Draft version July 29, 2019
Typeset using LaTeX **twocolumn** style in AASTeX61

THE PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLAsTiCC): SELECTION OF A PERFORMANCE METRIC FOR CLASSIFICATION PROBABILITIES BALANCING DIVERSE SCIENCE GOALS

A.I. Malz,[1,2] R. Hložek,[3,4] T. Allam Jr,[5] A. Bahmanyar,[4] R. Biswas,[6] M. Dai,[7] L. Galbany,[8] E.E.O. Ishida,[9] S.W. Jha,[7] D. Jones,[10] R. Kessler,[11] M. Lochner,[12,13] A.A. Mahabal,[14,15] K.S. Mandel,[16,17] J.R. Martínez-Galarza,[18] J.D. McEwen,[5] D. Muthukrishna,[16] G. Narayan,[19] H. Peiris,[6,20] C.M. Peters,[4] K. Ponder,[21] and C.N. Setzer[6]

(the LSST Dark Energy Science Collaboration and the LSST Transients and Variable Stars Science Collaboration)

[1] *Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, NY 10004, USA*
[2] *German Centre of Cosmological Lensing, Ruhr-Universitaet Bochum, Universitaetsstraße 150, 44801 Bochum, Germany*
[3] *Department of Astronomy and Astrophysics, University of Toronto, 50 St. George St., Toronto, ON M5S 3H4, Canada*
[4] *Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George St., Toronto, ON M5S 3H4, Canada*
[5] *Mullard Space Science Laboratory, Department of Space and Climate Physics, University College London, Holmbury Hill Rd, Dorking RH5 6NT, UK*
[6] *The Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, Stockholm, SE-106 91, Sweden*
[7] *Rutgers, the State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854 USA*
[8] *University of Pittsburgh, 300 Allen Hall, 3941 O'Hara St, Pittsburgh, PA 15260*
[9] *Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France*
[10] *University of Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA*
[11] *KICP, 5640 S Ellis Ave, Chicago, IL 60637*
[12] *African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg, 7945, South Africa*
[13] *South African Radio Astronomy Observatory, The Park, Park Road, Pinelands, Cape Town 7405, South Africa*
[14] *Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[15] *Center for Data Driven Discovery, California Institute of Technology, Pasadena, CA 91125, USA*
[16] *Institute of Astronomy and Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK*
[17] *Statistical Laboratory, DPMMS, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK*
[18] *Harvard-Smithsonian Center for Astrophysics, 60 Garden Street,Mail Stop 66 Cambridge, MA 02138*
[19] *Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218, USA*
[20] *Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[21] *Berkeley Center for Cosmological Physics, Campbell Hall 341, University of California Berkeley, Berkeley, CA 94720, USA*

ABSTRACT

Classification of transient and variable light curves is an essential step in using astronomical observations to develop an understanding of the underlying physical processes from which they arise. However, upcoming deep photometric surveys, including the Large Synoptic Survey Telescope (LSST), will produce a deluge of low signal-to-noise data for which traditional type estimation procedures are inappropriate. Probabilistic classification is more appropriate for the data but is incompatible with the traditional metrics used on deterministic classifications. Furthermore, large survey collaborations like LSST intend to use the resulting classification probabilities for diverse science objectives, indicating a need for a metric that balances a variety of goals. We describe the process used to develop an optimal performance metric for an open classification challenge that seeks to identify probabilistic classifiers that can serve many scientific interests. The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC) aims to identify promising techniques for obtaining classification probabilities of transient and variable objects by engaging a broader community beyond astronomy. Using mock classification probability submissions emulating realistically complex archetypes of those anticipated of PLAsTiCC, we compare the sensitivity of two metrics of classification probabilities under various weighting schemes, finding that both yield results that are qualitatively consistent with intuitive notions of classification performance. We thus choose as a metric for PLAsTiCC a weighted modification of the cross-entropy because it can be meaningfully interpreted in terms of information content. Finally, we propose extensions of our methodology to ever more complex challenge goals and suggest some guiding principles for approaching the choice of a metric of probabilistic data products.

