Characterization of Supernovae Based on the Spectral–Temporal Energy Distribution: Two Possible SN Ib Subtypes

Ofek Bengyat\textsuperscript{1,2} and Avishay Gal-Yam\textsuperscript{1}

\textsuperscript{1}Department of Particle Physics and Astrophysics, Weizmann Institute of Science, 76100 Rehovot, Israel
\textsuperscript{2}Faculty of Physics, University of Vienna, Boltzmanngasse 5, A-1090 Vienna, Austria

Received 2022 February 21; revised 2022 March 22; accepted 2022 March 25; published 2022 May 2

Abstract

A quantitative data-driven comparison among supernovae (SNe) based on their spectral time series combined with multiband photometry is presented. We use an unsupervised random forest algorithm as a metric on a set of 82 well-documented SNe representing all the main spectroscopic types, in order to embed these in an abstract metric space reflecting shared correlations between the objects. We visualize the resulting metric space in 3D, revealing strong agreement with the current spectroscopic classification scheme. The embedding splits Type Ib supernovae into two groups, with one subgroup exhibiting broader, less prominent, higher-velocity lines than the other, possibly suggesting a new SN Ib subclass is required. The method could be to classify newly discovered SNe into two groups, with one subgroup exhibiting broader, less prominent, higher-velocity lines than the other, possibly suggesting a new SN Ib subclass is required. The method could be to classify newly discovered SNe.

1. Introduction

The existing classification scheme of supernovae (SNe; Filippenko 1997; Gal-Yam 2017) has been successful in sorting out the majority of objects discovered, in a way that facilitates their study or practical use, e.g., as standard candles (Riess et al. 1998; Perlmutter et al. 1999). It does that by relying mainly on their spectra.

The current classification scheme divides SNe into Type I and Type II, comprised respectively of SNe lacking H lines and those containing them. The reason for this is mostly historical (Minkowski 1941), and modern understanding distinguishes between two other categories based on the progenitor star and the underlying physical explosion process. The first class is thermonuclear SNe, including mostly members of the spectroscopic Type Ia group, that were found (Nugent et al. 2011) to be a result of a white dwarf (WDs) progenitor, exploding in a manner still actively researched (Polin et al. 2019; Magee et al. 2021). Spectroscopically, normal Type Ia SNe are identifiable by the strong Si II 6355 Å line seen in peak spectra, though this is less clear in some subtypes. The rarer group of Ca-rich Type I SNe has also been associated with long-lived, WD progenitors (Perets et al. 2010; De et al. 2020). The second category is core-collapse SNe, originating from massive stars whose self-gravity ceases to be supported by nuclear fusion in their core. These are further divided into the H-rich Types II and IIn, the latter distinguished by their narrow Balmer emission lines ascribed to circumstellar material (CSM), and the stripped-envelope SN Types, which are the explosive deaths of massive stars that have already lost their outer hydrogen shell prior to explosion. These include Types Ib and Ic, the latter differing by the lack of He in peak spectra (in particular, the 5876, 6678, and 7065 Å lines). Type Ib SNe exhibit hydrogen lines only in early spectra, which later disappear, indicating a much thinner layer of H in their progenitor star than those of normal Type II SNe. He-poor stripped-envelope SNe with broad spectral lines belong to Type Ic-BL, corresponding to explosions with larger kinetic energy per unit mass, and hence higher ejecta velocity—up to about threefold compared to the \( \sim 10^4 \text{ km s}^{-1} \) of normal Type Ic SNe. While these are the main, most frequently observed Types, others exist, including stripped-envelope SNe with CSM interaction of Types Ibn (e.g., Pastorello et al. 2008) and Icn (Gal-Yam et al. 2022), superluminous SNe (e.g., Gal-Yam 2019), as well as numerous subdivisions of the Types mentioned here. A recent review regarding the current classification scheme of SNe can be found in Gal-Yam (2017).

