The Astrophysical Journal Supplement Series, 267:25 (22pp), 2023 August https://doi.org/10.3847/1538-4365/accd6a

Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC)

R. Hložek[1,2], A. I. Malz[3], K. A. Ponder[4,5,6], M. Dai[7,8], G. Narayan[9], E. E. O. Ishida[10], T. Allam Jr[11], A. Bahmanyar[2], X. Bi[12,46], R. Biswas[13], K. Boone[14,46], S. Chen[15,46], N. Du[16,46], A. Erdem[17,46], L. Galbany[18,19], A. Garreta[20,46], S. W. Jha[8], D. O. Jones[21], R. Kessler[22,23], M. Lin[24,46], J. Liu[25,46], M. Lochner[26,27,28], A. A. Mahabal[29,30], K. S. Mandel[31,32], P. Margolis[33,46], J. R. Martínez-Galarza[34], J. D. McEwen[11], D. Muthukrishna[35], Y. Nakatsuka[36,46], T. Noumi[37,46], T. Oya[38,46], H. V. Peiris[13,39], C. M. Peters[2], J. F. Puget[40,46], C. N. Setzer[13], Siddhartha[41,46], S. Stefanov[42,46], T. Xie[16,46], L. Yan[43,46], K.-H. Yeh[44,46], and W. Zuo[45,46]

[1] David A. Dunlap Department of Astronomy and Astrophysics, University of Toronto, 50 St. George St., Toronto, ON M5S 3H4, Canada; hlozek@dunlap.utoronto.ca
[2] Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George St., Toronto, ON M5S 3H4, Canada
[3] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4] Berkeley Center for Cosmological Physics, Campbell Hall 341, University of California Berkeley, Berkeley, CA 94720, USA
[5] Physics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720, USA
[6] SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., Menlo Park, CA 94025, USA
[7] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[8] Rutgers, the State University of New Jersey, 136 Frelinghuysen Rd., Piscataway, NJ 08854 USA
[9] Space Telescope Science Institute, 3700 San Martin Dr., Baltimore, MD 21218, USA
[10] Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France
[11] Mullard Space Science Laboratory, Department of Space and Climate Physics, University College London, Holmbury Hill Rd., Dorking RH5 6NT, UK
[12] Department of Electrical and Computer Engineering, Michigan State University, MI, USA
[13] The Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, AlbaNova, Stockholm, SE-106 91, Sweden
[14] DIRAC Institute, Department of Astronomy, University of Washington, 3910 15th Ave. NE, Seattle, WA 98195, USA
[15] DiDi, Building B1&B2, Digital Valley, Zhongguancun Software Park Compound 8, Dongbeiwang Road, Haidian District, Beijing 100000, People's Republic of China
[16] Department of Computer Science and Engineering, Michigan State University, MI, USA
[17] NVIDIA, Narcisstraat 8B, Utrecht, The Netherlands
[18] Institute of Space Sciences (ICE-CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain
[19] Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain.
[20] University of the Basque Country, Barrio sarriena s/n, E-48940 Leioa, Spain
[21] Gemini Observatory, NSF's NOIRLab, 670 N. A'ohoku Place, Hilo, HI 96720, USA
[22] Kavli Institute for Cosmological Physics, 5640 S Ellis Ave., Chicago, IL 60637, USA
[23] Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA
[24] MediaTek, No.1, Dusing 1st Rd., Hsinchu Science Park, Hsinchu, 30078, Taiwan
[25] NVIDIA, 6454 Living Place, Pittsburgh, PA 15206, USA
[26] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town, 7535, South Africa
[27] South African Radio Astronomy Observatory (SARAO), 2 Fir St., Observatory, Cape Town, 7925, South Africa
[28] African Institute for Mathematical Sciences, 6 Melrose Rd., Muizenberg, 7945, South Africa
[29] Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA
[30] Center for Data Driven Discovery, California Institute of Technology, Pasadena, CA 91125, USA
[31] Institute of Astronomy and Kavli Institute for Cosmology, Madingley Rd., Cambridge, CB3 0HA, UK
[32] Statistical Laboratory, DPMMS, University of Cambridge, Wilberforce Rd., Cambridge, CB3 0WB, UK
[33] Freelance Machine Learning Specialist, Kaggle Competitions Grandmaster, Zurich, Switzerlang
[34] Center for Astrophysics | Harvard & Smithsonian, 60 Garden St., Mail Stop 66, Cambridge, MA 02138, USA
[35] Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[36] Astamuse Company Ltd., BIZCORE Jinbocho 4F, 3-9-2, Kanda Ogawamachi, Chiyoda Ku, Tokyo To, 101-0052, Japan
[37] Keyence, 1-3-14, Higashinakajima, Higashiyodogawa-ku, Osaka 533-8555, Japan
[38] Department of Applied Physics, Waseda University, 1 Chome-104 Totsukamachi, Shinjuku City, Tokyo 169-8050, Japan
[39] Department of Physics and Astronomy, University College London, Gower St., London, WC1E 6BT, UK
[40] NVIDIA France, 350 Avenue de Boulouris, F-83700 Saint Raphael, France
[41] Hertzwell Pte., 19 Jalan Jintan, Singapore
[42] Independent researcher, Sofia, Bulgaria
[43] Independent researcher, Utåkravägen 6, SE-433 65 Sävedalen, Sweden
[44] H2O.ai, 2307 Leghorn St., Mountain View, CA 94043, USA
[45] Independent researcher, Guangzhou, Guangdong, 510175, People's Republic of China

Received 2021 March 29; revised 2023 April 4; accepted 2023 April 13; published 2023 July 21

Abstract

Next-generation surveys like the Legacy Survey of Space and Time (LSST) on the Vera C. Rubin Observatory (Rubin) will generate orders of magnitude more discoveries of transients and variable stars than previous surveys.

[46] PLAsTiCC top team member with publicly available code.

To prepare for this data deluge, we developed the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC), a competition that aimed to catalyze the development of robust classifiers under LSST-like conditions of a nonrepresentative training set for a large photometric test set of imbalanced classes. Over 1000 teams participated in PLAsTiCC, which was hosted in the Kaggle data science competition platform between 2018 September 28 and 2018 December 17, ultimately identifying three winners in 2019 February. Participants produced classifiers employing a diverse set of machine-learning techniques including hybrid combinations and ensemble averages of a range of approaches, among them boosted decision trees, neural networks, and multilayer perceptrons. The strong performance of the top three classifiers on Type Ia supernovae and kilonovae represent a major improvement over the current state of the art within astronomy. This paper summarizes the most promising methods and evaluates their results in detail, highlighting future directions both for classifier development and simulation needs for a next-generation PLAsTiCC data set.

Unified Astronomy Thesaurus concepts: Astrostatistics (1882); Observational cosmology (1146); Transient detection (1957); Astronomy software (1855)

1. Introduction

The Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al. 2009) of the Vera C. Rubin Observatory will be a survey of the full sky observable from its location in the deserts of northern Chile, with a combination of survey area and depth that is unprecedented. Rubin will make a complete map of the sky every few days. This repeated scanning of the sky will be used to generate difference images formed by subtracting that map from a reference template, leading to millions of "alerts" every night for objects (or detected sources) changing flux. The repeated scanning will also build up light curves (a time series of measured light observed in various filters) for the transients and variable stars (Ivezić et al. 2019), whose brightness changes with time.

This large and heterogeneous data set consisting of signals from different types of astronomical objects will need to be classified into different categories or types of transients and variables from the photometric light-curve data. Only a small fraction of objects will be followed up spectroscopically to confirm their types. The need for accurate photometric classification of light curves is therefore a longstanding goal in transient astrophysics. Classification challenges have often been designed with a focus on correctly identifying one class from a given data set.

Previous community efforts to simulate future surveys included the Supernova Photometric Classification Challenge (SNPCC; Kessler et al. 2010), which galvanized the astronomy community to develop classification methods for three classes of supernovae given a smaller spectroscopic training set. The SNPCC performance metric had a strong emphasis on returning a sample that was both pure (one which had a small number of false positive classifications) and complete (one which had a large fraction of true positive classifications), and its performance metric balanced those criteria. Classification challenges must often balance competing goals, with the balance between the two axes often depending on spectroscopic resource efficiency and the brightness limit of any survey, known as the photometric depth. While individual science groups have historically focused on their specific science needs and generated individualized simulations, any classification algorithm that will be run on real and heterogeneous data from LSST will need to be robust to that heterogeneity.

The motivations behind PLAsTiCC were somewhat different than those of groups focused on one scientific question. Given the heterogeneous types of objects that we expect LSST will provide, and the range of science questions we will want to answer with these data, the PLAsTiCC team instead asked the question: Can you classify a broad range of objects with a good overall/average classification of each type, and identify new objects that are not included in the training set? The PLAsTiCC challenge was presented to the broader public through the Kaggle[47] platform, a popular site for data challenges and machine-learning competitions. While the participants in PLAsTiCC submitted classification methods that were often a combination of different strategies, their metric performance was not tuned to any given type and allowed for the comparison of classifiers in cases of degeneracies between classes that had not been explored before.

The paper is organized as follows: In Section 2, we present a general PLAsTiCC overview including data simulation, validation, metric, and information about the competition. Section 3 details the top three submissions and more general aspects of the methods used in the top classifiers. We study alternative methods of exploring the performance of the classifiers in Section 4, and summarize the challenges presented by the data in Section 5. We conclude with key insights from this challenge in Section 6. The authors of the top 10 PLAsTiCC submissions (top 12 submissions overall) are included in the author list of this paper to recognize their contribution to the challenge and the community through their shared solutions.

2. Overview of PLAsTiCC

2.1. Data Simulation

The PLAsTiCC data were simulated to mimic 3 yr of an LSST-type survey (which is slated to run for a decade) using the SuperNova ANAlysis package (SNANA; Kessler et al. 2009). In 2017 May, an initial call was put out to the astronomy community to develop models of extragalactic transients, Galactic variables, and novel astronomical transients. In response to the call, 18 different models were generated according to estimated rates and distributions on the sky. The models included:

1. Extragalactic models including exploding supernovae (SNe), which have large variations in peak brightness, kilonova (KN) models, which are the electromagnetic output from two colliding neutron stars, the emission from the tidal disruption events (TDE) of stars due to the supermassive black hole at the center of a galaxy, and emission from gas falling onto the black hole creating an

[47] The PLAsTiCC challenge is found at https://www.kaggle.com/c/PLAsTiCC-2018.

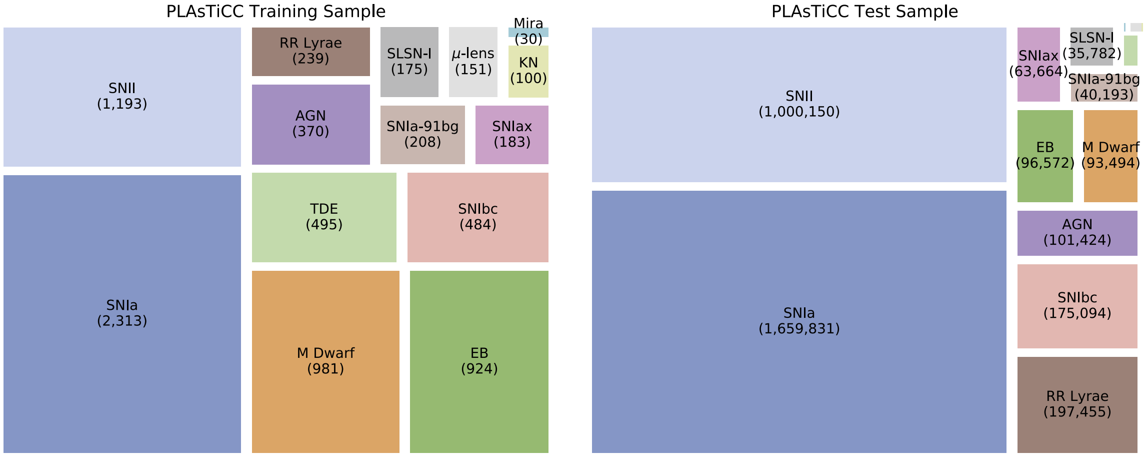


Figure 1. Relative numbers of objects of each class between the training (left) and test (right) data of PLAsTiCC. The size of the boxes is proportional to the relative numbers in each set and the absolute numbers are in parentheses. The simulated ≃8000 objects in the training set were distributed across the data classes in a different ratio compared to the test data of over 3.5 million objects.

active galactic nucleus (AGN). The supernova models represent a wide variety of scenarios, from the explosion of stars at the end of their lives (SN II, SN Ibc), exploding white dwarf progenitors (the SN Ia models including SN Ia, SN Ia-91bg and SN Iax), and the superluminous supernovae (SLSN).

2. Galactic models, which include variable stars (Mira variables, RR Lyrae, and eclipsing binaries (EB)), and nonrecurring models such as the single microlensing events (μlens-Single) and the M-dwarf flares.
3. "Novel" objects: pair-instability supernovae (PISN), calcium-rich transients (CaRTs), intermediate luminosity optical transients (ILOTs), and binary microlensing events (μlens-Binary). These objects do not appear in large numbers in the simulation, and so may represent a challenge to detection/classification algorithms.

The simulation generated 100 million sources, and a subset of 3.5 million were predicted to be detected by Rubin, and thus useful for classification and analysis. These data were provided to participants in PLAsTiCC as 3.5 million light curves in the *ugrizY* filters containing over 453 million observations. These data formed the "test" sample that PLAsTiCC participants were asked to classify. The training set consisted of 8000 objects constructed as a mock spectroscopic subset of the full data, meant to mimic the available data from current and near-term spectroscopic surveys that will be available at the start of Rubin science operations. A small subset of triggered events were flagged as spectroscopically identified and used for the training set. We modeled the spectroscopic training set on a 4-meter Multi-Object Spectrograph Telescope (4MOST; de Jong et al. 2019) which is a 2400-fiber spectroscopic instrument under construction for the development of the Visible and Infrared Survey Telescope for Astronomy. 4MOST will run a "Time Domain Spectroscopic Survey" (Swann et al. 2019). The training set is designed to be small, representing only 0.2% of the test set, and is biased toward brighter events leading to a nonrepresentative subsample of the full test set. The mock spectroscopic selection function is described in more detail in Section 6.4 of Kessler et al. (2019a). The differences in numbers and representation between the training and test data sets can be seen in Figure 1.

The PLAsTiCC data were simulated to mimic future Rubin observations in both the deep-drilling fields (DDFs), which are small patches of the sky that will be sampled often to achieve enhanced depth, and a Wide–Fast–Deep (WFD) survey that covers a larger part of the sky (at ≃ 18,000 square degrees it is almost 400 times the DDF area) that will be observed less frequently. Compared to DDFs, WFD will include many more objects but have lower signal-to-noise ratios. While the PLAsTiCC simulation was performed for one particular survey configuration that is not optimal for transients, classification performance can be used to evaluate survey design for the LSST. We leave this to future work. The extragalactic models were simulated from the PLAsTiCC models as described in Kessler et al. (2019a, see Figure 13 for the full simulation procedure) and Kessler et al. (2019b), by generating a spectral energy distribution (SED) at each epoch for the source, simulating a cosmological distance and peculiar velocity for the source, computing the lensing seen by that source, redshifting the SED, and accounting for Galactic extinction. Each extragalactic object was simulated with a unique redshift, host galaxy extinction, R.A. and decl., and cadence. The SED is integrated in each filter to produce a light curve for the object. The Galactic models are instead defined by a 4 yr time sequence of true magnitudes in the *ugrizY* filter bands, which are then sampled to form a light curve. The rates for the different objects were supplied by the modelers. For the Galactic models, these rates included a dependence on Galactic coordinates, while the extragalactic rates were assumed to be isotropic. Galactic and AGN models were simulated with the unique initial phase, reference-image flux, R.A., and decl. and cadence.