1. INTRODUCTION

The Large Synoptic Survey Telescope (LSST) will revolutionize time-domain astronomy and the study of transient and variable objects within and beyond the Milky Way. With its rapid scan strategy, exquisite depth, and multiple optical filters, LSST will deliver millions of light curves, comprised of time-series observations in six electromagnetic wavelength ranges divided into photometric bands in the visible regime. LSST's expansive catalog of light curves will enable unprecedented population-level studies of time-varying astrophysical sources, from asteroids to variable stars to active galactic nuclei, deepening our understanding of stellar aging processes, the evolution of the most massive galaxies, and the expansion history of the universe, to name but a few.

Science output from the LSST dataset is, however, contingent on distinguishing classes of astrophysical sources from one another. Though photometric light curves like those of LSST can be used for classification, costly observations of a high-resolution spectrum have traditionally served as the gold standard for classification. The volume of objects anticipated of LSST, as well as the potentially low signal-to-noise ratios of the faintest sources, likely exceeds the availability of spectroscopic follow-up resources; the great majority of LSST's time-varying discoveries will never be spectroscopically confirmed. As such, there is an acute need for classifiers of photometric light curves that can perform well on datasets that include a wide variety of sources including those that are at the limits of detection.

The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC[1]) aims[2] to identify and motivate the development of classification techniques that serve astronomical science goals by engaging the broader community outside astronomy. PLAsTiCC's dataset is comprehensive, including models for well-understood classes, newly observed classes, and classes that have only been proposed to exist, to simulate serendipitous discoveries anticipated of LSST. Additionally, PLAsTiCC joins the ranks of a handful of past astronomy classification challenges including (Kitching et al. 2011, Mapping Dark Matter[3]), (Harvey et al. 2013, Observing Dark Worlds[4]), and (Dieleman et al. 2015, the Galaxy Challenge[5]), all hosted on Kaggle[6], a platform that hosts data analytics competitions where seasoned professionals and amateurs alike can compete to classify, model, and predict large data sets uploaded by companies or scientific collaborations. Kaggle attracts a broad userbase, and those without domain knowledge may provide novel approaches to the problem at hand.

Classification in astronomy may proceed through images, as has been done in the contexts of galaxy classification (Hoyle 2016), supernova classification (Cabrera-Vives et al. 2017), identification of bars in galaxies (Abraham et al. 2018), weak lensing estimation[7](Mandelbaum et al. 2014), separation of Near Earth Asteroids from artifacts in images (Morii et al. 2016), as well as time-domain classification (Morii et al. 2016; Mahabal et al. 2017; Zevin et al. 2017), and even noise classification (Zevin et al. 2017; George et al. 2018). Classification may also proceed from time-series or spectroscopic data rather than images, as in Newling et al. (2011); Richards et al. (2012); Ishida et al. (2013); Richards et al. (2015); Armstrong et al. (2016); Lochner et al. (2016); Möller et al. (2016). Automated classification (Mahabal et al. 2008; Djorgovski et al. 2011; Bloom et al. 2012; Djorgovski et al. 2012; Narayan et al. 2018) is becoming increasingly important in time-domain astronomy due to its potential for speed relative to visual inspection by an expert; the sooner one can make follow-up observations of an interesting object, the more one can learn about its underlying physical processes and nature.

Classification is intrinsically *probabilistic* in that the goal is to constrain the class *conditioned* on limited data, thereby defining a *posterior probability density*, or *classification posterior* for short, over all classes for each classified light curve. Probabilities of classification that are reduced to an estimated class label (say, by rounding a probability $0 \leq p \leq 1$ up or down) without a notion of confidence become *deterministic* classifications. Such a reduction of a probability density to a deterministic

[1] http://plasticcblog.wordpress.com/, https://www.kaggle.com/c/PLAsTiCC-2018

[2] PLAsTiCC was run as a Kaggle challenge from 17 September 2018 to 17 December 2018. Though PLAsTiCC concluded prior to the final revision of this paper, the study herein was conducted entirely before the commencement of PLAsTiCC, and the draft was submitted to the journal prior to PLAsTiCC's conclusion, hence the use of the present and future tenses throughout this paper.

