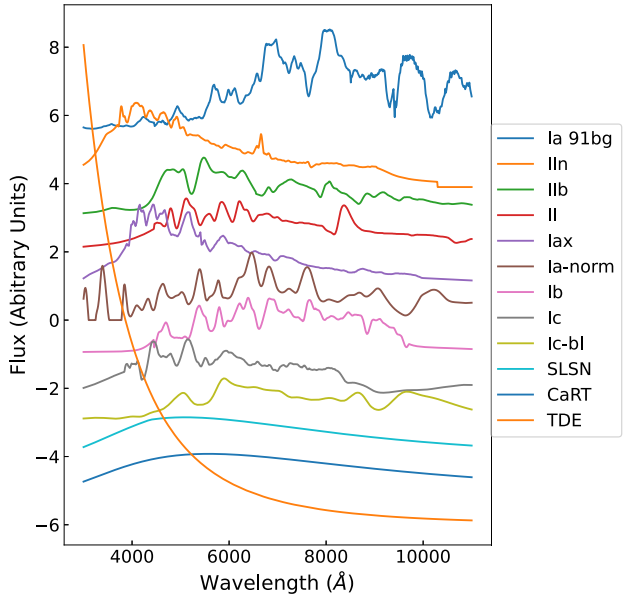


Figure C1. Example spectra for each distinct input transient class. Spectra such as these were used as the starting point to generate the simulated spectra in Section 3.

APPENDIX C: EXAMPLE SPECTRA

Fig. C1 shows an example of each of the twelve types of inputs transient spectra used in our blended spectra simulations. The spectra belong to the transient classes of: Ia-norm, Ia 91bg-like (faint, fast-declining), Iax (faint, progenitor-preserving white dwarf thermonuclear detonations), Ib, Ic, IIb, Ic-BL, II, IIn (all core-collapse SNe), SLSNe (incredibly bright transients), TDEs (star disrupted by black hole tidal forces), and CaRTs (SN Ia-related events, rich in calcium).

The spectra presented here are arbitrarily scaled and flux-shifted for presentation. No simulated fibre effects or observational noise from the 4MOST ETC has been added. As noted in Section 3, the primary purpose of the spectra is for simulating realistic light-curve information for LSST rather than accurately portraying the spectra of a given transient class. Rarely observed transient classes, such as TDEs and CaRTs are essential featureless blue-dominated continua, combined with a limited presence in classifier training samples/template banks, is likely partially responsible for their incredibly poor classification results.

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