Once the "pure" source model was obtained, it was combined with a noise model specific to the observational conditions of the Rubin Observatory (Biswas et al. 2020), including the cadence information, zero-points, sky noise, and point-spread function (PSF; Kessler et al. 2009). The objects have to be "detected" in order to be included in PLAsTiCC. We required that each source be detected twice at roughly 5σ in the difference image between two nights, so sources must be bright and vary in brightness to be valid PLAsTiCC objects. For detected events, a training subset was tagged as spectroscopically identified, and another subset was tagged as having a spectroscopically measured redshift of the host galaxy.

Rubin may well discover new transients that are not captured by current models, and it may be sensitive to theorized objects that have not yet been observed due to the limitations of current surveys. Thus, the training data will not be fully representative of the larger test data that will probe fainter depths and larger redshifts than the training data. To capture this nonrepresentativity, four of the models simulated in the PLAsTiCC test data, namely the PISN, CaRTs, ILOT, and μlens-Binary were not provided in the training set (see Kessler et al. 2019a, for individual model references).

As part of the challenge, participants were asked to classify objects in the test set into a different class of previously unknown objects described also as the "Other" class. Participants were not given any information about the "Other" class or the composite nature of this class, which is discussed in more detail in Section 3. The simulation procedure and the models involved are discussed in extensive detail in Kessler et al. (2019a). The models are provided for use by the community (PLAsTiCC-Modelers 2019).

2.2. Data Validation

One of the most critical parts of developing PLAsTiCC was validating the data, to ensure that all light curves had been simulated without irregularities or too large/small errors, and that unphysical correlations (e.g., catalog ID number versus decl.) were not present. In addition, we validated the physical models that were input to the simulations. We also validated and improved the SNANA simulation code. We list the validations performed on the simulations below.

Unphysical artifacts. When processing light curves and metadata (additional information on the objects including their sky position, redshift, etc.) in order to classify objects, unphysical artifacts or missing light-curve data can lead to leakage. As one example, modelers typically provided SED information starting at 1000 Å or 2000 Å, which will cause the u band (with a central wavelength around 3000 Å) to drop out at extreme redshift ($z \simeq 3$). To avoid UV dropouts in the simulation, we extrapolated the UV flux to 500 Å to ensure that all u-band observations were included (Kessler et al. 2019a). If left unchecked, classifiers might use this zero flux as a feature; however, this is unphysical and should be ignored. If we were unable to debug or account for an artifact, the affected data were removed.

Distribution tests. The PLAsTiCC data were simulated from model objects at a given rate and from SEDs that were redshifted, and then "observed" in the LSST bands with suitable noise properties and observing conditions. The first step in the validation procedure was to perform tests of the distribution properties of the objects to ensure that they matched those of known objects for models based on observations or expected distributions for theoretically proposed models. We performed checks of the distributions of the maximum and minimum flux for the objects, their redshift distribution, and rates. These tests caught a host of minor issues in reported rates; in many cases the model assumptions were revised for inclusion in the final suite of models. The real data from Rubin will not be validated in exactly the same way given that much of the initial classification will proceed from the alert stream directly; however, quality cuts on the data will ensure data quality for any light curves used for a specific scientific question.

Light-curve tests. Once the average properties of the population were inspected, individual light curves were plotted from the simulated populations. These tests were particularly relevant for models derived from objects that had previously been observed (rather than objects simulated from purely theoretical models). The light curves were compared to databases of known examples to diagnose model inconsistencies and shape discontinuities. This validation process led to decisions on how the PLAsTiCC simulation would handle saturation in the LSST bands for objects brighter than 16th magnitude. However, we did not want to introduce saturation issues in the challenge, and instead simulated a saturation level at 12th mag, or 4 mag brighter than nominal. The small number of saturated observations that remained were removed.

Prior to generating the PLAsTiCC sample, we simulated "perfect" examples of each model type at a high cadence and with no noise. These perfect light curves were compared with the input observations used to generate the model, and then with the results of small runs of the simulations in both the DDFs (which would probe higher redshift with higher signal-to-noise) and the WFD, to test the rates of objects more completely and whether or not there were objects that would not be detected given the cadence and brightness specifications of the survey. Perfect simulations were used to validate the models while the WFD and DDF light curves were used to validate the realistic simulations.

Specialized model tests. Depending on the type of transient or variable, specialized tests on object properties were carried out for individual models. For example the period–luminosity relationship was tested for variable objects like the RR Lyrae. These tests were typically performed with input from the model proposers directly, to ensure that the expected model relationships remained intact in the PLAsTiCC simulations.

Checks on metadata. Once the models were verified for correctness, the additional data about the objects, or metadata (the redshifts, sky position, Galactic coordinates, extinction) were analyzed to ensure that the data are simultaneously useful and exhaustive without rendering the challenge trivial. This involved iterating on the form and description of the metadata, knowing that the user interface/workbook provided to initiate the challenge would be the place where many participants got their information. The metadata were also checked for consistency by, for example, plotting R.A. and decl. per target.

Correlations between object variables. Related to checks of the metadata, we tested for spurious or obvious correlations between object metadata and classification type. A trivial example would be to ensure that the object ID was in fact randomly generated not only within a particular type but across various types. If the objects are simulated in order then there would be a type–ID correlation which would render the classification challenge trivial. Some astrophysical correlations do exist, for example the nature of the objects (Galactic versus extragalactic) is correlated with the redshift. A plot of the correlation between training set metadata parameters is shown in Figure 2, where the simulated WFD sample has been split into Galactic and extragalactic objects. There are correlations in the metadata, but they are expected. For example, the redshift is correlated with the redshift error and the target is associated with the dust reddening of the Milky Way (Galactic objects will have more dust than extragalactic). We also do not see unphysical correlations such as with object ID or row ID.

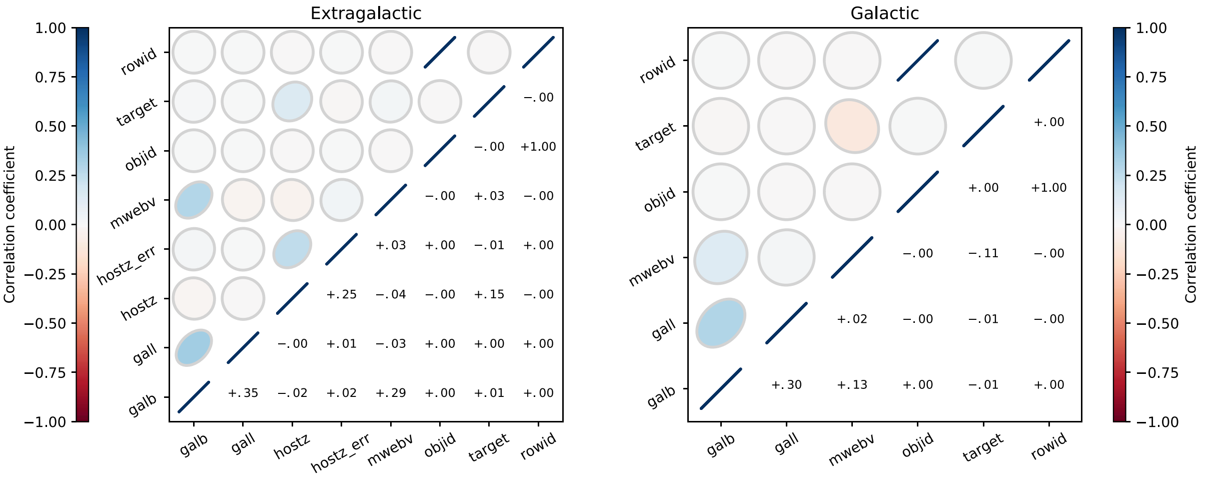


Figure 2. Correlation tests for the simulated WFD PLAsTiCC data. We used this test to detect any spurious correlations between object ID and the type and other properties provided to the contestants. We disambiguate Galactic and extragalactic sources and consider the WFD and DDF data separately as there are physical correlations that are introduced when aggregating these groups, e.g., all Galactic variables will have a host redshift of $z = 0$, and all sources within the four DDFs will show strongly correlated values of the Milky Way reddening.

Simple classification of objects. An additional test of the robustness of the training set was to train a "simple" random forest classification algorithm (Ho 1995; Narayan et al. 2018) on the PLAsTiCC data to ensure that the problem was not trivially solvable, and also provided a rough benchmark for comparing results throughout the challenge. The 11 extracted features per passband are described in Table 2 of Narayan et al. (2018), and included the autocorrelation length, the kurtosis and skewness of the magnitude distribution. The classifier was run on the DDFs and the WFD separately during the validation phase and as such was not run on the complete data set after production. Given that some rare objects such as the kilonovae were not found in the DDFs, a direct comparison of the performance of the simple classifier to the competition entries is less straightforward; however, the simple classifier was merely testing for obvious problems in the simulated sample. In addition, the simple classifier was not run on the "Other" class objects, as testing anomaly detection was not part of the validation process. This classification testing (and the other model validation) was only performed by members of the PLAsTiCC team who had agreed not to submit an entry to the challenge, and not to help challenge participants. Each model type was validated by a subset of the PLAsTiCC development team and was revalidated for each new version of the simulated data.

2.3. The PLAsTiCC Metric

The choice of metric for PLAsTiCC was motivated by two main drivers: the importance of classification uncertainties and the lack of a single, unifying science goal. Unlike its predecessor, the SNPCC, which was motivated by SNe Ia cosmology, PLAsTiCC was motivated by dozens of science drivers from Galactic population statistics to gravitational wave electromagnetic counterparts. The metric was designed to be sufficiently generic so that it would disfavor classifiers that neglected any classes, especially those that are rare in the Universe.

Because the data from Rubin is anticipated to be inherently noisy, the PLAsTiCC metric did not reduce the classification posteriors to deterministic class estimates, thereby including classification uncertainty. In contrast to traditional classification tasks that request a single estimated class $\hat{m}_n$ for each classified object n, PLAsTiCC required participants to submit tables of classification posterior probabilities $0 \leqslant p(m|d_n) \leqslant 1$ where object $n = 1, \ldots, N$ belongs to class $m = 1, \ldots, M$, given the light-curve data d_n. The PLAsTiCC metric was a weighted log-loss

$$L \equiv -\sum_{m=1}^{M} w_m \sum_{n=1}^{N_m} \tau_{n,m} \ln[p(m|d_n)], \tag{1}$$

where

$$\tau_{n,m} \equiv \begin{cases} 0 & m_n \neq m \\ 1 & m_n = m \end{cases} \tag{2}$$

is an indicator variable for the true class. This particular metric strongly disfavors assigning very low probabilities to the classes that the classifier thinks are not the correct ones, implying that very low entropy classifications (i.e., with lots of zeros/small values in the vector of class probabilities) are disfavored.

Table 1
Table of Weights and Target Numbers for the Various Classification Types Included in PLAsTiCC

Class Name	Class ID Number	Weight
Point source μ-lensing	6	2
Tidal disruption event (TDE)	15	2
Eclipsing binary event (EBE)	16	1
Core-collapse supernova Type II (SN II)	42	1
Supernova Type Ia-x (SN Iax)	52	1
Mira variable	53	1
Core-collapse supernova Type Ibc (SN Ibc)	62	1
Kilonova (KN)	64	2
M dwarf	65	1
Supernova Type Ia-91bg (SN Ia-91bg)	67	1
Active galactic nucleus (AGN)	88	1
Supernova Type Ia (SN Ia)	90	1
RR Lyrae	92	1
Superluminous supernova (SLSN)	95	2
"Other" class	99	2

Note. Rare or interesting objects were upweighted to increase their contribution to the total metric.

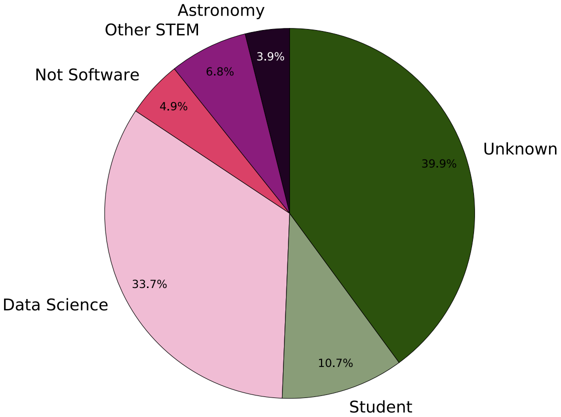


Figure 3. Demographic information for entrants to PLAsTiCC. Only a small fraction of 1315 participants with known demographics were part of the astronomy community. Many of the participants have some data science background. These data were determined from the Kaggle entries or GitHub user pages of the entrants, future studies will include self-identification information from participants.

The weights $w_m = J/N_M$, where $J = 1$ for most classes, but $J = 2$ for rare objects, as shown in Table 1. These weights were chosen by the PLAsTiCC team to discourage participants from classifying all objects as the most populous class, ensuring that appropriate attention was directed to the classification of rare classes that otherwise would not have much influence on a metric that gave equal weight to all objects. A detailed description of the PLAsTiCC metric may be found in Malz et al. (2019), along with an investigation of its assessment of archetypal classification pathologies.

2.4. The PLAsTiCC Competition

PLAsTiCC ran from 2018 September 28 until 2018 December 17, and received entries from 1094 teams around the world from a variety of different groups. The demographic information of the 1315 participants is summarized in Figure 3 and was determined from the Kaggle or GitHub profiles of the participants in each team. Some of the participants have subsequently deleted their accounts, or do not have complete profiles, and so Figure 3 provides an estimate of the group demographics. While a substantial fraction of participants has a background in data science/software, only a fraction of entrants had formal astronomy training. A reasonable subset of the PLAsTiCC participants were students. These included a group from the University of Warwick in the United Kingdom who entered PLAsTiCC as part of a class project.

PLAsTiCC led to publications by participating groups, including presentations of the winning avocado algorithm (Boone 2019) and that of the group "Day meets night" (Gabruseva et al. 2020). The data set itself has also been used for several publications on classification methods and anomaly detection approaches. Table 2 provides a table of publications from the PLAsTiCC team, and a subset of publications enabled by this challenge. The full list of 96 papers related to PLAsTiCC is given in an ADS link.[48]

Though participants submitted their classifications on the full PLAsTiCC test data set, the metric was evaluated on representative subsets whose distinctions were not made known to participants. The metric computed on one-third of the full data set was posted to a "public leaderboard," whereas the metric evaluated on the remaining two-thirds of the test data was the basis for the official ranking Kaggle used to award the cash prize. This procedure could be applied to subsets of the data to determine a classification "error," however Kaggle did not include any such error estimation in their application of the PLAsTiCC metric.

This approach of keeping data hidden from participants is in place to discourage "probing the leaderboard" (hereafter LB probing), wherein one's position on the leaderboard (LB) is used as a metric on its own to tweak a submitted algorithm until it is tuned to the LB performance. While this will lead to improved performance in one particular challenge/realization of the data, this approach does not yield algorithms robust to changes in data simulation, or more general intuition-building about the problem.