[3] https://www.kaggle.com/c/mdm

[4] https://www.kaggle.com/c/DarkWorlds

[5] https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

[6] https://www.kaggle.com/

[7] http://great3challenge.info/

label discards information, the impact of which depends on how the classification results are subsequently used.

Probabilistic classifications could inform decisionmaking regarding allocation of limited spectroscopic follow-up resources. To reduce wasting spectroscopic resources dedicated to a common class whose science use requires spectra, one might only attempt follow-up observations of the objects with the highest classification probabilities. Spectroscopic follow-up of a rare class, on the other hand, may be useful enough that an object with even a moderate probability of being of a very rare class could be worth the risk.

Perhaps more significantly, classification probabilities may be propagated through a hierarchical inference of population-level parameters, enabling scientific investigations to proceed even when spectra are unavailable. The efficacy of this application of classification probabilities in the context of supernova cosmology is an active field of research (Rubin et al. 2015; Roberts et al. 2017; Jones et al. 2018, Malz, Peters, and Hložek in prep). Thus the impact of a photometry-only survey like LSST can be greatly enhanced by probabilistic classifications.

In light of the aforementioned benefits of classification probabilities, PLAsTiCC will thus accept classifiers producing classification posteriors.[8] However, probabilistic classifications are incompatible with the *metrics*, any quantification of the performance of a classifier, of deterministic label assignments used in previous classification challenges (Kessler et al. 2010a,b) and efforts to develop supernova classifiers (Narayan et al. 2018). Accuracy, purity, completeness, and contamination are examples of metrics of deterministic classification estimates that are commonly used in astronomical applications.

Many deterministic classification metrics can be modified for evaluation on classification posteriors (Gieseke et al. 2010; Lochner et al. 2016; Möller et al. 2016; Hon et al. 2017, 2018b), but only by reducing class probabilities to deterministic labels via evaluation at different cutoffs, the choice of which may ultimately affect the value of the metric and thus assessment of the classifier. Furthermore, many such metrics are restricted to binary classifications ("yes" or "no") and thus do not meet the diverse needs of PLAsTiCC.

If the data are simulated using a fully self-consistent forward model, a metric of the accuracy of classification posteriors relative to the true, underlying probabilities would be straightforward. However, such a simulation procedure would require beginning with a fully populated probability space over all classes and all possible light curves, which is an insurmountable challenge. Therefore, attention must be directed toward defining the criterion for identifying a winning classifier. In the context of astronomy, concerns about the choice of metric for probabilistic classifications have been investigated (Kim & Brunner 2017; Florios et al. 2018), though most studies focus on the standard metrics of purity and completeness. Even within that subset, metric consistency over a range of classifiers and between different analyses is not always ensured (Bethapudi & Desai 2018), indicating a need for further study.

This work explores the problem of how to choose a metric of probabilistic classifications with intended application to many science applications. The PLAsTiCC metric must respect the information content of probabilistic classifications without reduction to point estimates of class; it must be well-defined for non-binary classes, going beyond a positive/negative dichotomy inherent to some traditional metrics. The winning classifier should not favor one science application above all others, necessitating robustness against significant class imbalance, both between and within the training set and test set, as well as other concerning systematics. Finally, in order for the metric to satisfy the challenge requirements, the metric must return a single, scalar value.

We perform a systematic exploration of the sensitivity of metrics of probabilistic classification to anticipated classifier failure modes using the PRObabilistic CLAssification Metric (`proclam`) code (Malz 2018), which is publicly available on GitHub[9]. The mock classification submissions that we use for this study are described in Section 2. The metrics we consider are presented in Section 3. The behavior of the metrics as a function of mock classification results is presented in Section 4. We discuss extensions of this exploratory framework to more complex challenge goals in Section 5.