In practice, an experienced individual could examine a spectrum of some SN, mainly by looking at emission or absorption lines, determine its type, and enter it to some database. The problem of subjectivity that could possibly emerge in this human classifier scenario has been treated by developing tools such as Superfit (Howell et al. 2005) or SNID (Blondin & Tonry 2007), both relying on a template bank of SN spectra to which a query spectrum is compared. Considerations such as redshift, reddening through Milky Way or host dust, photometric calibration of the spectrum and contamination by host lines are present in these methods as fit parameters or have their effects diminished by smoothing or continuum estimation and subtraction. These methods, as opposed to human classification, standardize the treatment of different SNe such that different reddening or contamination conditions have less impact on the classification, and they are also more objective, being independent of the user, one’s attention to detail, or other forms of human error. The use of such automatic tools also streamlines the task of classification.
of large amounts of data. Another quantitative approach to classification of Type I SNe, also relying on peak spectra, was presented by Sun & Gal-Yam (2017). This classification uses the estimation of the depths of a line around 6150 Å attributable to Si II or H, depending on the SN, and of O1777 Å, and achieves significant separation between the main Type I SN subtypes (Ia, Ib, and Ic), as well as between different Ia subtypes.

However, one drawback in the presently available schemes and methods that still awaits remedy is that they do not rely on the entirety of information available for an SN, but on a single spectrum of it, i.e., from a single epoch. This could also lead to ambiguity, this time not a subjective, user-dependent one, but the mere fact that the current scheme could assign different types to the same object when faced with different spectra of it. An example is Type IIb, which includes core-collapse SNe exhibiting H in early spectra but not in later ones. Prentice & Mazzali (2017) have approached the study of this type-continuum by distinguishing between four subtypes ranging from Ib to IIb based on the Hα emission and absorption lines. They also treated the continuum between the He-poor stripped core-collapse SNe Types Ic and Ic-BL, the latter being defined by significantly higher explosion velocities than the former, evident in the breadth of spectral lines. This difference in velocities is, just as in the Ib–IIb case, not only a continuous parameter discretized into two groups, but also a time-dependent property, as they show that it changes throughout the lifetime of the transient. They do this by examining in practice the related (though only weakly correlated) property of feature count, and divide the continuum into five subtypes depending on the average of the feature count over a certain time interval. Examples for He-poor stripped core-collapse events of uncertain classification include SN2004aw (Mazzali et al. 2017) and SN2016coi (Kumar et al. 2018). Additionally, a certain confusion exists with regard to the classification of stripped core-collapse SNe as Ia or Ic. One reason for it is the presence of He, observed in some events initially classified as Type Ic (e.g., SN2007gr and SN2009jf). This has been perceived as a flaw in the current classification scheme, rather than an incomplete physical understanding of the objects (Blondin & Tonry 2007; Valenti et al. 2011; Dessart et al. 2012).

There exists then a certain need for a classification scheme that both quantitative and taking into account the entire spectral–temporal energy distribution of a SN (i.e., the spectral flux $f(t, \lambda)$ along both wavelength and time) in order to provide a more definite answer about the degree of similarity between, and thus the proper taxonomic sorting of SNe. Such a method could perhaps further unveil some structure in the space of SNe, possibly in the form of a few hidden parameters controlling their properties. Those might, ideally, be later ascribed to physical properties through correlation analysis.

1.1. Machine-learning Methods in Astronomy

Various methods of data analysis have been used in astronomy, the emphasis in recent years having been placed on machine-learning methods. A few examples for uses in the context of SN classification include Sasdelli et al. (2014, 2016) and Williamson et al. (2019). Fremling et al. (2021) present a binary deep classifier named SNIAscore, which succeeds in classifying SNe as Type Ia with very low error rates. See also Baron (2019) for a useful review on machine-learning methods in astronomy, including those used in this work. The question of quantitative SN classification could be approached by means of some machine-learning algorithm that would be trained on feature vectors describing a set of SNe. This training set should then ideally consist of well-sampled SNe, whose classification and comparison to other objects is well-understood. After training, the algorithm should be able to provide new insight about this taxonomy of and similarity between objects when input SNe from or outside of the training set. The method we use in this work is an unsupervised Random Forest (RF) serving as a metric (Shi & Horvath 2006), which has been found viable in defining a distance between spectra (Baron & Poznanski 2016; Reis et al. 2018). Section 3.3 describes the algorithm and its application.

2. Objective

In this work, we present a method to characterize, or classify, SNe based on their spectral–temporal energy distribution by means of a data-driven embedding in an abstract metric space. We compare the spectroscopic types in the current scheme—as determined by a human expert after examining the spectral sequence available for the SN in question—with the resulting embedding. Other than this comparison, we examine possible type-continua or new subtypes that may arise from the embedding.