The distribution of submissions over time and their corresponding metric values are shown in Figure 4. Interest in the challenge grew over time as more participants joined. Kaggle restricts participating teams to a maximum of five submissions per day, a limitation that proved critical near the end of the challenge, when participants had to be selective in making submissions that would best inform final adjustments to their classifiers.

The performance of the three winning solutions are highlighted in the plot as colored lines. The horizontal clustering of the PLAsTiCC metric near scores of $\log[\text{weighted log}-\text{loss score}] \simeq 1$ and $\simeq 1.6$ (where the log-loss score is given in Equation (1)) can be seen in the one-dimensional histogram of scores on the right-hand panel of Figure 4 and could be a result of overfitting ("playing the leaderboard"), particularly where weighted combinations of classification probabilities for some objects are used to determine the probabilities of other objects (e.g., the "Other" class objects).

3. Solutions to PLAsTiCC

PLAsTiCC ended on 2018 December 17. The competition was won by Kyle Boone which is described in Boone (2019), with the second-place prize awarded to Mike and Silogram[49] and third place was awarded to "Major Tom, mamas & nyanp"[50]. The top three teams were selected to present their methods at a meeting with the PLAsTiCC team and the staff at Kaggle.

Figure 5 shows the performance of the top three entries according to the metrics considered in Malz et al. (2019), namely the weighted log-loss used in PLAsTiCC, an unweighted log-loss, and the Brier score (Brier 1950, equivalent to the mean square error of the prediction), normalized to the Kyle Boone score. The left panel shows the metrics evaluated on the full test set, rather than the one-third subset that formed the basis of the PLAsTiCC competition results, while the right panel performance when the "Other" class objects were excluded from the test data. The scores of the top three entries tighten in that case and the top two places are reversed, highlighting the importance of returning a classification probability for objects whose class was not present in the training set. All three top finishers were very close in score throughout the competition in both the public and the private

[48] The shortened search link is https://bit.ly/PLAsTiCC_Papers.

[49] The second-place solution is described in a Kaggle post (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75059).

[50] The third-place solution is described in two separate posts, one describing the CNN (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75116) and another describing the model including code repositories (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75131).

Table 2
A Selection of Publications Resulting from PLAsTiCC

Reference	Topic	Comment
The PLAsTiCC team et al. (2018)	PLAsTiCC Release Note	Main PLAsTiCC data note
Kessler et al. (2019a)	PLAsTiCC Models	Main PLAsTiCC publication
Malz et al. (2019)	PLAsTiCC Metric	Main PLAsTiCC publications
This work	PLAsTiCC Results	Main PLAsTiCC publications
Boone (2019)	Challenge solution	PLAsTiCC submission write-up
Gabruseva et al. (2020)	Challenge solution	PLAsTiCC submission write-up
Ishida et al. (2021)	Paper using PLAsTiCC to test classification method	Paper by PLAsTiCC member
Soraisam et al. (2020)	Paper using PLAsTiCC to test classification method	General paper
Sravan et al. (2020)	Paper using PLAsTiCC for imbalance tests	General paper
Dobryakov et al. (2021)	Paper using PLAsTiCC for imbalance tests	General paper
Hosenie et al. (2020)	Paper using PLAsTiCC for imbalance tests	General paper
Villar et al. (2020)	Paper using PLAsTiCC to test the method	General paper

Note. The papers are sorted into ones coming directly from the PLAsTiCC team, and publications enabled by the challenge and its associated simulation set.

LBs. We summarize their solutions and their relative performance in the subsections that follow.

3.1. Summary of Methods Used

The Kaggle discussion board was an excellent forum for the dissemination and discussion of methods used to provide solutions to PLAsTiCC. The entrants from the top 12[51] entrees discussed their solutions on the forum, and here we briefly summarize the elements used in their solutions in Tables 3 and 4 shows the classification algorithms that were implemented. An introductory explanation of the methods/concepts used in those algorithms is provided in Table 5.

Combining multiple approaches. The unifying characteristic of the solutions is the fact that they employed a mixture of classification approaches from neural networks to gradient boosting. The entrants often tried different combinations of approaches and adjusted their mixture within different iterations, effectively playing multiple classification methods "against" each other internally before submitting their solutions at each step. We refer to this as "Ensemble" in Table 3. Common among most solutions was the use of the LightGBM (Ke et al. 2017). Approaches varied when it came to feature selection. Some groups built on previous attempts by performing template light-curve fits to extract features, using known fitting software such as the SALT2 package (Guy et al. 2007) or the Bazin et al. (2011) model. Some used the more time consuming but less model-dependent Gaussian process (GP) fitting. Originally developed in the context of mining surveying (first published in Krige 1951), GP interpolation relies on the assumption that the interpolated values can be modeled by a Gaussian process, or one in which the process is defined by a prior covariance, or kernel. This interpolation allows one to easily map not only the interpolated values, but the uncertainty band around the interpolation without assuming any physical relationship between the variables. This makes it well suited to light-curve fitting. A more thorough discussion of the techniques used in geostatistics can be found in Chilès & Jean-Paul (2012).

A few groups described fitting for hundreds of features and then using feature importance ranking to reduce the feature space over which classification was performed.

Data augmentation. Central to the PLAsTiCC challenge was data augmentation, the supplementation of training data with new light curves derived through small changes to the original training set, and other approaches to mitigate nonrepresentativity of the training set and class imbalance in both the training and test sets. Several groups tried a range of approaches to augment the data and rectify the class imbalance, which are discussed in more detail in Section 5, where we illustrate the degeneracies between the "Other" class objects and the rest of the classes).

The "Other" class objects. One of the most popular methods of computing probabilities for the "Other" class was to probe the LB. All of the top models that provided information on their handling of "Other" class objects used LB probing. One approach ("Weighted Average") produced an "Other" class probability formed from a weighted combination of the supernovae class probabilities, and then determined the best weights by optimizing over their score or position on the LB. The "Tuned Combination" from the MTMN method (see Section 3.4) used a power law combination of supernovae probabilities and tuned the power and multiplicative factor using the LB. They also had two different functions for Galactic and extragalactic objects. Ahmet Erdem found which classes that "Other" class objects were not like (TDE, KN, AGN, SN Ia, SN Ia-91bg) and applied a constant multiplication factor found through LB probing depending on which class had the highest probability outside of those classes. Four of the top participants used a method that was shown on a kernel on Kaggle that involved scaling all the class probabilities with weights based on the LB probing from another user ("Probability Scaling").

Redshift. As described in Kessler et al. (2019a), the SED models input in PLAsTiCC are redshifted, peculiar velocity and lensing effects are applied and the SED is convolved with the Rubin passbands and "observed" with various telescope noise properties and calibrations applied. These operations can by construction not be rerun by participants in PLAsTiCC, and so any augmentation of the training data due to redshift effects was done at the light-curve level.

Dependence on simulated cadence. While the PLAsTiCC data were simulated for one proposed observing strategy for the LSST, we note that the features themselves will in general be strongly dependent on the particular cadence or observational strategy of a survey. We leave a full investigation of classification performance for different observing cadences to

[51] The seventh and 10th place entries did not provide their solutions on the Kaggle forums so they are not included here.

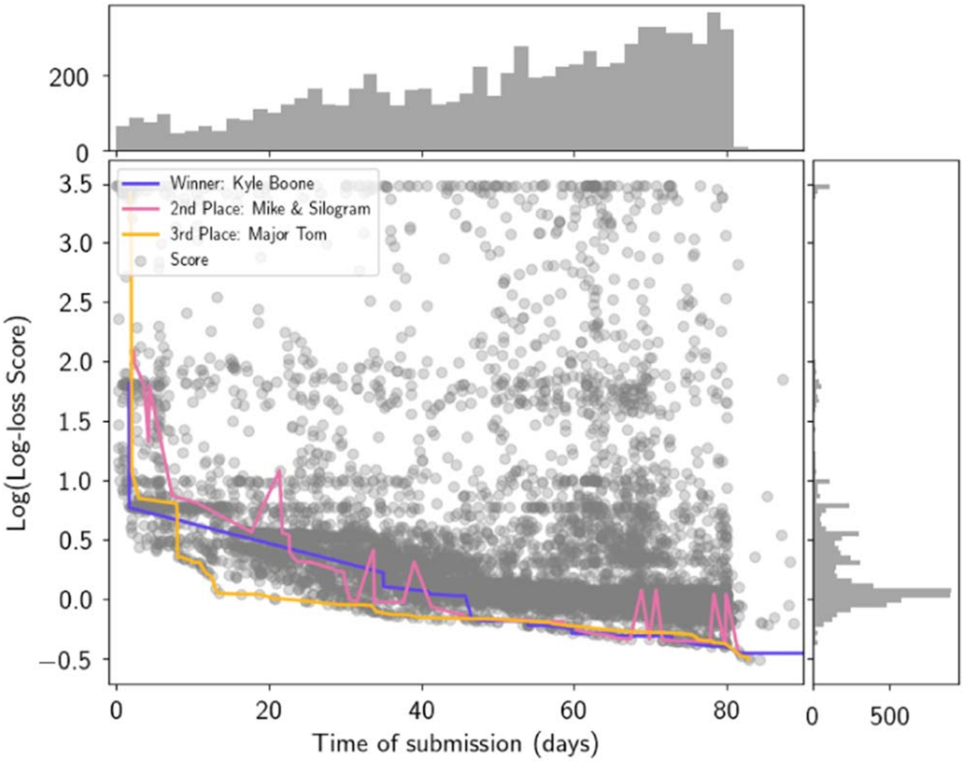


Figure 4. The distribution of the submissions to PLAsTiCC and the log of the PLAsTiCC metric score as a function of time since the start of the challenge relative to the start of the challenge. The three colored lines indicate the performance of the three prize-winning entries. As the challenge neared completion, the scores started to asymptote to metric scores of near unity. Note that the competition remains "live" to this day, hence the few submissions received after the closing deadline.

future work. In the subsections below we summarize in more detail the winning solutions.

3.2. Kyle Boone

The overall winner of PLAsTiCC was Kyle Boone, then a graduate student at the University of California at Berkeley. His method using GP light-curve fitting, data augmentation, and a gradient boost (Friedman 2001) classification strategy is described in detail in Boone (2019), which presented a slightly improved version of his submission to the PLAsTiCC competition. For the discussion here, we focus on the version of `avocado` that was submitted to PLAsTiCC. Boone was the only astronomer in the top three finishers, and his astronomical domain knowledge played a role in the choices he made for feature selection. For example, Boone extracted the color of the object at maximum light and computed the time to decline to 20% of the maximum light of the transient, which resulted in greater accuracy of classification of SN Ia-like objects in `avocado` than other challenge entries.

Premaximum observations are a useful indicator of the type of supernova. As a corollary, the lack of premaximum observations can serve as a useful flag of how discerning the data will be in separating out different classes, and hence are a measure of the accuracy in the classification of the simulations. `avocado` includes a flag for "incomplete" simulated light curves, which can help remove spurious classifications that might degrade the overall metric score.

Data augmentation also features strongly in the `avocado` classifier from Kyle Boone. The nonrepresentativity of PLAsTiCC led to a pristine but small training data set and a large (3.5 M-object sized) test data set, with much larger measurement uncertainties. Boone focused on augmenting the training data by shifting the data in time (without changing the shape of any light-curve features), and removing or "dropping" observations at random time positions. Boone also increased the scatter in the data to make it more representative of the test data quality, as shown in Figure 6. In order to supplement the training data, Boone generated high-z light curves from low-z analogs by degrading the signal-to-noise ratio as a function of distance. This data augmentation greatly improved the accuracy of his `avocado`. The change of light-curve colors with the redshifting of the SED was not included in `avocado`.

3.3. Mike and Silogram

The second-place winners of PLAsTiCC were from a pair of competitors (Mike and Silogram) who joined forces to produce an ensemble solution of seven recurrent neural network classifiers (RNN; Rumelhart et al. 1986) and two light gradient boosting machine (LightGBM; Ke et al. 2017) models. Rather than focus on selecting one classifier, the ensemble provided diversity across model space, to avoid overfitting the training data.

The RNN model was a one-layer bidirectional gated recurrent unit (GRU; Cho et al. 2014). The seven different RNN models used between 80 and 160 units. The input data for the RNN was the flux and its associated uncertainty, the intervals between the measurements, the wavelength, and the passband, which were included with "one-hot" encoding, where combinations of the data are only those with a single value at a time (see Harris & Harris 2012, for details on this method). The models reduced the dimensionality by downsampling the layers and then combined the max pooled layer with light-curve metadata, specifically the host galaxy photo-z and associated error, the Milky Way extinction, and flags for the observational area/field. This model was novel in the simplicity of the individual NNs, and did not involve the computational overhead

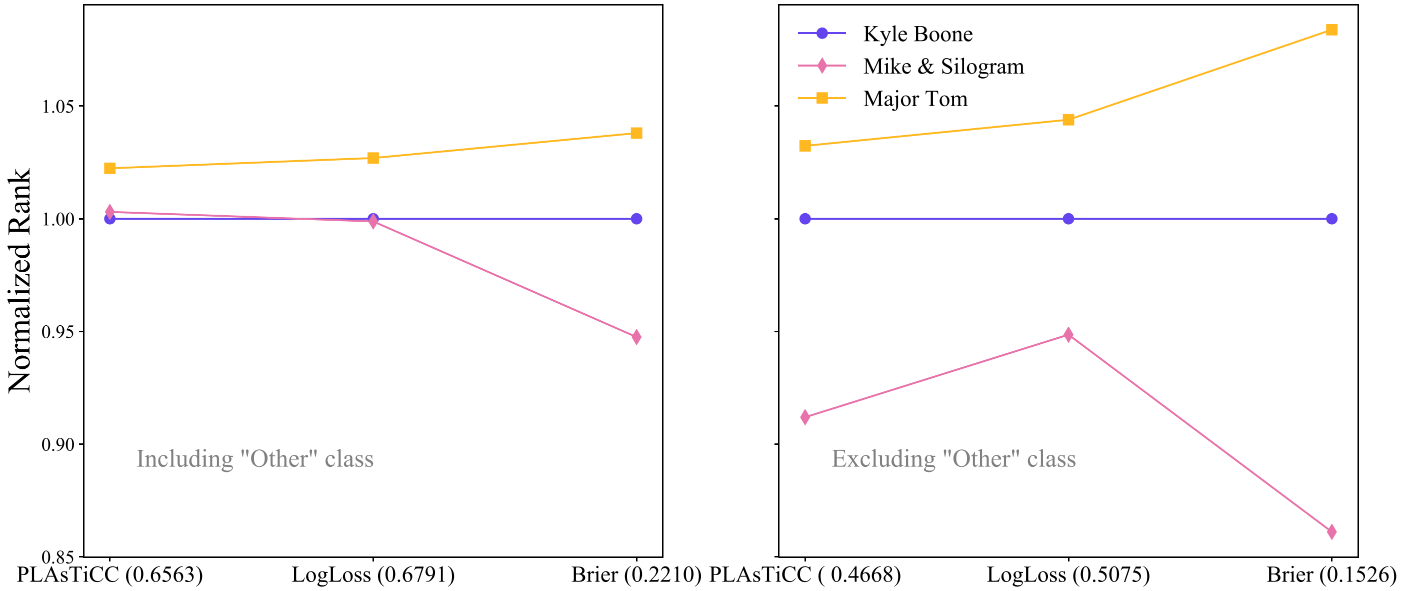


Figure 5. The performance of the top three entrants to the PLAsTiCC across three different metrics. In each case a lower score indicates better performance of the classifier. The left panel shows the metrics for the full test data set including the unseen "Other" class objects. All classifiers were asked to return probabilities for the "Other" class objects. The right panel shows the scores on the test data set excluding the "Other" class objects. The scores have been normalized to the score of Kyle Boone, indicated in parenthesis on the x-axis label. This plot illustrates how close the first- and second-place finish was, and the importance of the unknown objects on the final classifier performance. In the case where the "Other" class objects were not included, the first- and second-place positions would have been reversed.

associated with light-curve fitting. The RNN implementation was released as a kernel on the Kaggle site.[52]

As with the other solutions, the issues of data augmentation and feature selection were perhaps the most pressing for Mike and Silogram. This method was supplemented to include augmentation near the closing date for submissions to the competition and so did not have the opportunity to retrain on a comprehensively augmented data set. The Mike and Silogram method dropped 30% of the measurements and adjusted the redshifts and the flux of the training data.