2. DATA

We explore the behavior of metrics on mock classification probabilities with isolated strengths and weaknesses as well as realistic mock classification probabilities from a publicly available light curve catalog. Throughout this paper, *data* always refers to mock classification submissions to PLAsTiCC, not the PLAsTiCC light curves; no light curves were simulated, viewed, or classified in the preparation of this paper.

[8] Classifiers that only provide deterministic or binary classifications (including some of the most prevalent classifiers in the field of time-domain astronomy) will have to convert their results to probability vectors to compete in PLAsTiCC.

[9] https://github.com/aimalz/proclam

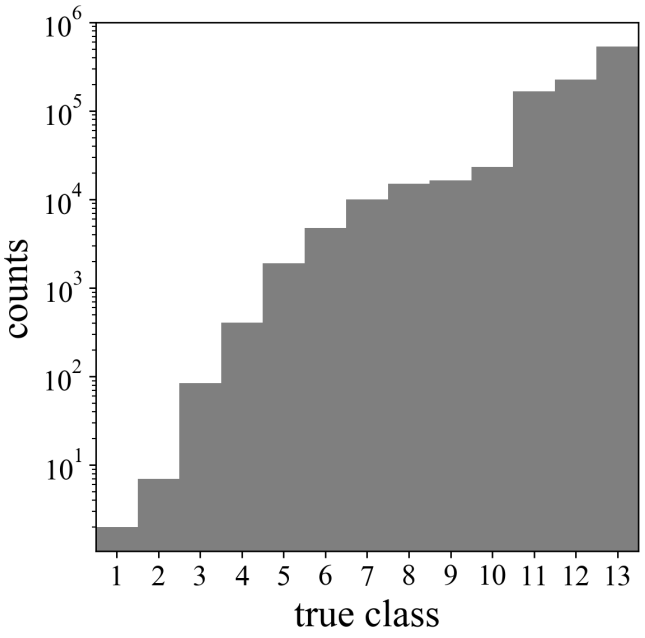

Figure 1. ~~The number of objects in a given class as a function of class population size. The true class populations are logarithmically distributed.~~ The number of members of each of thirteen mock classes considered in this work. Class populations were simulated by drawing the number of members of a given class from a logarithmic distribution to emulate the extreme class imbalances typical of astronomical samples.

Our data is in the form of catalogs of N posterior probability vectors $p(m \mid d_n, D, \mathcal{C})$ over M classes with labels m conditioned on each observed light curve d_n, the training set D, and some parameters $\mathcal{C}$ concerning the behavior of the classifier. We motivate $\mathcal{C}$ here before deferring its detailed explanation to later in Section 2.1.

If a mock classifier produced $p(m \mid d_n)$, it would take solely the light curve and produce a posterior over classes. Since such a situation involves no information besides the light curve d_n, every classifier would produce identical classification submissions $\bar{p}(m \mid d_n)$. Including the training set D would not remedy the problem, as every classifier for PLAsTiCC has access to the same training set and so would still have no way to produce different classification submissions $p(m \mid d_n, D)$. Thus there must be some other parameters $\mathcal{C}$ that are specific to each classifier and contribute to the mock classification posteriors it produces.[10] We describe below the way in which mock data is synthesized and return to the classifier parameters $\mathcal{C}$ later.

As is anticipated of the real LSST dataset, we use class populations that are logarithmically distributed such that they span many orders of magnitude. We then take M draws $u_m \sim \mathcal{U}(0,1)$ from the standard continuous uniform distribution. These draws $\{u_m\}$ are used to establish a discrete probability distribution $p(m) = b^{u_m} / \sum_m b^{u_m}$ such that $\sum_{m=1}^{M} p(m) = 1$. From $p(m)$ we draw $N = 10^b$ instances $\{m'_n\}$ of a true class m' for each light curve n in the catalog.