3. Method

3.1. Estimating the Spectral–Temporal Energy Distribution of an SN

The raw data of each SN, i.e., its spectra and multiband photometry, are processed with PyCoCo, a program for template creation of spectral energy distribution of SNe developed by Vincenzi et al. (2019, hereafter V19). It uses Gaussian Process Regression to interpolate the flux in the intervals between the input spectra, with the photometry data used both for flux calibration of the spectra and as data points. PyCoCo also corrects the flux for Milky Way extinction and host extinction values the user inputs, and removes tellurics and host lines, as well as transforming the data into the rest frame of the SN. The final output from PyCoCo is the (normalized) rest-frame spectral energy distribution $f(t, \lambda)$ of the input SN on some grid of time and wavelength.

We keep the power $n$ of the light-curve rising fit in PyCoCo, $f \propto (t - t_0)^n$, (V19, Equation (2)) as a free parameter for all the SNe we apply it on, and we also choose to not extend and oversample the light curves at late times, as our analysis concentrates on relatively early times (see next section). As the majority of PyCoCo outputs used in this work were generated by V19 (see Section 4), it should be mentioned that they only left $n$ as a free parameter in cases where the photometric data shortly after explosion was good enough to allow a reasonable fit. When that was not the case, $n$ was fixed to a value that depends on whether the SN is H-rich (Types II and IIn) or not (the rest of the types), as indicated by the user. This is the only place where input regarding classification is used in PyCoCo. It should also be noted that a different light-curve rise model was used for the data sets of Type IIb SN2019J3, SN2011dh, SN2011fu, and SN2013df generated by V19, in order to take

---

5. Host uncorrected output is also produced, though in this work we only use the host corrected versions.  

---
into account their double early peaks. This is also true for SN2006aj, but in this case its double peak is likely a result of its unusually early detection, rather than an actual peculiarity (Mirabal et al. 2006), so we omit the first peak data for this object. Readers are referred to Section 2.1 in V19 for details about the light-curve and photometry are not yet calibrated or dereddened.

The code used in this section is available on GitHub. See Figure 1 for an example of the process.

3.2. Preparation of Feature Vectors

We use an algorithm from Carnall (2017) to rebin in wavelength and perform linear interpolation in time to obtain the fluxes on a uniform grid $T \times \Lambda$ over all SNe, where

$$T = 0 \text{ to } 50 \text{ days at } 1 \text{ day intervals}$$

$$\Lambda = 4000 \text{ Å to } 8000 \text{ Å at } 40 \text{ Å intervals}.$$  

Times are counted since the estimated explosion of each template. We choose this wavelength interval $\Lambda$ because it is covered available photometric data for almost all SNe in our data set, making the output of PyCoCo more credible. The time interval $T$ is chosen such that it includes times where most SNe are not too faint.

As others have done (e.g., Sasdelli et al. 2014, or less directly in Blondin & Tonry 2007), we use the wavelength derivative of the logarithm of the flux of every SN,

$$\tilde{f}(t, \lambda) \equiv \frac{d}{d\lambda} \log f(t, \lambda),$$

instead of the flux itself, in order to emphasize spectral lines and diminish the effect of the unaccounted for reddening. The data are then sorted into an $N$-row matrix $X$, where each row $X_i$ corresponds to the $i$th SN in our data set:

$$X_i = (\tilde{f}(t_j, \lambda_j))_{(t_j,\lambda_j) \in T \times \Lambda}.$$ 

In order to account for some missing data points, we then use the Expectation Maximization Principal Component Analysis (EMPCA) algorithm presented in Bailey (2012) on our feature matrix $X$ to produce a principal component matrix, $P_C$, which is then immediately transformed back to the feature space to produce $X_C$, now without missing data. The number of principal components we choose is $C = 50$, which preserves 93% of the variance.

3.3. Unsupervised Random Forest as a Metric

We now train an unsupervised RF on the data in $X_C$ and use it as a metric on the set of SNe as described by the same matrix $X_C$. This section describes this process.