In the LightGBM models, adjusting the flux values according to the photometrically derived redshift (photo-z) value helped "normalize" the flux, effectively correcting for the cosmological redshift of the object, which stretches spectrum of the object and dims its brightness. While the photo-z was provided in PLAsTiCC, the spectroscopic redshift (spec-z) is the more useful quantity as it is free of modeling errors. For both the RNN and the LightGBM models in this submission, the authors were able to build a separate model to predict the spec-z of the host galaxy from the other properties of the object. This approach is useful particularly to flag potential outliers in photo-z space (incorrectly predicted photo-z values).[53] This photo-z modeling will be one of the key systematic issues that Rubin observations will face, and any mitigation of this issue by classifiers is very valuable. The Mike and Silogram authors stated that they converted all time and wavelength-related features to redshift-independent versions before training the RNN.

3.4. Major Tom, Mamas, and Nyanp

The third-place solution also went to a combination of teams, "Major Tom" and "mamas," themselves made of individuals with different implementations that were combined. In addition, "nyamp" provided feature engineering expertise to the combined group, hereafter known as the MTMN method. The final solution was a one-dimensional convolutional neural network (CNN) with 256 8,5,3 convolution neurons followed by global max pooling as described above, with a stride of unity. The inputs to the CNN in this case was a series of 128-long vectors over different "channels." Six of these channels were constructed out of linear interpolated flux values (as a function of time) in the six bands, while another six were made of linearly interpolated vectors of flux multiplied by time, which forms a crude integral in the band. The final six channels in the set described the distance of the current point (in time) from the true input, effectively providing a vector describing the interpolation error.

Three different CNNs were constructed based on different metadata features sent to the multilayer perceptron (MLP) which was the basis for the backpropagation: a minimal CNN based on using only the galaxy photo-z, distance modulus, MJD, and flux error as metadata; a minimal CNN supplemented by the 16 best features obtained from features extracted in a separate CatBoost model framework; a minimal model supplemented by features obtained through template fitting from the Supernova Cosmology package (*sncosmo*; Barbary et al. 2016). The models are described in more detail on the Kaggle discussion boards,[54] and the CNN architecture is shown in Figure 1 of Wang et al. (2016).

4. Comparison of Top-ranked Classifiers by Additional Metrics

The PLAsTiCC metric was designed to evaluate submissions of classification probability mass functions (PMFs) without favoring any particular science case. For example,

[52] https://www.kaggle.com/zerrxy/plasticc-rnn

[53] This is described in a note on the Kaggle discussion board online (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75059#latest-462457).

[54] See, e.g., the online links for the third-place CNN (https://www.kaggle.com/yuval6967/3-place-cnn), third-place CatBoost (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75131), and third-place sncosmo entries (https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75222).

Table 3
Methods Employed by Classifiers in PLAsTiCC

Final Rank	Name	Public Score	Private Score	Class Imbalance Mitigation	Nonrepresentativity Mitigation	"Other" Class	Ensemble	Light-curve Fit
1	Kyle Boone	0.6706	0.6850	Light-curve augmentation	Data degradation	Weighted Average	No	GP
2	Mike and Silogram	0.6937	0.6993	⋯	Model spec-z, Data degradation	Probability Scaling	Yes	⋯
3	Major Tom mamas & nyanp	0.6804	0.7002	⋯	Model spec-z, Data degradation	Tuned Combination	Yes	SALT2
4	Ahmet Erdem	0.6913	0.7042	⋯	Model spec-z	Weight based on top prediction	Yes	Bazin
5	SKZ Lost in Translation	0.7397	0.7523	Flux error augmentation	Removed photo-z	Probability Scaling	Yes	⋯
6	Stefan Stefanov	0.7933	0.8017	Flux augmentation	Dropped light curve points, photo-z augmentation	⋯	Yes	⋯
8	rapids.ai	0.7922	0.8091	Pseudo label on test data	⋯	⋯	Yes	⋯ ⋯
9	Three Musketeers	0.7921	0.8131	⋯	Dropped light-curve points	Probability Scaling	Yes	⋯
11	Simon Chen	0.7948	0.8225	⋯	⋯	⋯	No	SALT2
12	Go Spartans!	0.8112	0.8265	⋯	⋯	Probability Scaling	Yes	Bazin

Note. The scores are taken from the Kaggle site, and illustrate how many methods were tuned to the public data/scores, and performed worse on the private data. The class imbalance techniques were used to address the fact that the different classes were present in very different numbers in the data set. Techniques include augmenting the light curve or fluxes, and adding approximate/pseudo "labels" to the test data to supplement the training data. The nonrepresentativity techniques modified data to account for the differences in the test and training data (e.g., that the test data were noisier and more sparse) by either degrading the data (error bars), removing light-curve points and modeling the spectroscopic redshift (spec-z) from the photometric redshifts (photo-z). The light-curve fitting techniques included Gaussian Process (GP) interpolation and template fitting (using the Bazin et al. 2011; Guy et al. 2007, light-curve fits). As discussed in Section 3.1, the entrants used a range of methods listed here to derive a probability for the "Other" class objects, and then probed the LB to optimize their procedure. In all columns above, if there is no entry for a given column, it indicates either that the model did not address the column or that the information was not provided in the write-ups presented on Kaggle.

cosmological classification efforts have traditionally focused on removing contamination of non-Ia objects to yield a "pure" SN Ia photometric sample, but metrics that reward only the purity of this class would be inappropriate for other science goals, such as novelty detection.

The caveats to the nominally science-agnostic global metric were twofold. The classification of rare and novel classes was encouraged by increasing the relative weight of the contribution of these classes to an overall metric that was otherwise evenly weighted across classes. This division into classes did not encode class hierarchy; had all supernova subclasses been combined into a single class, or had galactic and extragalactic classes been evaluated separately, it is possible that a metric would have favored different classifiers.

While the Kaggle metric was required to be a scalar so that a cash prize could be awarded to an overall winner, the resulting classifications in the form of submitted PMF catalogs are a much richer data set. In this section, we compare the top-ranked classifiers by some more nuanced and higher-dimensional metrics of estimated classification PMFs, including the precision recall and the Kullback–Leibler divergence (KLD). This list is not exhaustive, and several other metrics can be used to compare classifier performance (e.g., the Mathews correlation coefficient, the F-score, and the false detection rate). We leave a comprehensive comparison across a range of metrics for future work. Similarly, we do not focus on investigating metrics tuned to any specific science goals (e.g., cosmology constraints given a classified sample). We leave such an exploration for future work.

4.1. Pseudoconfusion Matrices

The confusion matrix is a classic metric of deterministic classifications (see e.g., Lochner et al. 2016, for one such astronomical example). To put the PLAsTiCC results into context for those familiar with the confusion matrix, we present an adaptation thereof to probabilistic classifications. For each classifier, we assign each light curve a deterministic classification corresponding to the mode (maximum) of its PMF. To remove the visual domination of class imbalance, we normalize the rows (corresponding to predicted classifications for a given true class) of a standard confusion matrix, whose cells contain counts rather than proportions relative to the number of objects in the true classes.

The pseudoconfusion matrices for the solutions of the top three classifiers are shown in Figure 7. The matrices show a distinct difference in the performance of the classifiers when distinguishing between different types of supernovae, which

Table 4
Classification Algorithms of the Top Entrants to PLAsTiCC, Sorted into Common Approaches

Name	Boosted Decision Trees			Neural Nets			
	LightGBM	CatBoost	XGBoost	NN	CNN	RNN	MLP
Kyle Boone	✓	×	×	×	×	×	×
Mike and Silogram	✓	×	×	×	×	✓	×
Major Tom, mamas & nyanp	✓	✓	×	×	✓	×	×
Ahmet Erdem	✓	×	×	✓	×	×	×
SKZ Lost in Translation	✓	×	×	×	×	✓	✓
Stefan Stefanov	×	×	×	✓	×	×	×
rapids.ai	✓	×	×	×	×	✓	✓
Three Musketeers	✓	✓	✓	×	✓	×	×
Simon Chen	✓	×	×	×	×	×	×
Go Spartans!	✓	×	✓	×	×	×	×

Note. A key result of the challenge was the observation of how different solutions combined various machine-learning techniques. The various methods are summarized in Table 5.

Table 5
A Brief Description of the Methods Used and Acronyms in PLAsTiCC

Method	Description	Reference
Random Forest	Also known as random decision trees, random forests are a method of classification that constructs many decision trees between different classes. The mode of multiple trees is computed to obtain a final classification.	Ho (1995)
Gradient Boosting	Gradient boosting produces a combined model from an ensemble of prediction models (often but not always comprised of decision trees). Each new tree is fit on a modified version of the original data set, and the weights of subsequent trees are adjusted according to which objects were difficult to classify by previous trees (upweighting difficult classifications).	Friedman (2001)
Light Gradient Boosting Machine (LightGBM)	LightGBM is a gradient-boosting framework originally developed by Microsoft, but is now open source. It is based on decision trees.	Ke et al. (2017)
CatBoost	CatBoost is a popular open-source algorithm for gradient boosting on decision trees. It includes an implementation of ordered boosting, and an algorithm for processing categorical features in data.	Prokhorenkova et al. (2017)
XGBoost	XGBoost is another gradient boosting algorithm, which performs tree boosting in parallel.	Chen & Guestrin (2016)
Neural network (NN)	A neural network mimics the way the human brain operates by connecting a system of connected nodes (neurons) connected by edges. The inputs to neurons are real numbers, and the output of each neuron is computed by some nonlinear function of its inputs. The separate neurons are combined with different weights.	See Schmidhuber (2015) for a recent review
Recurrent neural network (RNN)	This is a NN where node edges/connections have a direction to them. This means the RNN can display temporal dynamic behavior.	Rumelhart et al. (1986)
Gated recurrent unit (GRU)	A popular variant of RNN, GRU includes gating units which control the amount of information flow inside each recurrent unit (controlling when the hidden state in a unit should be updated).	Cho et al. (2014)
Multilayer perceptrons (MLP)	MLPs are NNs where each neuron in one layer is connected to all neurons in the next layer.	Rosenblatt (1963)
Convolutional neural network (CNN)	CNNs are regularized versions of MLPs and are commonly used on images.	Fukushima (1980)

are clustered in the top left-hand corner of the diagram. In addition, the degeneracy between the classification of true "Other" class objects (right column) and Type II supernovae (fourth row of the matrix) is shown in Figure 7. The degeneracy arises because three of the four novel objects included in the "Other" class had supernova-like light curves and were difficult to disentangle from the supernovae. In addition, some entrants formed a composite classification probability for this class from a weighted average combination of probabilities of the other classes (notably the SN Ibc and the SN Ia classes), leading to the correlation between the classification probabilities of specific objects and the "Other" class.

We show the confusion matrices considered without the "Other" class objects in Figure 8. The classifiers were not rerun