The true class membership distribution of our tests with $M = 13$ and $b = 6$ is shown in Figure 1. Though the class labels for PLAsTiCC are expected to be randomized, we artificially order our mock class labels by their prevalence for ease of visual interpretation. Once the true classes have been set, mock classification probabilities for each class are derived using the procedure described in Section 2.1.

2.1. *Mock classification schemes*

In order to observe metric performance on different classification schemes, we simulate some archetypical mock classifiers, devised to produce generic responses to a classification challenge, without any interaction with actual challenge data, nor any other light curves. We use these mock classifiers to investigate how the performance under each metric changes in the presence of certain types of failure modes, or *systematics*. A robust metric should not reward classification schemes that display these systematic effects.

The archetypical systematics can be seen as modifications to the confusion matrix, a measure of deterministic classification (Bloom et al. 2012). The confusion matrix is an $M \times M$ table of observed counts (or, if normalized, rates) of pairs of estimated class labels $\hat{m}$ (columns) and true classes m' (rows) computed after a deterministic classification has been performed on some data set with N objects.

Under a binary deterministic classification between positive and negative possibilities, the confusion matrix contains the numbers of true positives TP, false positives FP (Type 1 error), true negatives TN, and false negatives FN (Type 2 error), which can be turned into rates relative to the true numbers of positive and negative instances. These rates may serve as building blocks for more sophisticated metrics of multi-class deterministic classifiers addressed in Section 3. Though probabilistic classifications are not compatible with the confusion ma-

[10] It should be noted that classification submissions may not be derived in this way, i.e. the parameters $\mathcal{C}$ may not be explicitly known or may indicate a procedure that does not produce posteriors but, rather, scores of some kind. However, we assume for these purposes that classifiers produce the classification posteriors PLAsTiCC seeks.

trix, regardless of normalization, we design tests around proposed normalized confusion matrices exhibiting various systematics that we anticipate being problematic for LSST.

Under a deterministic classification scheme with a normalized confusion matrix with elements $p(\hat{m}, m')$, an object with true class m' would have an assigned class $\hat{m}$ drawn from $p(\hat{m} \mid m') = p(\hat{m}, m')/p(m')$, via Bayes' Rule. We note that the elements of the confusion matrix have values of $Np(\hat{m}, m')$ and that $p(m') = N_{m'}/N$, where $N_{m'}$ is the number of true members of class m', must be known in order to produce a confusion matrix. We refer to the matrix $\mathbb{C}$ composed of $p(\hat{m} \mid m')$ as the *conditional probability matrix* (CPM), and we use it to derive mock classification posteriors.

Assuming the light curves contain information about the true class (an assumption that underlies classification as a whole), we can use the appropriate row $\mathbb{C}_{m'_n} = p(\hat{m} \mid m', \mathcal{C})$ of the CPM $\mathbb{C}$ as a proxy for $p(m \mid d_n, D, \mathcal{C})$, without directly classifying light curves themselves.[11] To emulate the effect of natural variation of information content in different light curves (e.g. a noisy lightcurve has less information to recover than one with a higher signal-to-noise ratio) using the above, we generate a posterior probability vector $\vec{p}(m \mid m', \mathbb{C})$ by taking a Dirichlet-distributed draw

$$\vec{p}(m \mid d_n, D, \mathcal{C}) \sim \mathrm{Dir}[\mathbb{C}_{m'_n} \delta] \tag{1}$$

about $\mathbb{C}_{m'_n}$, with a small nonnegative perturbation factor $\delta = 0.01$. In this way, the posterior probability vector has an expected value equal to the appropriate row in the CPM, with a variance set by δ. We impose one restriction in addition to the normalization factor of Equation 1, namely that all elements of $p(m \mid d_n, D, \mathcal{C})$ exceed 10^{-8}, to ensure numerical stability in light of the limitations of floating point precision.