RF is originally a supervised machine-learning classifier (Ho 1995). A brief description of it follows, regarding only binary classification as relevant for this work. First, we describe binary decision trees. A decision tree is a supervised classifier by itself, structured as a tree graph—a directed graph with a single start node, several terminal nodes, and several “generations” of intermediate nodes in between, where every intermediate node has exactly one arrow pointing to it from its parent and exactly two arrows that it points to its children. When an object is queried into the decision tree, each nonterminal node $k$ hands it over to one of its two child nodes, chosen according to the truth value of a condition of the form $f_k^i < c_k$, where $f_k^i$ is the value of the $i$th feature of the tree input, and $c_k$ is a real number. The training process consists of the simultaneous input of all objects in the training set. Before propagating a set of objects that have arrived to any node $k$, the training algorithm chooses the parameters $t_k$ and $c_k$ providing the best separation between the known labels for the objects. The definition of goodness of separation is a matter of choice (a so-called hyperparameter). When all of the objects that arrived in a node are of the same label, this node is set as a terminal node and is assigned the label. Training ceases when all objects arrived at terminal nodes. Classifying a queried object occurs by propagating it through the trained decision tree and returning the label corresponding to the terminal node it reaches. Note that decision trees and their training process are deterministic. See Figure 2 for an illustration and the classification scheme for Type I SNe presented in Sun & Gal-Yam (2017, Section 4.2) for a working example.

An RF consists of multiple decision trees. Every tree is (i) only trained on a random subset of the training set, and (ii) only allowed to use a random subset of the features. The class assigned to a queried object is determined by a majority vote of all the decision trees in the forest. An RF is thus an ensemble learning method, aggregating the results of the trees to achieve better accuracy and robustness. In order to use an RF as a metric, one performs the following unsupervised training process. The training data set in question is used to create a
synthetic data set of the same size. The samples in the synthetic data set have the same number of features as in the original data set and the same marginal distributions for each feature, but zero correlations between features. Supervised training follows, as described above, using the real and synthetic sets, where the real set is assigned the label \textit{real} and the synthetic set is assigned the label \textit{synth}. Note that these labels have nothing to do with SN types or with the data set itself in any way, and they merely differentiate between the real, correlation-exhibiting data, and the synthetic, uncorrelated data. After training, the similarity returned for two queried objects is the fraction of trees in the forest for which the two objects end up in the same terminal node, out of only the trees that assign label \textit{real} for both objects (i.e., the correct label, as one tries to find the similarity measure between two real objects). The metric, distance, or dissimilarity (used interchangeably in this work) is obtained by subtracting the similarity from unity.

This similarity measure thus learns the most prominent correlations in the training data, and measures to what extent do two given objects share such correlations. One such correlation could be, for instance, the presence of all three strongest He lines in a spectrum, as opposed to only some of them.

The algorithm we use was written by Baron & Poznanski (2016). The number of trees used is 2000, and the separability criterion is the Gini Impurity. The output of our analysis is the distance matrix for our SN data set, with entries between zero and one.

The code used in Sections 3.2 and 3.3, as well as in the following visualization of the results is available on GitHub.\(^7\)

4. Data

Alongside the description of the PyCoCo code, \textit{V19} also includes the code output for 67 core-collapse SNe, listed in Table 1 of their publication. Our data set is built upon those SNe, with some exclusions and type changes of SNe listed in Table 2. The final list of 82 SNe used in this work is given in Table 3, along with the Data Quality Index (DQI) described next. This data set will be made available upon request.

Two SN Types missing from the \textit{V19} set, which we also did not add, are Ca-rich SNe and H-poor superluminous SNe (SLSN-I). In the case of Ca-rich SNe, the reason is the

\(^7\) https://github.com/dalya/WeirdestGalaxies
\(^8\) https://github.com/ofek-b/spectra_in_time

**Figure 2.** An illustration of a trained binary decision tree, multiple different instances of which comprise an RF. The input object propagates according to its features \(f_i\).