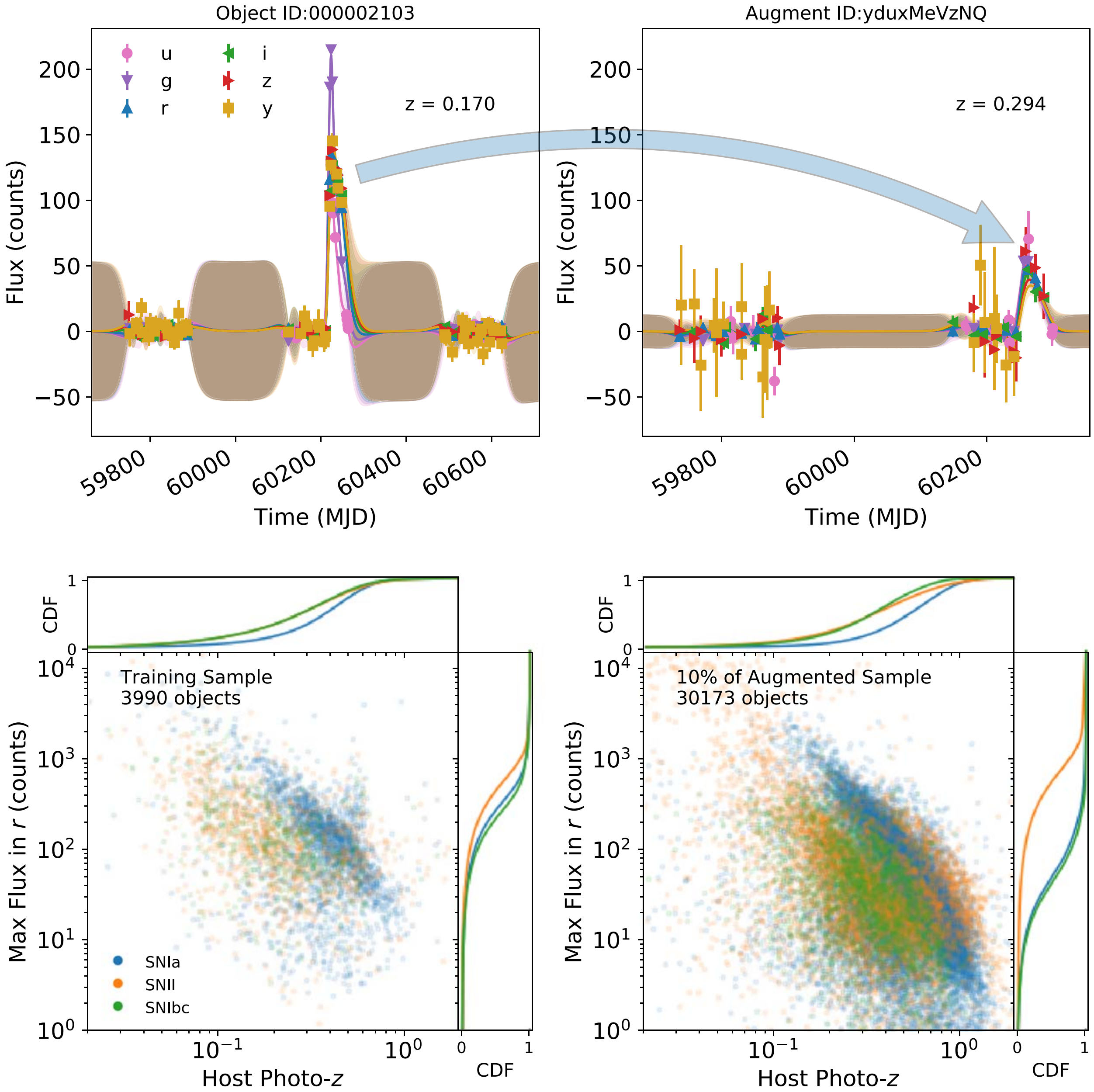


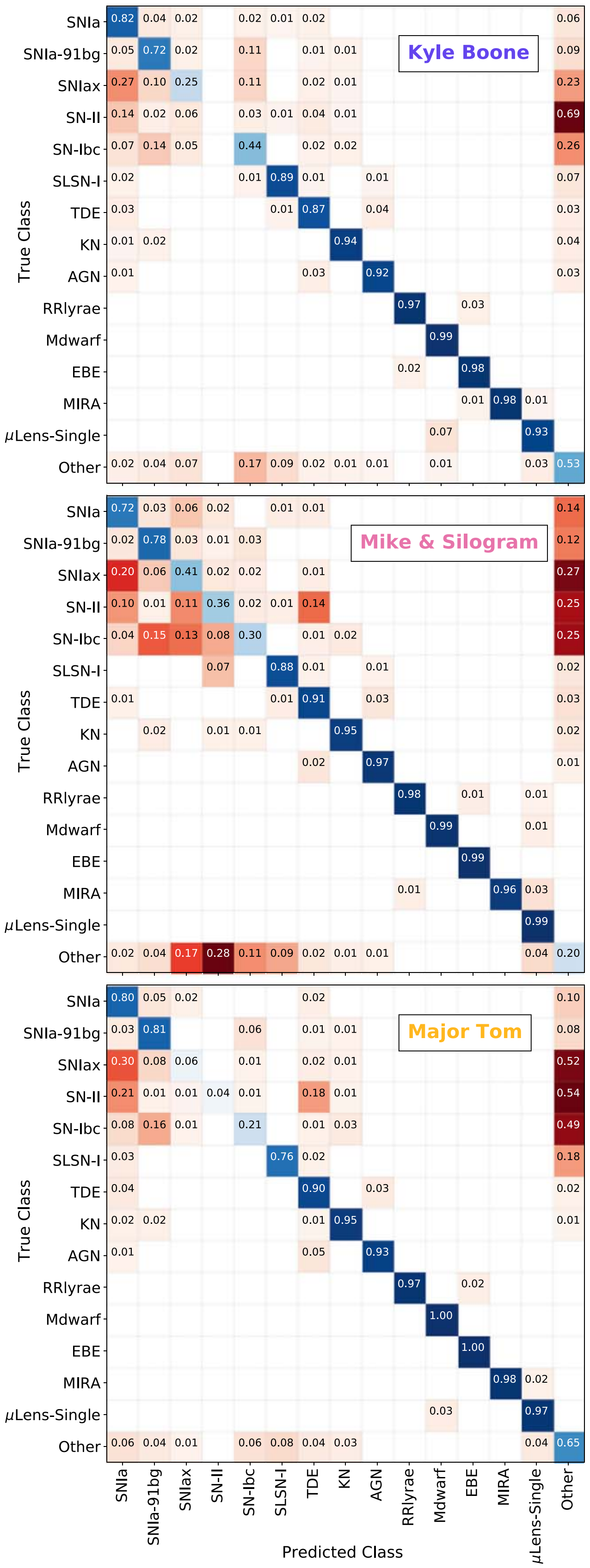
Figure 6. Data augmentation as part of the KB method. The light curves have been interpolated using GP interpolation, shown as shaded regions with the largest spread where there is no data. The top panel shows the augmentation in light-curve space, where points are removed, the object is moved to a higher redshift and the errors on the light-curve points are increased. The bottom panel illustrates the augmentation in feature space for a subset of the PLAsTiCC models. The augmented sample not only has a larger spread in the maximum r-band flux, one of the features used by the `avocado` classifier but extends each class to a higher redshift and helps increase the sample.

without the "Other" class objects. We removed the "Other" class objects from the submissions and renormalized the probabilities. None of the classifiers shown here attempted to classify "Other" class objects through their machine-learning algorithms but instead combined the probabilities from other classes. The validation classifier was also run without "Other" class objects such that it is a better comparison to these confusion matrices than the ones in Figure 7.

4.2. Distributions of Assigned Classification Probabilities

PLAsTiCC required submissions of posterior PMF vectors f^i for each light curve i satisfying $f^i(m) > 0$ and $\sum_{m=1}^{M} f^i(m) = 1$ for each class m out of M total classes. This requested data product comprises a much richer set of resultant data than solely deterministic metrics such as those related to the numbers of objects classified as true/false positives/negatives. The violin plots of Figure 9 depict the probability density functions (PDFs) of estimated PMF values across all predicted classes for each true class's light curves, illustrating more than just a summary statistic of the PDF of estimated posterior PMF values, such as the mode or mean. Though the per-light-curve covariance is lost, such a visualization still conveys some of the degeneracies between different true and predicted classes, like an expanded form of each row in a confusion matrix.

For a light curve of a given true class, one would hope that a classifier would assign it a high probability for being of that class and low probabilities for being all other classes. Two



Kyle Boone

True Class	SNIa	SNIa-91bg	SNIax	SN-II	SN-Ibc	SLSN-I	TDE	KN	AGN	RRlyrae	Mdwarf	EBE	MIRA	μLens-Single	Other
SNIa	0.82	0.04	0.02		0.02	0.01	0.02								0.06
SNIa-91bg	0.05	0.72	0.02		0.11		0.01	0.01							0.09
SNIax	0.27	0.10	0.25		0.11		0.02	0.01							0.23
SN-II	0.14	0.02	0.06		0.03	0.01	0.04	0.01							0.69
SN-Ibc	0.07	0.14	0.05		0.44		0.02	0.02							0.26
SLSN-I	0.02				0.01	0.89	0.01		0.01						0.07
TDE	0.03					0.01	0.87		0.04						0.03
KN	0.01	0.02						0.94							0.04
AGN	0.01						0.03		0.92						0.03
RRlyrae										0.97		0.03			
Mdwarf											0.99				
EBE										0.02		0.98			
MIRA												0.01	0.98	0.01	
μLens-Single											0.07			0.93	
Other	0.02	0.04	0.07		0.17	0.09	0.02	0.01	0.01		0.01			0.03	0.53

Mike & Silogram

True Class	SNIa	SNIa-91bg	SNIax	SN-II	SN-Ibc	SLSN-I	TDE	KN	AGN	RRlyrae	Mdwarf	EBE	MIRA	μLens-Single	Other
SNIa	0.72	0.03	0.06	0.02		0.01	0.01								0.14
SNIa-91bg	0.02	0.78	0.03	0.01	0.03										0.12
SNIax	0.20	0.06	0.41	0.02	0.02		0.01								0.27
SN-II	0.10	0.01	0.11	0.36	0.02	0.01	0.14								0.25
SN-Ibc	0.04	0.15	0.13	0.08	0.30		0.01	0.02							0.25
SLSN-I				0.07		0.88	0.01		0.01						0.02
TDE	0.01					0.01	0.91		0.03						0.03
KN		0.02		0.01	0.01			0.95							0.02
AGN							0.02		0.97						0.01
RRlyrae										0.98		0.01		0.01	
Mdwarf											0.99			0.01	
EBE												0.99			
MIRA										0.01			0.96	0.03	
μLens-Single														0.99	
Other	0.02	0.04	0.17	0.28	0.11	0.09	0.02	0.01	0.01					0.04	0.20

Major Tom

True Class	SNIa	SNIa-91bg	SNIax	SN-II	SN-Ibc	SLSN-I	TDE	KN	AGN	RRlyrae	Mdwarf	EBE	MIRA	μLens-Single	Other
SNIa	0.80	0.05	0.02				0.02								0.10
SNIa-91bg	0.03	0.81			0.06		0.01	0.01							0.08
SNIax	0.30	0.08	0.06		0.01		0.02	0.01							0.52
SN-II	0.21	0.01	0.01	0.04	0.01		0.18	0.01							0.54
SN-Ibc	0.08	0.16	0.01		0.21		0.01	0.03							0.49
SLSN-I	0.03					0.76	0.02								0.18
TDE	0.04						0.90		0.03						0.02
KN	0.02	0.02					0.01	0.95							0.01
AGN	0.01						0.05		0.93						
RRlyrae										0.97		0.02			
Mdwarf											1.00				
EBE												1.00			
MIRA													0.98	0.02	
μLens-Single											0.03			0.97	
Other	0.06	0.04	0.01		0.06	0.08	0.04	0.03						0.04	0.65

Figure 7. Pseudoconfusion matrices for the top three entries to PLAsTiCC. The metric scores for these entries were 0.68503 (Kyle Boone), 0.69933 (Mike and Silogram), and 0.70016 (Major Tom, mamas, and nyanp). The top entries are described fully in Tables 3 and 4.

panels of Figure 9 show that the winning classifier exemplifies this desirable behavior for true SN Ia and SN Ibc by assigning high probabilities for their true class and low probabilities for all other classes. The degeneracies between classes can be seen as PDFs of the estimated posterior PMFs where there is substantial density for a predicted class that is not the true class (i.e., the peak of the PDF is skewed away from zero for a class other than the true class). The remaining two panels of Figure 9 provide illustrations of this undesirable effect: objects with true SN Iax and SN Ia-91bg type are assigned comparable probabilities for several classes including the true class. Although the centering of the PDF density around zero PMF for some mismatched classes indicates that the winning classifier did indeed easily rule out those incorrect classes, there is also low PDF density at high PMF for the true class (consider that the PDF drops off sharply for probability >0.15 for the SN Iax in the top right panel of Figure 9), meaning that the classifier was unable to distinguish between those SN subtypes.

Figure 10 isolates the true KN to compare the performance of the top three classifiers to that of the naïve "validation" classifier, whose inclusion is facilitated by including only light curves evaluated by the validation classifier that also appeared in the PLAsTiCC data set and omitting the probabilities assigned to the "Other" class prior to renormalizing the probabilities assigned to the remaining classes. All of the top three classifiers were able to classify the true KN objects, as they assigned high PMF values for the predicted KN class and low PMF values for all other classes. The validation classifier, on the other hand, assigns nonzero PMF values over all the SN subtypes, effectively misclassifying KN as SN without being able to distinguish the SN subtype.

Figure 11 shows the performance of the top three classifiers in the "Other" class. All three top classifiers successfully isolated the "Other" class light curves from most of the named classes and for the most part, successfully assigned a higher posterior PMF value for the correct class. However, all three top classifiers also had high PDF density at nonzero PMF values for SN II and SN Ibc, meaning that they frequently allocated nontrivial probability to the possibility that "Other" class light curves were SN II and SN Ibc.

4.2.1. Classifier Disagreement

To compare classifier performance, we need to define the level of disagreement between the estimated posterior PMFs across classifiers and classes on the basis of information content. Since we do not have access to the true PMF values corresponding to the PLAsTiCC data set, we cannot measure the loss of information due to using an estimated PMF $g(x)$ as opposed to the true PMF $f(x)$. Instead, we can calculate the continuous KLD (cKLD) of the probability density functions (PDFs $g(x), f(x)$ from the approximating PMF values relative to those of the true PMF values (e.g., the "violin plots" from Figure 9–11):

$$\mathrm{cKLD}(f,\, g) = \int f(x)(\log[f(x)] - \log[g(x)])dx, \qquad (3)$$

where x is the set of real numbers between 0 and 1, and the KLD is measured in nats, the unit of base-e information. The appendix of Malz et al. (2018) is a good primer on the KLD including a demonstration to build intuition for the behavior of the KLD in

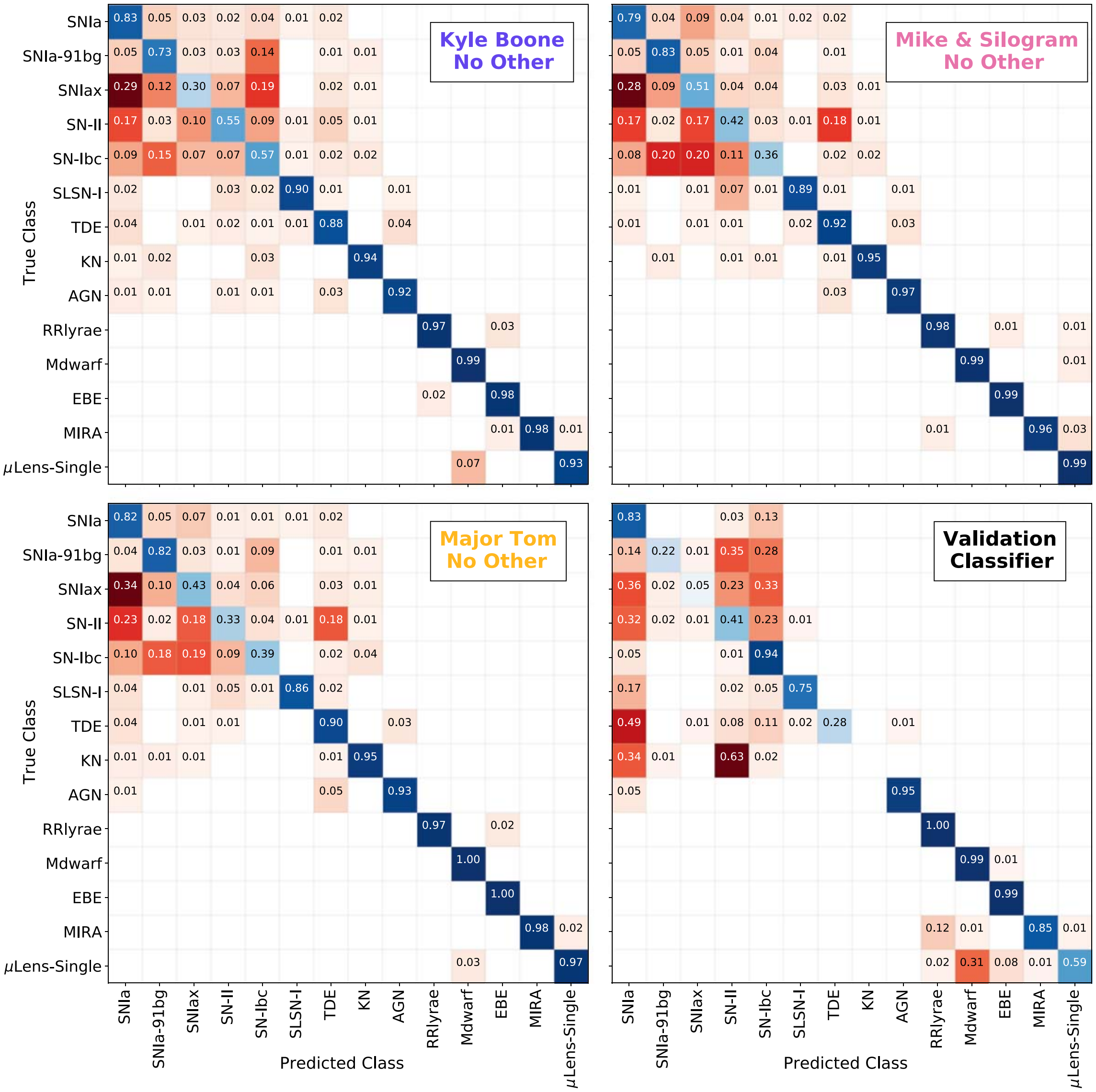


Figure 8. Similar to Figure 7, but in this case the probabilities assigned to "Other" class objects in the test set are omitted and the remaining probabilities are renormalized in the test data set. This illustrates more clearly the degeneracies between the known and the "Other" class objects. The large "subsuming" classification systematic visible in Figure 7 that arises since many classifiers provided a probability for "Other" class objects based on a weighted sum of probabilities from the known objects is no longer present (consider the SN II-"Other" class subsuming systematic in the Kyle Boone entry of Figure 7). We also show the results from the "validation classifier" run on the WFD data (without "Other" class objects).

limiting cases. While we lack a notion of a ground truth for comparison in the simulation, we can establish an ordering of probabilistic classifications "closer to the truth" by assuming that the Kaggle ranking metric correctly sorts classifiers by how close their submitted PMFs are to the (unknown) true PMFs. This assumption implies that the Kaggle scope is sufficient to define a directionality for evaluating the KLD:[55]

$$\mathrm{cKLD}(f^b, g^a) = \int f^b(x)(\log[f^b(x)] - \log[g^a(x)])dx, \quad (4)$$

where the index b indicates the better of the two classifiers being compared and index a indicates the approximating classifier ranked lower than b. For a specific class, a given classifier b might perform better than another a, even if a has a higher metric score than b overall/across all classes. We also note that the raw leaderboard data are available at https://www.kaggle.com/competitions/PLAsTiCC-2018/leaderboard.