We consider eight mock classifiers, each characterized by a single systematic affecting their CPM. Figure 2 shows the CPMs corresponding to each systematic considered, discussed in detail below.

For each of our archetypical mock classifiers, we address:

1. What characteristic behavior defines this classifier?

2. Under what conditions does this behavior arise in real classifications?

3. What are our expectations of and desires for response of the metric to this archetypical classifier?

An actual classifier is expected to be more complex than the simplified cases of Figure 2, with different systematic behavior for each class. An example of a combined CPM across different classes and systematics is given in the top panel of Figure 3. The rows of this CPM correspond to rows of the archetypical classifiers of Figure 2. To demonstrate the procedure by which mock classification posteriors are generated from rows of the CPM, we provide 22 examples of draws of the posterior CPM in the bottom panel of Figure 3. Given a set of true class identities, the mock classification posteriors of the bottom panel are Dirichlet draws from the corresponding row of the CPM of the top panel.

2.1.1. *Uncertain classification*

A CPM $\mathbb{U}$ with uniform probabilities for all classes, as shown in the leftmost top panel of Figure 2, would correspond to uniform random guesses for deterministic classification, but in accordance with Equation 1, the classification posteriors are perturbations away from a uniform distribution across all classes. The peak values of one such classification posterior would correspond to random classification drawn from a uniform distribution, with $p(m' \mid d_n, D, \mathcal{C}_{\mathbb{U}}) \approx M^{-1}$. We can consider the *uncertain* classifier as an experimental control for the least effective possible classification scheme, bearing in mind that if classifications were anticorrelated with true classes, the experimenter could simply reassign the classification labels to improve performance under any metric.

2.1.2. *Accurate classification*

The *perfect* classifier has a diagonal CPM $\mathbb{I}$ (left-center top panel of Figure 2), which would correspond to deterministic classifications that are always correct. In terms of probabilistic classifications, a perfect result would be a classification posterior with 1 for the true class and 0 for all other classes. In accordance with the classification posterior synthesis scheme of Equation 1, the class with maximum probability is almost always still the true class, and indeed with $N \sim 10^6$ and $\delta = 0.01$, this is always true. This case is also a control, in that PLAsTiCC would not be necessary if we believed the perfect classifier were potentially achievable.

In addition to a perfect classifier, we test linear combinations $\mathbb{C} = (s+1)^{-1}(s\mathbb{I} + \mathbb{U})$ of the perfect and uncertain CPMs where the contribution of the perfect classifier is greater than that of the uncertain classifier by a factor of $s > 0$. Deterministic classifications drawn from such a CPM would be correct s times as often as they

[11] This assumption is key to the generality of this work, which was conducted without any knowledge of the PLAsTiCC dataset simulation procedure.