**Table 1**

| Name      | Type Made to SNe from V19 | Comments                                                                 |
|-----------|---------------------------|---------------------------------------------------------------------------|
| Excluded: |                           |                                                                           |
| SN1987A   | 87A-like                  | Long rise time with peak outside our temporal range                       |
| SN2005bf  | Ib                        | Very anomalous, double-peaked light curve                                 |
|           |                           | (Folatelli et al. 2006)                                                  |
| SN2008D   | Ib                        | Highly extinguished (Rabinak & Waxman 2011); unique early time data       |
| SN2011bm  | Ic                        | Very wide light curve (Valenti et al. 2012)                               |
| SN2016bkv | II                        | Faint, unusually long light curve, weaker lines (Nakaoka et al. 2018)     |
| SN2008in  | II                        | DQI < 0.8                                                                 |
| SN2009dd  | II                        | DQI < 0.8                                                                 |
| SN2013fs  | II                        | Added again manually for the analysis in Section 5.3. Appears on Table 2.|
| SN2008aq  | IIn                       | DQI < 0.8                                                                 |
| SN2009ip  | IIn                       | Peculiar object, may not be an SN (Mauerhan et al. 2013)                   |
| SN2011ht  | IIn                       | DQI < 0.8                                                                 |
| Type changed: |                 |                                                                           |
| SN2009jf  | Ib                        | Changed to Ic; see Gal-Yam (2017)                                         |
| SN2010al  | IIn                       | Changed to Ibn (Pastorello et al. 2015)                                   |
| SN2008eq  | IIn                       | Changed to II                                                             |
| SN2007pk  | IIn                       | Changed to II                                                             |
poor data sets available for those objects. In the case of SLSN-I, peak light often occurs late and outside our temporal window, such that our selected time grid would miss important information about those objects; the same holds also for SN1987A-like Type II SNe. One could envision including these groups once more data are available, and including a temporal renormalization for SLSNe-I and 87A-like events.
For the SNe we added, spectra were taken from WISEREP\(^9\) (Yaron & Gal-Yam 2012) and photometry from the Open Supernova Catalog\(^{10}\) (Guillochon et al. 2017), where available. The values for Milky Way reddening $E(B-V)_{\text{MW}}$ were taken from Schlafly & Finkbeiner (2011). The host reddening estimations $E(B-V)_{\text{host}}$ were taken from the provided references. When no estimation could be found, the median value for the SN type was taken.

### 4.1. Data Quality Index

The quality of the output of PyCoCo depends on the coverage of the spectral–temporal region $T \times \Lambda$ by the spectroscopic and photometric data of each object. A DQI between 0 and 1 is thus calculated for each SN in the following manner. For each point on the $T \times \Lambda$ grid, the distance to the nearest input data point (spectroscopic or photometric) is calculated as the usual Euclidean distance, where the time and wavelength axes are normalized by their total length. We define the DQI as the fraction of grid points that are closer than 0.1 to a data point. We only included SNe with DQI $\geq 0.8$ in our analysis. The impact of this choice is minimal (the removal of four objects; Table 1).

### 5. Results

After applying the EMPCA algorithm, the fraction of data variance explained is 93%. Figure 3 shows the distance matrix $D$ resulting from an unsupervised RF as described. It is not completely symmetrical, due to computation errors—the median absolute symmetric difference of off-diagonal terms is 0.005—so we enforce symmetry by using $D + D^T$ as the dissimilarity matrix. A certain degree of block-diagonality is noticeable, expressing agreement with the current classification scheme for these well-documented SNe.

---

\(^9\) https://wiserep.weizmann.ac.il

\(^{10}\) https://sne.space
5.1. Visualization of the Metric Space

For visualization of this dissimilarity matrix inside a low-dimensional Euclidean space while trying to preserve structure, we use t-Distributed Stochastic Neighbor Embedding (tSNE), developed by van der Maaten & Hinton (2008). The visualization in 3D space is shown in Figure 4, with colors...
representing the spectroscopic type assigned by a human expert after manually inspecting the evolution of spectra available for the SN. Examining the visualization, one can notice the aforementioned block-diagonality in the form of well-separated clusters that are relatively homogeneous in type. We remind the reader that the spectroscopic types represented by the colors were not used in the analysis that led to the dissimilarity matrix or the visualization (the only exception is mentioned in Section 3.1).

Another way to visualize the results is to look at the Minimum Spanning Tree (MST) of the fully connected graph defined by the distance matrix. The MST is a fully connected subgraph minimal in the sum of edge weights (here—distances) and is shown in Figure 5. The algorithm by Yamada et al. (2010) was used. The concept of MST in the context of an abstract dissimilarity space could be employed, as was done by Baron & Ménard (2020), to extract sequences from the data, which could later be compared with physical parameters.

Looking at Figure 4, we see several remarkable features. Objects of Type Ibn are close to those of Type IIn, probably due to sharing narrow features. In contrast, the proximity of Type Ic-BL objects to the Type II cluster is unexpected. A possible explanation could be that the features those two groups share with one another are the dominance of the continuum.

The very close pair of objects SN2015ap (IIb, purple) and SN2004gq (Ib, blue) is particularly noticeable. A possible explanation is that SN2015ap, although exhibiting H-absorption initially, loses it at a relatively early stage, ~10 days after explosion, leaving no H evident in its spectrum. Another noteworthy object is SN2004gt (Ibc, cyan), in between the main Ib and Ic clusters, which fits previous classification (Eldridge et al. 2013). As for SN2007Y (main Ib cluster, blue), its initial classification was Ib despite Hα absorption lines observed at early times, and it has been claimed that Ib would be a better classification (Stritzinger et al. 2009; Chevalier & Soderberg 2010; Maurer et al. 2010). The embedding suggests that it is part of the main Ib cluster, though it is on its edge.