To visualize the disagreements between classifiers, we consolidate the cKLD values of Equation (4) for each true class (e.g., SN-Ia, KN, etc.) and each predicted class pair as follows. For a given true class, we compute the summed cKLD of the PDFs of the predicted class PMFs, which is interpretable as a measure of the information loss for that true class.

To produce a scalar value (that is shown in Figure 12 we compute the weighted sum of these cKLD terms across the true classes using the original per-true-class weights of the original Kaggle metric (as defined in Table 1).

This integrated cKLD may then be interpreted as the value of the information loss due to using the approximating classifier rather than the one that is closer to the truth. This scalar value is shown for all pairs of classifiers in Figure 12, where the color indicates the scalar cKLD score for a pair of classifiers. The

[55] The validity of this assumption is equivalent to KLD$[i+1; i-1]$ = KLD $[i+1; i]$ + KLD$[i; i-1]$, which is formally untrue for this data set, but the assumption is useful in this context.

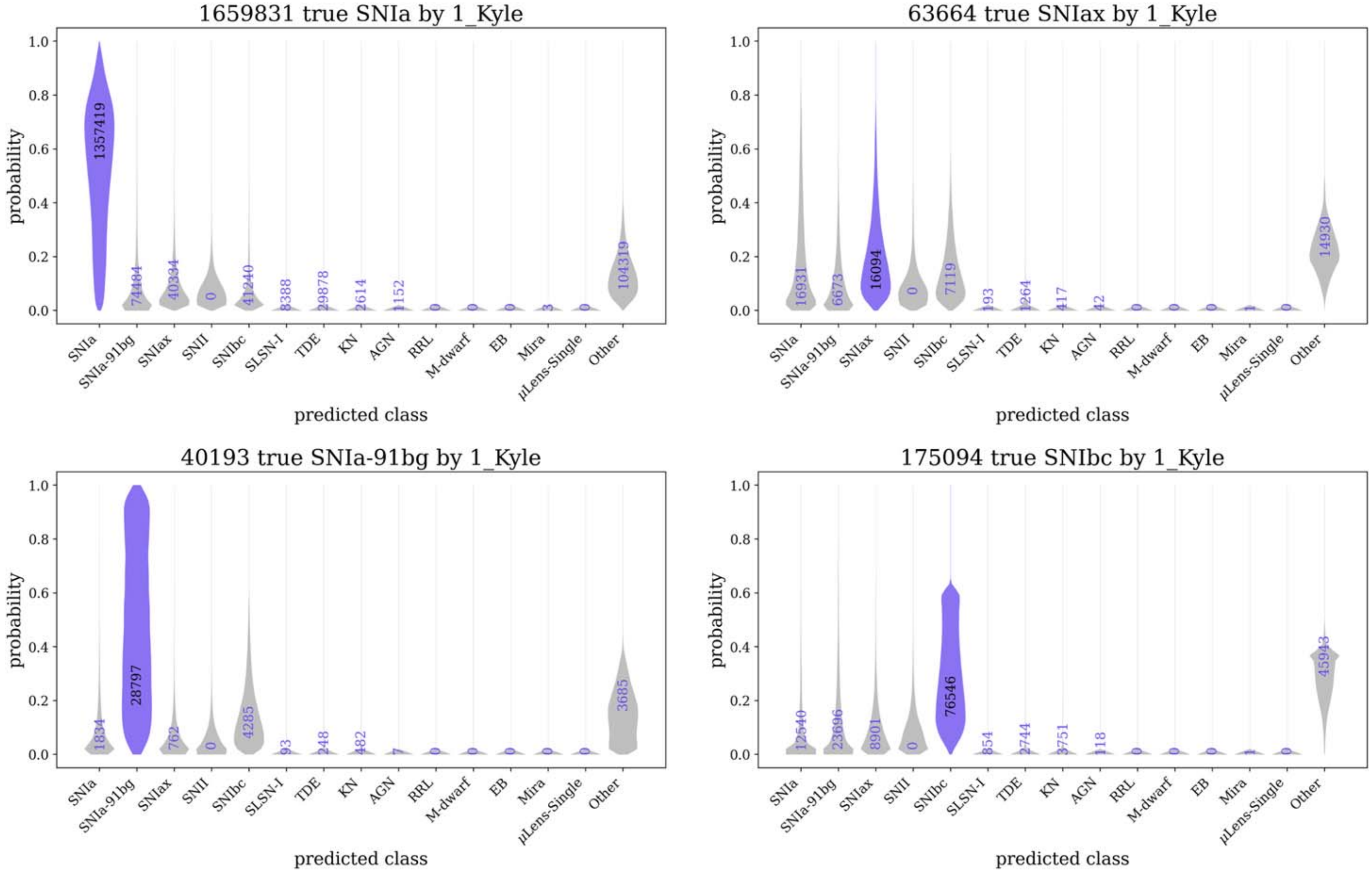


Figure 9. The estimated posterior probabilities (violins) for each predicted class (columns) among a given true class (panels) for the Kyle Boone method, where the assigned class corresponding to the true class is highlighted in color. Top row: the estimated probabilities for well-classified SN Ia and those of the less well-classified SN Iax. The assigned probability of being the true class for SN Ia has a long tail but a high mode of ∼0.8, whereas SN Iax is degenerate with the "Other" class type, whose mode is ∼0.2. Bottom row: the classification probabilities for SN Ia-91bg and SN Ibc (core-collapse SN). The estimated classification probabilities for true SN Ia-91bg objects reflects uncertainty; while roughly two-thirds of the true SN Ia-91bg objects have a mode corresponding to the predicted SN Ia-91bg class, their classification probabilities span the entire range of probability. The uncertainty of estimated probabilities for SN Ibc objects manifests differently; the probability for predicted class SN Ibc has a less pronounced tail, but there is considerable probability mass assigned to the predicted "Other" class. This is not surprising given that the Kyle Boone classifier determined the "Other" class classification probability from a combination of the probabilities of the SN Ia, SN Ibc, and SN II.

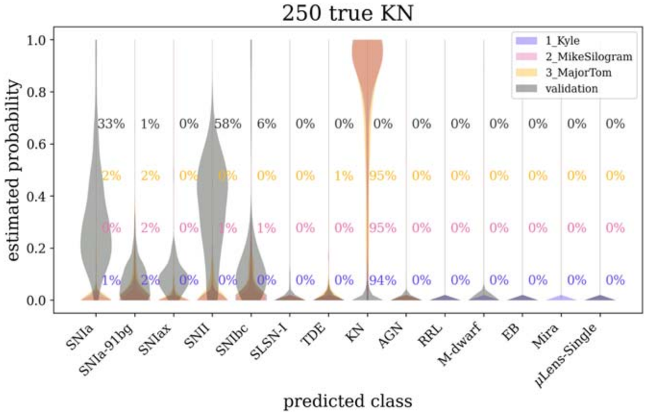


Figure 10. Performance of the "validation" classifier compared to the top three entries to PLAsTiCC. The naïve classifier does not correctly classify the rare KN objects, while a classifier more optimized to find rare objects through data augmentation is a very accurate classifier of these bright transients.

name of the class that is the largest contributor to the score is listed in the matrix entry, and the numerical value shows the percentage that that class contributes to the total cKLD score for the classifier pair.

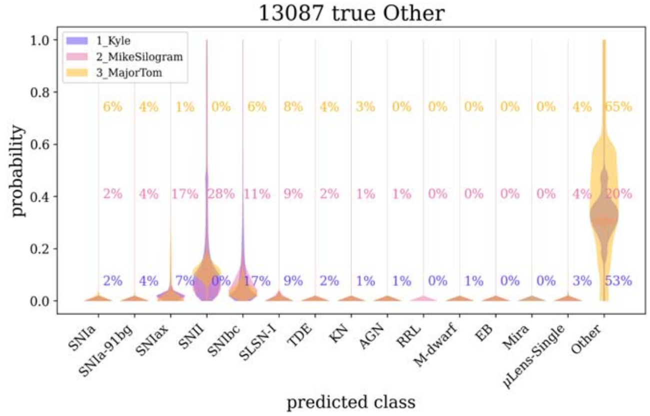


Figure 11. Performance of the top three classifiers for the "Other" class. The predicted probability distributions for the 13087 true type "Other" class objects. The numbers in different colors reflect how many objects were incorrectly classified as a given type by each of the three classifiers, and the shaded regions show the predicted probability distributions for each class. Note how the "Other" class is most often confused with SN II and SN Ibc. This occurred as a result of many of the classifiers generating their "Other" class probability as a combination of probabilities from the other classes.

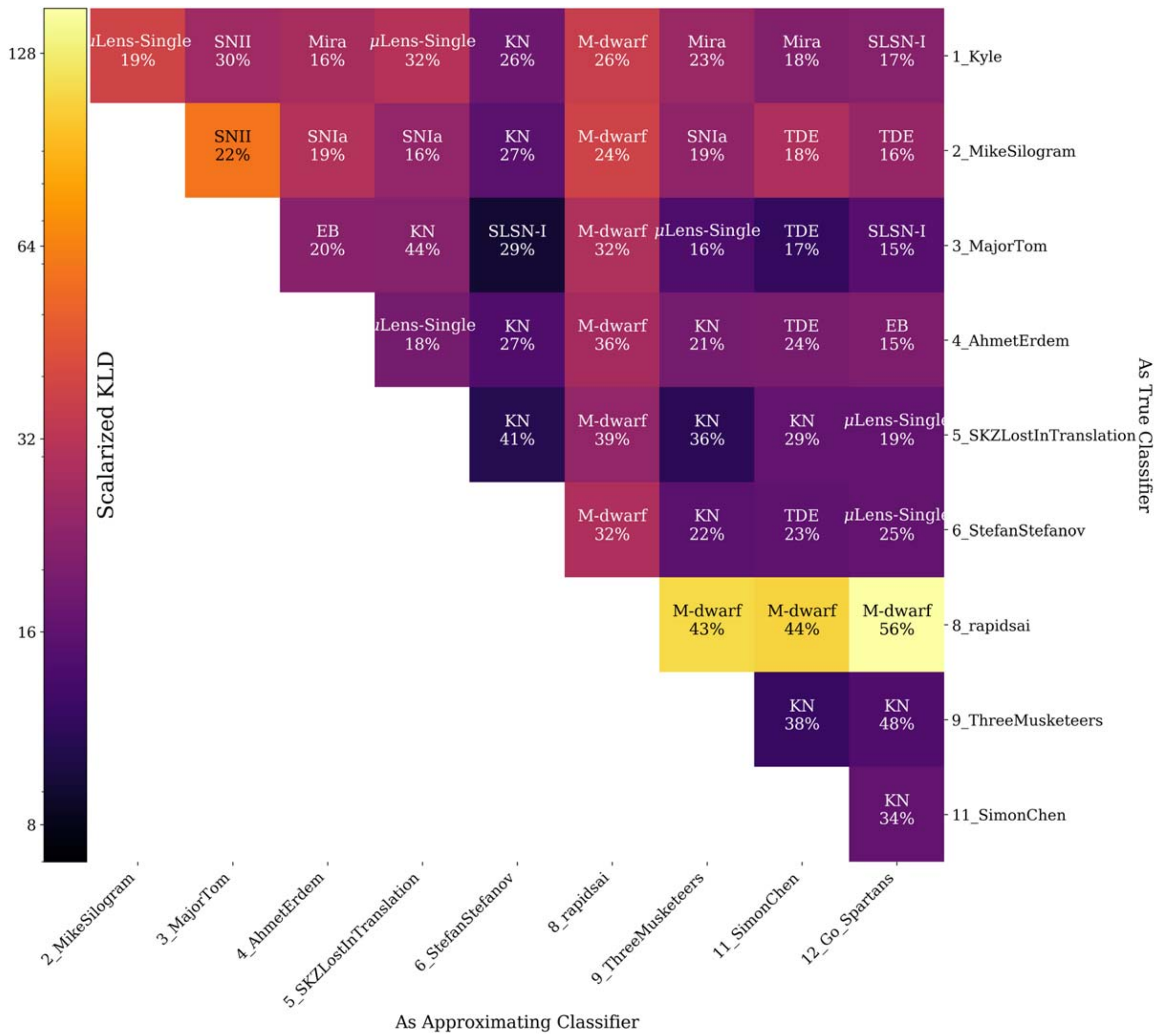


Figure 12. The disagreement matrix for the top classifiers. Colors indicate the disagreement metric between each classifier (x-axis) and all classifiers ranked better than it (y-axis), where darker cells correspond to lower Kullback–Leibler Divergence from the worse-ranked classifiers' probability densities of assigned posterior classification probabilities to those of the better-ranked classifier. The higher- and lower-ranked classifier pairs' cells are labeled with the class that was most different between the two classifiers and the percentage of the total disagreement between the two classifiers for which that class is responsible. The methods used by each classifier are summarized in Table 3.

The scalar KLD values of the pairs span 3 orders of base-e magnitude; however, if we exclude the eighth ranked classifier, then the pairwise information losses in the PDFs of PMFs only span two base-e orders of magnitude. The reason for this large range in KLD values is that the affected classifier ("8_rapidsai") assigned nearly deterministic classifications to the M-dwarf class, meaning the PMF values tended to be approximately 1 for one predicted class and approximately 0 for all other predicted classes. This choice induced long tails in the per-assigned class PDFs of true M-dwarf PMF values, to which the KLD is known to be sensitive, particularly when the PDF taken as closer to truth is affected. While the PLAsTiCC metric was designed with probabilistic classifications in mind, classifiers submitted both probabilistic (aka "soft") and deterministic ("hard") classifications to PLAsTiCC. Interestingly, Liu et al. compare the merits of deterministic and probabilistic classifier approaches (and suggest a unified/hybrid scheme that incorporates the best of both in Liu et al. 2011). They note that objects with smoother features may be better suited to soft classifications, while objects with sharp features may be better classified by hard algorithms, and could explain some of the performance of "8_rapidsai" on the M-dwarf objects. We leave an explicit comparison of the same classifiers on PLAsTiCC simulations using a deterministic metric for future study.