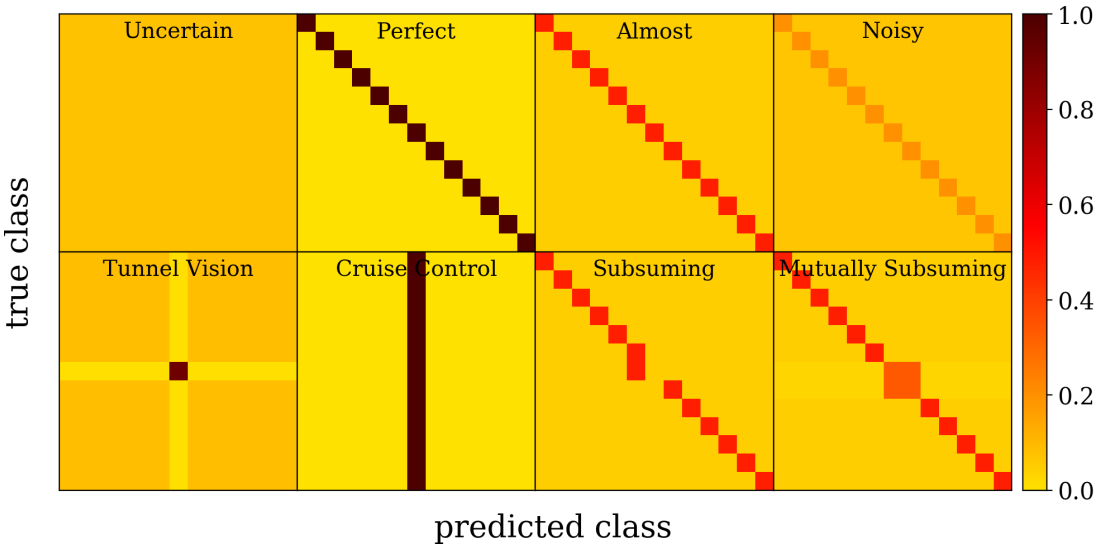

Figure 2. Conditional probability matrices (CPMs) for eight mock classifiers. Top row: the uncertain classifier's uniform CPM; the perfect classifier's identity CPM; the almost perfect classifier's CPM, a linear combination of one part uniform and four parts identity; the noisy classifier's CPM, a linear combination of one part uniform and two parts identity. Bottom row: the tunnel vision classifier's CPM is uniform except at the row and column corresponding to one class, where it takes the values of the identity matrix; the cruise control classifier's CPM, which has the every row equal to a particular row of the identity; the subsuming classifier's CPM, which has two or more rows equal to one another; the mutually subsuming classifier's CPM, a symmetric case of the subsuming classifier. The top row shows CPMs that serve as unbiased control cases. The CPMs of the bottom row represent concerning systematics that we would like to ensure are not rewarded by the PLAsTiCC metric.

take any one wrong label, and the incorrect labels would be uncorrelated across classes. The classification posteriors drawn from such CPMs would have some probability at classes other than the true class, but almost all would still have their peak value at their true class. We consider the case of the *almost perfect* classifier with $s = 4$ (right-center top panel of Figure 2) and the *noisy* classifier with $s = 2$ (rightmost top panel of Figure 2).

A classifier with different accuracy for each class may be considered a systematic in its own right. An extreme example of such a classifier is one with perfect classification performance on one class and uncertain classification on all others. This classifier's CPM would be uniform except for one row, which would take a value of unity on the diagonal and zero elsewhere; if the classifier were also resilient against Type 1 errors, the CPM would also take zeros along the column in question, aside from the value of unity on the diagonal. For a single science application, this type of classifier is desirable, but the goal of PLAsTiCC is to serve the needs of those who study a wide variety of classes for different purposes. Hence, from the perspective of PLAsTiCC, we seek a metric that disfavors the *tunnel vision* classifier (leftmost bottom panel of Figure 2).

2.1.3. *Inaccurate classification*

If a deterministic classifier is systematically inaccurate, its CPM has significant off-diagonal contributions. We model inaccurate probabilistic classifications of class m' by using the row of the CPM corresponding to class $\tilde{m}$ as the basis for the perturbed probability vector $p(m \mid m') = p(m \mid \tilde{m})$. Class m' is said to be *subsumed* by class $\tilde{m}$ by a classifier that absorbs class m' into class $\tilde{m}$ (right-central bottom panel of Figure 2). The subsuming classifier may be asymmetric, or the classes may be mutually subsumed (rightmost bottom panel of Figure 2) if one already has significant off-diagonal probability, as is true for the uncertain classifier.

Subsuming is not always the mark of a poor classifier and may be insurmountable by more sophisticated classification techniques. Real classification posteriors $p(m \mid d_n, D, \mathcal{C})$ are conditioned on light curves, training data, and assumptions necessary for the classification algorithm, and there may simply not be enough information in a light curve and/or training set to distinguish between classes.