As could also be observed in the tSNE visualization, the Type Ic SN2009jf and SN2007gr (two dark green points in the upper main Ic cluster) are very close by, which fits the fact that they are almost identical spectroscopically in the optical and IR wavelengths throughout their evolution (Valenti et al. 2011). They are also close to the Ib cluster, with the He-poor SN2007gr slightly differing from SN2009jf, which shows stronger He lines, leading Valenti et al. (2011) to select a classification of Ib, though they acknowledge the possibility that Ic could be a better choice. Likewise, Dessart et al. (2012) have found that the data of SN2007gr, classified there as Type Ic, may be well-explained by a He-rich model, but with weak lines. These authors conclude from their analysis of Type Ib and Ic SNe that the property usually assumed to define Type Ic events, the lack of He, is justified for very few, if any, objects.

---

11 The MST is generally not unique, though it is unique in this case.
12 https://github.com/dakota-hawkins/yamada

Figure 5. The MST of the graph of SNe defined by the dissimilarity matrix. Note that edge lengths are arbitrary.
Indeed, a continuum such as the one visualized in Figure 4 may be a more suitable way to understand those objects.

5.2. The Split of Type Ib SNe

It can be noticed in the visualizations that some Type Ib SNe are set apart from the main cluster formed by this type. The SNe in the main cluster are SN2004gv, SN2007Y, SN2006ep, SN2015ah, SN1999dn, iPTF13bvn, SN2005hg, and SN2009iz. The offset Type Ib SNe are SN2004gq, SN2007uy, SN2016jdw, SN2017bgu, SN2004dk, and SN2015ap, which is of Type IIb but still included in this list (see Section 5.1). Figure 6 shows a comparison between the mean spectra of these two groups.

The figure suggests that this split could be correlated with the SN expansion velocity, as the features in the mean spectrum of the offset cluster are wider and bluer, especially before and at peak. To validate this effect, we examine the raw spectra of each event near peak (i.e., prior to any processing of the data). We find that the average expansion velocities derived from the location of the absorption minimum of the He I lines at $\lambda\lambda 5876, 6678, 7065$ Å are $9700 \pm 1000$ and $12300 \pm 1200$ km s$^{-1}$ for the main and offset clusters, respectively. The velocities were measured using the dedicated tool on WISeREP. This supports the results of our analysis, and suggests that there may be a split between populations of low-velocity and high-velocity SNe Ib. This of course brings to mind the well-known separation between spectroscopically normal SNe Ic and broad-line, high-velocity SNe Ic-BL (e.g., Modjaz et al. 2016; Gal-Yam 2017; Prentice & Mazzali 2017), though for SNe Ib the velocity spread between normal and high-velocity events may be less extreme, making this division less obvious.

5.3. An Application to SNe with Less Data

While the SNe used for training are some of the best observed objects of their types, we wish to test to what extent it is useful to embed a new SN with a lower amount of data. Figure 7 shows the effect of only including two spectra in the input to PyCoCo on the displacement of the embedding relative to the “true” one produced by using all of the data for the SN. We test this on the Type Ia SN2005cf and the Type II SN2013fs. We do not degrade the photometric data in this test. It can be seen that, at least for those SNe, using only two spectra does not have much effect on the quality of embedding. This is highly relevant because future data sets from ongoing and future surveys such as the Zwicky Transient Facility (ZTF; Bellm et al. 2019; Graham et al. 2019) and the Rubin Observatory will include high-quality multicolor light curves. Spectroscopy is often obtained as soon as possible after discovery (e.g., Gal-Yam et al. 2011) and again around peak, so a typical future data set may likely include this combination of two spectra and good light curves.

6. Conclusion

We present a method for quantitative comparison of SNe by means of an unsupervised embedding in a metric space. This data-driven method identifies the important correlations/spectral features evident in the spectral–temporal energy distributions it is trained on. Other methods that have been used for quantitative classification of SNe (e.g., Prentice & Mazzali 2017; Sun & Gal-Yam 2017) can be understood as supervised decision trees constructed by experienced astrophysicists rather than by an automated training process. Our approach is in principal similar, but chooses an unsupervised method in order to detect the most important correlations—
deemed as such by an objective measure—in a given training set. The method succeeds in defining some clusters, which mostly agreed with the known spectroscopic types, and in visualizing the IIb-IIc-Ic-IIc-BL continuum. There, objects like SN2007Y, SN2009jf, and SN2015ap could get a quantitative answer regarding their Type by this method, while the classical element-based classification is not as conclusive.