Figure 12 illustrates a few additional trends. The top two classifiers disagree more strongly with all other classifiers than the other classifiers do with one another. The class for which the discrepancy is greatest varies quite a bit among classifier pairs, but aside from the outlier classifier discussed above, the KN and Mira objects appear most discrepant frequently, indicating that the other ranked classifiers' strategies resulted in very different classification performance on those classes in particular.

4.3. Precision-recall Plots

Deterministic classifiers are often evaluated using functions of the true and false positive and negative rates. The true positive (TP) rate is the ratio of TP classifications to the sum of both the TP and false negative (FN) classifications $N_{\rm TP}/(N_{\rm TP}+N_{\rm FN})$. The false positive (FP) rate is the ratio of the FPs to the sum of both the FP and true negative (TN) rates.

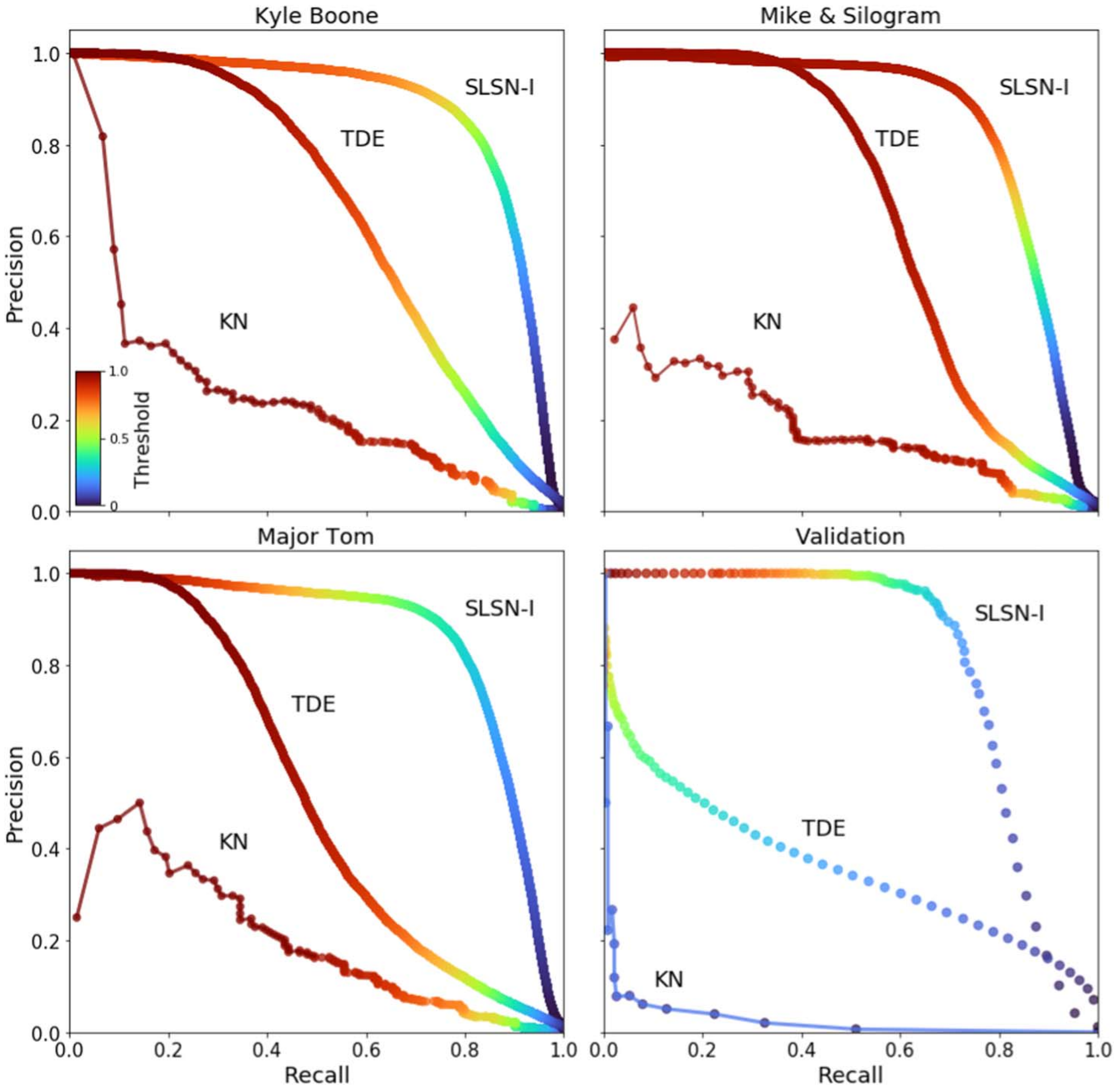


Figure 13. The precision-recall plots for the top three classifiers and the "validation" classifier as a function of threshold (color) for three classes of interest: KN, TDE, and SLSN.

One example of such a function is the receiver operating characteristic (ROC) curve, the TP rate as a function of the FP rate for a given classifier.

Another is the precision-recall (also referred to as "completeness-purity") plot of the precision $P = N_{\rm TP}/(N_{\rm TP} + N_{\rm FP})$ as a function of the recall $R = N_{\rm TP}/(N_{\rm TP} + N_{\rm FN})$. Recall is a performance measure of the TP compared to total actual positives, whether TP or FN, whereas precision is a performance measure of TP predictions compared to the total positive predictions (whether true or false). We choose to use precision-recall plots over ROC curves because they are appropriate for imbalanced data sets (see, e.g., Saito & Rehmsmeier 2015). The precision and recall are calculated for different values of the threshold of the classifier to map the performance of both as a function of the threshold. We require higher classification confidence with increasing threshold values. As the threshold increases, our precision increases, and recall (akin to sensitivity) decreases. Good performance on these plots would be indicated by having high precision and high recall, which would indicate that the classifiers are returning a lot of labels that are accurate.

A perfect classifier in these plots would have a precision of unity for all values of the recall. A bad classifier will output scores whose values are only slightly associated with the true outcome, and so will achieve high recall for low values of precision. Figure 13 shows examples of precision-recall plots

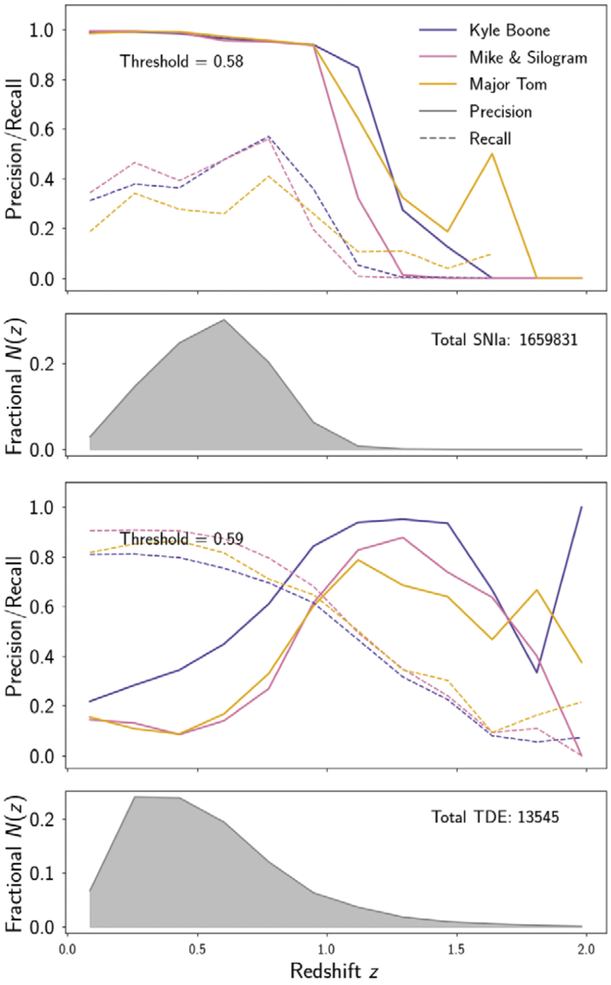


Figure 14. The precision and recall as a function of redshift for the SN Ia objects (top two panels) and TDE (bottom two panels). The average value of the threshold used across the classifiers is indicated in the text on the figure. The shaded distributions in each panel show the fractional distribution of the numbers of the transients as a function of redshift. The total number of transients in the sample is also indicated. The evolution with redshift of the metrics depends both on the redshift distribution of the objects and the evolution of the light curves with changing redshift.

for three different objects classified by the top three classifiers and the validation classifier. For the top three participants, SLSN are more well classified than KN or TDE but are not perfectly classified.

The classifiers also did approximately the same in precision and recall for TDEs. Of the transient events shown in this example, the KN are the most easily misclassified. This object has low precision, meaning that true positive classifications are a small fraction of total positive classifications. As the threshold value increases (needed to improve the precision performance), the recall (true positives relative to true positives and false negatives) drops relatively slowly as the number of false negatives is high for the KN class. The greatest contaminants to KN classification are SN Ia and SN II, which is seen in both the PLAsTiCC simulations and in the Dark Energy Survey observations (Doctor et al. 2017; Morgan et al. 2020). The MTMN and MS methods have a poor performance on this class throughout the plot whereas the KB method has good early "retrieval."

The SLSN has high values of recall for a range of thresholds, with $P > 0.9$ for all thresholds >0.7, indicating it is a well-classified object. TDE requires a high threshold to be correctly classified, but then plateau at constant P for threshold $\gtrsim 0.9$ for all classifiers. For the SLSN, the validation classifier performed significantly better than the top three classifiers, reaching $P \simeq 1$ for thresholds $\gtrsim 0.25$.

4.3.1. Redshift Evolution of Metrics

While the global precision-recall plots for various classifiers and types are shown in Figure 13, it is worth considering the evolution of the classifier performance with redshift. In Figure 14, we show the precision and recall of the top three classifiers as a function of redshift for a given threshold value of ∼0.6 for two object types, namely SNe Ia and TDEs. Also shown are the redshift distributions of the transients. The relative numbers of TDE peak at lower redshift (while there are still greater total numbers of SN Ia than TDE at all redshifts), leading to a peak in recall. The precision of the classifier peaks later, with the corresponding drop in the recall. This behavior is also seen in Figure 13 which shows the precision-recall curves as a function of threshold value rather than redshift. The ability to classify objects well with increasing redshift is driven in part by the redshift distribution of the sources themselves, and by the quality of data augmentation of the training set. As noted in Boone (2019), prioritization of completeness of a given survey as a function of redshift will lead to a rather different tuning of classifier performance.

5. Discussion of Challenges in the PLAsTiCC Data

The PLAsTiCC data set was rich in complexity and highly nonrepresentative.

Detection of novel objects. An additional class in the test set but not present in the training set, named the "Other" class, added the task of anomaly detection to the challenge. This unseen class proved challenging and led to confusion classifications that exhibit the "subsumed-to" behavior as described in Malz et al. (2019), where a first class is consistently mistaken for a second, but the second class is not commonly mistaken for the first. In Figure 8 we show the confusion matrices for the same entries as in Figure 7. In this case while the "Other" class objects were included in the test set, the classifiers did not return a classification for the objects (hence the predicted probabilities are set to zero for these objects). Hence while the true "Other" class objects have predicted probabilities of the other classes by construction, the subsuming probability systematic where true objects are confused for "Other" class objects is reduced. This can be most easily observed for the SN II objects in, e.g., the Kyle Boone classifier. We do not investigate separately the confusion between the four subclasses that made up the "Other" class, however future work and larger simulations could tease out the relative confusion of these interesting objects with well-known types.

This confusion between the other classes and the "Other" class objects is illustrated in Figure 9 through violin plots of the probability density of the posterior probability assigned to each predicted class, for a given true class. There are a few things worth noting in the violin plots. First, the distribution of probability extends to zero for some true objects, like the SN Ia-91bg which is shown in the bottom left panel of Figure 9, indicating that this true type is intrinsically uncertain. For the true type SN Ibc, the violin plots are even more interesting.

Here the violin plot for the "Other" class has a peak at P(type) =0.4 significantly away from zero. This arises from the fact that the assigned probability of the "Other" class in Boone's solution was derived as a combination of probabilities from SN Ia, SN Ibc, and SN II.

Feature selection. The participants in the PLAsTiCC challenge were unanimous that the biggest challenge they faced was the extraction of features from the large training and test sets of PLAsTiCC. Feature extraction is often the limiting factor in these classification challenges. Given the existence and use of the Bazin et al. (2011) light-curve fitting model (which is a four-parameter model that tends to zero flux as $t \to \pm\infty$) which is in many commonly used astronomical codes, many groups started with fits to the model, and trained their classifiers on the fit parameters. Consistently one of the top features was either the host galaxy photometric redshift or a proxy for the spectroscopic redshift modeled from the photometric redshift. Only two of the top 10 PLAsTiCC entries did not include an estimate of the redshift or the distance modulus. The maximum flux was also frequently used as a feature for classification. The Kyle Boone classifier included many signal-to-noise features whereas most other submissions traced the time-dependent movements of the light curves between the maximum and minimum flux. Many people used some form of the ratios of flux in different filters (essentially a "color") as classification features.

GP regression proved a more model-independent approach, but added the complexity of ensuring the correct choice of the kernel to do the regression.

Nonrepresentativity of the PLAsTiCC test data. The Rubin Observatory will map out the sky to a coadded magnitude depth of 27–28 and a single-visit depth of 24 in the various bands. This depth will yield a survey that has orders of magnitude more objects than current wide-field surveys. No current survey will be capable of providing distributions of classes as a function of redshift that will be representative of LSST. Any classifications performed based on low-z training sets will by definition be incomplete. To emphasize the importance of developing classifiers robust to this, the PLAsTiCC training set is nonrepresentative by design. The training set of 7486 objects was chosen to represent a probable spectroscopic survey made from a combination of objects from the Dark Energy Survey (DES) and other low-redshift surveys like the Foundations survey (Foley et al. 2018). The procedure for determining the spectroscopic training sample is outlined in Section 6.4 of Kessler et al. (2019a). The roughly 3 million objects in the test set not only had a different redshift distribution, but included the "Other" class as a way to test the robustness of classifiers to new objects and provide a sample for anomaly detection, as discussed above.

A small fraction of the 3 million test objects included spectroscopic "confirmation" and therefore had a small redshift error, while others had their redshifts assigned according to an example of the projected LSST photometric redshift performance. A small bias to the photo-z values is introduced as a result of the broadband data and training process. Many groups used the photo-z and/or the flux information to determine a proxy for the more informative spectroscopic redshift, which was used to determine an effective distance for the objects. These are labeled as "Model spec-z" in Table 3. We leave it to future work to investigate the impact the photo-z bias had on classification performance explicitly. Several groups performed data augmentation to address some of the issues with nonrepresentativity of the data by resampling in time to match the cadence of the test data, and adding noise to the training data that was similar to the (larger) scatter in the test data, as summarized in Table 3.

Class imbalance in PLAsTiCC. In addition to the test data being nonrepresentative of the training data set, the classes within the training and test data sets had significantly different proportions relative to the total population of objects. This class imbalance reflects true imbalances in nature, and was hence a designed feature of PLAsTiCC. Successful classifiers accounted for these imbalances by oversampling rare light curves, augmenting fluxes and errors, or including test data with pseudo labels. The approaches employed by the various classifiers are shown in Table 3.

6. Conclusion

The PLAsTiCC studies evaluate the classification probabilities derived by a range of methods for classifying a heterogeneous assortment of astronomical transients and variables. While previous challenges and data sets have focused on classifier performance related to one type of object or science case, PLAsTiCC represents the most complete simulation of future photometric surveys to date and challenged the community to take into account degeneracies between different types of interesting objects without a unifying science application and to prepare for future data that will contain novel/unknown objects that require classification and flagging for observational follow-up.

Accurate classification of transients in future surveys remains one of the key challenges for time domain observational astrophysics. While astronomers with domain knowledge are able to develop classification/selection methods tailored to individual science cases, the creativity and flexibility of the data science entries to PLAsTiCC suggest a collaborative future across disciplines will be key to making progress.

The PLAsTiCC challenge yielded a few key insights:

1. data augmentation of sparse nonrepresentative training data is key to accurate performance on the full test set;
2. a metric weighted by the number of objects in a class ensures that classification performance in the most populous class does not dominate overall performance;
3. data fitting and feature selection are the most computationally time-consuming parts of any classification challenge;
4. one successful approach is pulling out a larger set of features over which to train, and then determining the feature importance before classifying the objects on the smaller subset of features;
5. objects with similar light-curve shapes are broadly degenerate (e.g., SN Iax and SN Ia) and may present challenges to future surveys, motivating the inclusion of additional information about the host galaxy, where applicable;
6. no strong preference for any one machine-learning architecture was found; successful classification procedures often combine a range of methods from neural networks to tree-based approaches;
7. domain knowledge helped mainly in identifying more useful features over which to train, which was an asset,

given that this was often the most time-consuming part of the classification;

8. all classifiers struggled to classify objects with low frequency in the training data, or novel (unknown) objects

Classifying objects from their photometric information alone will unlock the potential of using these interesting objects in a range of science applications.

This challenge assumed that the full light curves for the objects were available. Long-lived transients can be followed up with ground-based resources within a few days of detection in a photometric survey. For some of the most interesting objects, however, waiting even a few days will mean losing valuable information about the evolution of the transient. As such, early classification of these bright objects becomes essential. We did not focus on the problem of early classification in PLAsTiCC, but the next step in using these simulations to prepare for the LSST will be to provide an extra performance score to classifiers capable of rapidly classifying transient sources, based on both their light curves and any additional data about their environment and location. This may prove more challenging or prone to bias than classification using the full light curve.

Acknowledgments

This paper has undergone internal review in the LSST Dark Energy Science Collaboration.

Author contributions are listed below.

R. Hložek: data curation, formal analysis, funding acquisition, investigation, project administration, software, supervision, validation, visualization, writing–editing, lead writing–original draft.

K. Ponder: software development and implementation for validation, testing, visualization, and analysis of results, writing–editing, writing–original draft.

A.I. Malz: conceptualization, formal analysis, investigation, methodology, visualization, writing–editing, writing–original draft.

M. Dai: software development and implementation for validation, testing, analysis of results, writing–editing.

G. Narayan: simulation, software, example kit, validation, visualization, writing–editing, writing–original draft.

E.E.O. Ishida: software development and implementation for validation, testing, Kaggle liaison, analysis of results, writing–editing.

T. Allam Jr: investigation and development of software for simulation validation.

A. Bahmanyar: investigation, implementation of metric methodology and software.

X. Bi: participated in the PLAsTiCC competition, released code publicly for analysis.

R. Biswas: conceptualization, methodology, software for validation and testing, writing–editing.

K. Boone: participated in the PLAsTiCC competition, released code publicly for analysis.

S. Chen: participated in the PLAsTiCC competition, released code publicly for analysis.

N. Du: participated in the PLAsTiCC competition, released code publicly for analysis.

A. Erdem: participated in the PLAsTiCC competition, released code publicly for analysis.

L. Galbany: implementation of validation software, writing–editing.

A. Garreta: participated in the PLAsTiCC competition, released code publicly for analysis.

S.W. Jha: conceptualization, simulation validation.

D. Jones: conceptualization, simulation validation, writing–editing.

R. Kessler: simulation development and validation testing, data generation, writing–editing.

M. Lin: participated in the PLAsTiCC competition, released code publicly for analysis.

J. Liu: participated in the PLAsTiCC competition, released code publicly for analysis.

M. Lochner: conceptualization, simulation validation, writing–editing.

A.A. Mahabal: conceptualization, simulation validation.

K.S. Mandel: conceptualization, simulation validation, writing–editing.

P. Margolis: participated in the PLAsTiCC competition, released code publicly for analysis.

J.R. Martínez-Galarza: conceptualization, simulation validation, writing–editing.

J.D. McEwen: conceptualization, simulation validation.

D. Muthukrishna: conceptualization, simulation validation, results analysis, and testing, writing–editing.

Y. Nakatsuka: participated in the PLAsTiCC competition, released code publicly for analysis.

T. Noumi: participated in the PLAsTiCC competition, released code publicly for analysis.

T. Oya: participated in the PLAsTiCC competition, released code publicly for analysis.

H.V. Peiris: conceptualization, supervision, funding acquisition, writing–editing.

C.M. Peters: conceptualization, simulation validation.

J.F. Puget: participated in the PLAsTiCC competition, released code publicly for analysis.

C.N. Setzer: conceptualization, simulation validation.

Siddhartha: participated in the PLAsTiCC competition, released code publicly for analysis.

S. Stefanov: participated in the PLAsTiCC competition, released code publicly for analysis.

T. Xie: participated in the PLAsTiCC competition, released code publicly for analysis.

L. Yan: participated in the PLAsTiCC competition, released code publicly for analysis.

K.-H. Yeh: participated in the PLAsTiCC competition, released code publicly for analysis.

W. Zuo: participated in the PLAsTiCC competition, released code publicly for analysis.

The Photometric LSST Astronomical Time-series Classification Challenge data set generation relied on numerous members of the astronomical community to provide models of astronomical transients and variables. The Photometric LSST Astronomical Time-series Classification Challenge is made possible through an intercollaboration agreement between LSST's Dark Energy Science Collaboration and the Transient and Variable Stars Science Collaboration. These models are described in Kessler et al. (2019a). The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science and Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne–France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515. We are grateful to Federica Bianco and the staff at NYU for hosting the PLAsTiCC team meeting. The awards for the PLAsTiCC contest were generously supported by Kaggle. We thank the staff at Kaggle, particularly Elizabeth Park, Maggie Demkin, and Sohier Dane, for their help in making the contest a success. We thank all participants in PLAsTiCC. We thank Federica Bianco, Francisco Förster Burón, Chad Schafer, and Melissa Graham for providing an internal review of this document for the DESC and TVS collaborations, and thank Seth Digel for comments on this draft. We thank the anonymous referee for the comments that greatly improved the manuscript. This work was supported by an LSST Corporation Enabling Science grant. The financial assistance of the National Research Foundation (NRF) toward this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. UK authors acknowledge funding from the Science and Technology Funding Council. Canadian coauthors acknowledge support from the Natural Sciences and Engineering Research Council of Canada. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. The authors at the University of Toronto acknowledge that the land on which the University of Toronto is built is the traditional territory of the Haudenosaunee, and most recently, the territory of the Mississaugas of the New Credit First Nation. They are grateful to have the opportunity to work in the community, on this territory. R. H. acknowledges support from CIFAR, and the Azrieli and Alfred. P. Sloan Foundations. A.I. M. acknowledges support during this work from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. A.I.M. is a member of the LSST Interdisciplinary Network for Collaboration and Computing (LINCC) Frameworks team; LINCC Frameworks is supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Institute of Astrophysics (VIA). K.A.P. acknowledges support from the Berkeley Center for Cosmological Physics; the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; and U.S. Department of Energy Office of Science under Contract No.DE-AC02-76SF00515. E.E.O.I. acknowledges support from CNRS 2017 MOMENTUM grant. L.G. acknowledges financial support from the Spanish Ministerio de Ciencia e Innovacíon (MCIN), the Agencia Estatal de Investigacíon (AEI) 10.13039/501100011033, and the European Social Fund (ESF) "Investing in your future" under the 2019 Ramón y Cajal program RYC2019-027683-I and the PID2020-115253GA-I00 HOST-FLOWS project, from Centro Superior de Investigaciones Científicas (CSIC) under the PIE project 20215AT016, and the program Unidad de Excelencia María de Maeztu CEX2020-001058-M. J.R.M.G acknowledges funding from the Chandra Science Research project code 40716015HH4600, under a contract with NASA. D.O.J. acknowledges the support provided by NASA Hubble Fellowship grant HST-HF2-51462.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. D.O.J. also acknowledges support from the Gordon and Betty Moore Foundation. This work has been partially supported by the Spanish grant PGC2018-095317-B-C21 within the European Funds for Regional Development (FEDER). C.N.S., R.B., and H.V.P. were partially supported by the research environment grant "Gravitational Radiation and Electromagnetic Astrophysical Transients (GREAT)" funded by the Swedish Research Council (VR) under Dnr 2016-06012. R.B. and H.V.P. were additionally supported by the research project grant "Understanding the Dynamic Universe" funded by the Knut and Alice Wallenberg Foundation under Dnr KAW 2018.0067. M.D. acknowledges support from the Horizon Fellowship at Johns Hopkins University. At Rutgers University (M.D., S.W.J.), this research was supported by NSF award AST-1615455 and DOE award DE-SC0011636. A.G acknowledges ERC grant PCG- 336983, Basque Government grant IT974-16, and MINECO grant PID2019-107444GA-I00.

ORCID iDs

R. Hložek https://orcid.org/0000-0002-0965-7864
A. I. Malz https://orcid.org/0000-0002-8676-1622
K. A. Ponder https://orcid.org/0000-0002-8207-3304
M. Dai https://orcid.org/0000-0002-5995-9692
G. Narayan https://orcid.org/0000-0001-6022-0484
E. E. O. Ishida https://orcid.org/0000-0002-0406-076X
R. Biswas https://orcid.org/0000-0002-5741-7195
K. Boone https://orcid.org/0000-0002-5828-6211
N. Du https://orcid.org/0000-0003-2855-7452
L. Galbany https://orcid.org/0000-0002-1296-6887
S. W. Jha https://orcid.org/0000-0001-8738-6011
D. O. Jones https://orcid.org/0000-0002-6230-0151
R. Kessler https://orcid.org/0000-0003-3221-0419
M. Lochner https://orcid.org/0000-0003-2221-8281
A. A. Mahabal https://orcid.org/0000-0003-2242-0244
K. S. Mandel https://orcid.org/0000-0001-9846-4417
J. R. Martínez-Galarza https://orcid.org/0000-0002-5069-0324
D. Muthukrishna https://orcid.org/0000-0002-5788-9280
H. V. Peiris https://orcid.org/0000-0002-2519-584X
C. N. Setzer https://orcid.org/0000-0002-7439-2735

References

Barbary, K., Barclay, T., Biswas, R., et al. 2016, SNCosmo: Python library for supernova cosmology, Astrophysics Source Code Library, ascl:1611.017
Bazin, G., Ruhlmann-Kleider, V., Palanque-Delabrouille, N., et al. 2011, A&A, 534, A43
Biswas, R., Daniel, S. F., Hložek, R., et al. 2020, ApJS, 247, 60
Boone, K. 2019, AJ, 158, 257
Brier, G. W. 1950, MWRv, 78, 1
Chen, T., & Guestrin, C. 2016, arXiv:1603.02754
Chilès, J. P., & Delfiner, P. 2012, Geostatistics: Modeling Spatial Uncertainty (2nd ed.; New York: Wiley)
Cho, K., van Merrienboer, B., Gulcehre, C., et al. 2014, arXiv:1406.1078
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, Msngr, 175, 3

Dobryakov, S., Malanchev, K., Derkach, D., & Hushchyn, M. 2021, A&C, 35, 100451

Doctor, Z., Kessler, R., Chen, H. Y., et al. 2017, ApJ, 837, 57

Foley, R. J., Scolnic, D., Rest, A., et al. 2018, MNRAS, 475, 193

Friedman, J. H. 2001, AnSta, 29, 1189

Fukushima, K. 1980, Biol. Cybern., 36, 193

Gabruseva, T., Zlobin, S., & Wang, P. 2020, JAI, 9, 2050005

Guy, J., Astier, P., Baumont, S., et al. 2007, A&A, 466, 11

Harris, D., & Harris, S. 2012, Digital Design and Computer Architecture (San Francisco, CA: Morgan Kaufmann)

Ho, T. K. 1995, in Proc. of 3rd Int. Conf. on Document Analysis and Recognition (Los Alamitos, CA: IEEE Computer Society), 278

Hosenie, Z., Lyon, R., Stappers, B., Mootoovaloo, A., & McBride, V. 2020, MNRAS, 493, 6050

Ishida, E. E. O., Kornilov, M. V., Malanchev, K. L., et al. 2021, A&A, 650, A195

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Ke, G., Meng, Q., Finley, T., et al. 2017, Advances in Neural Information Processing Systems 30, ed. I. Guyon et al. (Long Beach, CA: NeurIPS), 3146, https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kessler, R., Bassett, B., Belov, P., et al. 2010, PASP, 122, 1415

Kessler, R., Bernstein, J. P., Cinabro, D., et al. 2009, PASP, 121, 1028

Kessler, R., Brout, D., D'Andrea, C. B., et al. 2019b, MNRAS, 485, 1171

Kessler, R., Narayan, G., Avelino, A., et al. 2019a, PASP, 131, 094501

Krige, D. G. 1951, Master's thesis, Univ. Witwatersrand

Liu, Y., Zhang, H. H., & Wu, Y. 2011, J. Am. Stat. Assoc., 106, 166

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201

Malz, A. I., Hložek, R., Allam, T. J., et al. 2019, AJ, 158, 171

Malz, A. I., Marshall, P. J., DeRose, J., et al. 2018, AJ, 156, 35

Morgan, R., Soares-Santos, M., Annis, J., et al. 2020, ApJ, 901, 83

Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, ApJS, 236, 9

PLAsTiCC-Modelers, Kessler, R., Narayan, G., et al. 2019, Libraries & Recommended Citations for Using PLAsTiCC Models, v1, Zenodo, doi:10.5281/zenodo.2612896

Prokhorenkova, L., Gusev, G., Vorobev, A., Veronika Dorogush, A., & Gulin, A. 2017, arXiv:1706.09516

Rosenblatt, F. 1963, Am. J. Psychol., 76, 705

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Natur, 323, 533

Saito, T., & Rehmsmeier, M. 2015, PLoSO, 10, e0118432

Schmidhuber, J. 2015, NN, 61, 85

Soraisam, M. D., Saha, A., Matheson, T., et al. 2020, ApJ, 892, 112

Sravan, N., Milisavljevic, D., Reynolds, J. M., Lentner, G., & Linvill, M. 2020, ApJ, 893, 127

Swann, E., Sullivan, M., Carrick, J., et al. 2019, Msngr, 175, 58

The PLAsTiCC team, Allam, T. J., Bahmanyar, A., et al. 2018, arXiv:1810.00001

Villar, V. A., Hosseinzadeh, G., Berger, E., et al. 2020, ApJ, 905, 94

Wang, Z., Yan, W., & Oates, T. 2016, arXiv:1611.06455