One result our method produces is the possible split of Type Ib SNe into two subgroups, of normal and higher-velocity events, perhaps mimicking the more prominent division between spectroscopically normal SNe Ic and SNe IIc-BL. As discussed earlier, this separation could be of a discrete nature and should be investigated further (preferably with a larger sample) to further validate or dismiss its existence and test the underlying differences in the physics of the explosions.

Preliminary tests show that even a small number of spectra, when combined with multiband photometric measurements, suffice for obtaining a relatively accurate embedding. This could make the method useful in the context of surveys such as those to be conducted by the Rubin Observatory.

The application as a classifier for less typical SNe should be approached with caution, as the training set needs to contain a wide enough selection of SNe (which in turn need to have enough data) in order to be able to produce a meaningful result.

We thank Dalya Baron, Ido Irani, Barak Zackay, and the ZTF collaboration for useful advice, and an anonymous referee for a helpful review.

Software: This work made use of the Python packages NumPy (Harris et al. 2020), Matplotlib (Hunter 2007), scikit-learn (Pedregosa et al. 2011), pandas (McKinney 2010), and NetworkX (Hagberg et al. 2008).

ORCID iDs

Ofek Bengyat https://orcid.org/0000-0001-5547-9176
Avishay Gal-Yam https://orcid.org/0000-0002-3653-5598

References

Bailey, S. 2012, PASP, 124, 1015
Baron, D. 2019, arXiv:1904.07248
The Astrophysical Journal, 930:31 (11pp), 2022 May 1

Bengtay & Gal-Yam

Hunter, J. D. 2007, CSE, 9, 90
Jencson, J. E., Prieto, J. L., Kochanek, C. S., et al. 2016, MNRAS, 456, 2622
Kumar, B., Singh, A., Srivastav, S., Sahu, D. K., & Anupama, G. C. 2018, MNRAS, 473, 3776
Magee, M. R., Maguire, K., Kotak, R., & Sim, S. A. 2021, MNRAS, 502, 3533
Matheson, T., Filippenko, A. V., Li, W., Leonard, D. C., & Shields, J. C. 2001, AJ, 121, 1648
Mauerhan, J. C., Smith, N., Filippenko, A. V., et al. 2013, MNRAS, 430, 1801
Mauer, I., Mazzi, P. A., Taubenberger, S., & Hachinger, S. 2010, MNRAS, 409, 1441
Mazzali, P. A., Sauer, D. N., Pian, E., et al. 2017, MNRAS, 469, 2498
McKinney, T., Filippenko, A. V., Li, W., Leonard, D. C., & Shields, J. C. 2001, AJ, 121, 1648
Mauerhan, J. C., Smith, N., Filippenko, A. V., et al. 2013, MNRAS, 430, 1801
Maurer, I., Mazzi, P. A., Taubenberger, S., & Hachinger, S. 2010, MNRAS, 409, 1441
Mazzali, P. A., Sauer, D. N., Pian, E., et al. 2017, MNRAS, 469, 2498
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56
Minkowski, R. 1941, PASP, 53, 224
Mirabal, N., Halpern, J. P., An, D., Thorstensen, J. R., & Terndrup, D. M. 2006, ApJL, 643, L99
Modjaz, M., Kirshner, R. P., & Challis, P. 2008, ApJL, 687, L9
Modjaz, M., Liu, Y. Q., Bianco, F. B., & Graur, O. 2016, ApJ, 832, 108
Modjaz, M., et al. 2014, AJ, 147, 99
Munari, U., Henden, A., Belligoli, R., et al. 2013, NewA, 20, 30
Nakaoka, T., Kawabata, K. S., Maeda, K., et al. 2018, ApJ, 859, 78
Nugent, P. E., et al. 2011, Natur, 480, 344
Ostman, L., et al. 2011, A&A, 526, A28
Perrett, J. T., et al. 2012, ApJ, 752, L26
Pastorello, A., Benetti, S., Brown, P. J., et al. 2015, MNRAS, 449, 1921
Pastorello, A., Mattila, S., Zampieri, L., et al. 2008, MNRAS, 389, 113
Pastorello, A., et al. 2007, MNRAS, 376, 1301
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pereira, R., et al. 2013, A&A, 554, A27
Perets, H. B., Gal-Yam, A., Mazzali, P. A., et al. 2010, Natur, 465, 322
Perlmuter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565
Polin, A., Nugent, P., & Kasen, D. 2019, ApJ, 873, 84
Prentice, S. J., Ashall, C., James, P. A., et al. 2019, MNRAS, 485, 1559
Prentice, S. J., Ashall, C., Mazzi, P. A., et al. 2018, MNRAS, 478, 4162
Prentice, S. J., & Mazzi, P. A. 2017, MNRAS, 469, 2672
Prentice, S. J., Mazzi, P. A., Pian, E., et al. 2016, MNRAS, 458, 2973
Qiu, Y., Li, W., Qiao, Q., & Hu, J. 1999, AJ, 117, 736
Rabinak, I., & Waxman, E. 2011, ApJ, 728, 63
Reis, I., Pozanski, D., Baron, D., Zasowski, G., & Shahaf, S. 2018, MNRAS, 476, 2117
Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009
Sako, M., Basset, B., Becker, A. C., et al. 2018, PASP, 130, 064002
Sas'dellii, M., Hillebrandt, W., Aldering, G., et al. 2014, MNRAS, 447, 1247
Sas'dellii, M., Ishida, E. O., Vilkaitis, R., et al. 2015, MNRAS, 461, 2044
Schlaflly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Shi, T., & Horvath, S. 2006, J. Comput. Graph. Stat., 15, 118
Shivvers, I., et al. 2017, PASP, 129, 054201
Shivvers, I., et al. 2019, MNRAS, 482, 1545
Silverman, J. M., Ganeshalingam, M., & Filippenko, A. V. 2013, MNRAS, 430, 1030
Silverman, J. M., et al. 2012, MNRAS, 425, 1789
Simon, J. D., et al. 2007, ApJ, 671, L25
Smith, N., Silverman, J. M., Filippenko, A. V., et al. 2012, AJ, 143, 17
Smithit, M. T., Brown, P. J., Kuin, P., & Suntzeff, N. B. 2016, PASP, 128, 034501
Stahl, B. E., Zheng, W., de Jaeger, T., et al. 2019, MNRAS, 490, 3882
Stahl, B. E., Zheng, W., de Jaeger, T., et al. 2020, MNRAS, 492, 4325
Sternberg, A., Gal-Yam, A., Simon, J. D., et al. 2011, Sci, 333, 856
Stritzinger, M., Mazzi, P., Phillips, M. M., et al. 2009, ApJ, 696, 713
Stritzinger, M. D., Anderson, J. P., Contreras, C., et al. 2018, A&A, 609, A134
Stritzinger, M. D., et al. 2018, A&A, 609, A135
Stritzinger, M. D., et al. 2002, AJ, 124, 2100
Sun, F., et al. 2017, PASP, 130, 034501
Terreran, G., Margutti, R., Bersier, D., et al. 2019, ApJ, 883, 147
Valenti, S., et al. 2011, MNRAS, 416, 3138
Valenti, S., Howell, D. A., Stritzinger, M. D., et al. 2016, MNRAS, 459, 3939
Valenti, S., et al. 2012, ApJL, 749, L28
Valenti, S., et al. 2008, MNRAS, 383, 1485
van der Maaten, L., & Hinton, G. 2008, JMLR, 9, 2579
Vincenzi, M., Sullivan, M., Firth, R. E., et al. 2019, MNRAS, 489, 5802
Vinkò, J., et al. 2018, PASP, 130, 064101
Wang, X., et al. 2009, ApJ, 697, 380
Williamson, M., Modjaz, M., & Bianco, F. B. 2019, ApJL, 880, L22
Yamada, T., Kataoka, S., & Watanabe, K. 2010, Int. J. Comput. Math., 87, 3175
Yamanaka, M., Maeda, K., Kawabata, M., et al. 2014, ApJL, 782, L35
Yaron, O., & Gal-Yam, A. 2012, PASP, 124, 668
Yaron, O., et al. 2017, NatPh, 13, 510
Zhang, J.-J., et al. 2014, ApJL, 782, L35
Zhang, J., et al. 2018, ApJL, 863, 109
Zhang, K., et al. 2016, ApJ, 820, 67
Zhang, T., et al. 2012, AJ, 144, 131