For example, based on only the first few light curve points, it is sometimes impossible to separate cataclysmic variables (stars that are not destroyed and can brighten and fade many times) from supernovae, which are stars that are completely destroyed in their explosions. Even with observations over extended periods, it can still be impossible to distinguish cataclysmic variables from active galactic nuclei that result from activity near a galaxy's central black hole. Similarly, tidal dis-

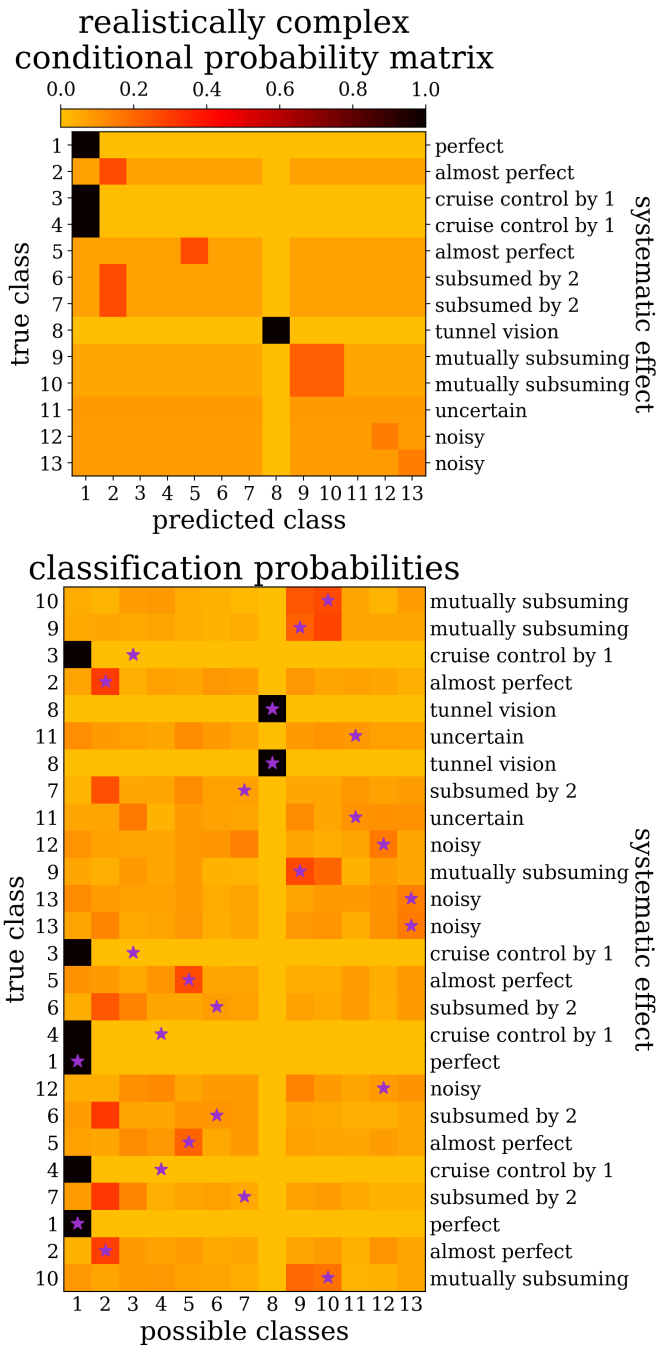

Figure 3. A realistically complex conditional probability matrix (CPM) and classification posteriors drawn from it. Top: An example of a realistically complex conditional probability matrix, constructed by selecting a systematic for each individual class. This illustrates (for example), how a classifier may exhibit multiple systematics from Figure 2 for each true class. Bottom: Example classification probabilities, drawn from the above CPM, with their true class indicated by a red star and the systematic, characterized by its row in the CPM, affecting that true class described on the right. The Dirichlet process emulates the variation in classification posteriors due to differences between light curves within a given class, leading to different classification posteriors even among rows sharing a true class.

ruption events that occur when stars are destroyed by proximity to the central black hole of a galaxy can look much like supernovae that simply happen to be near a galaxy's center. When the prior information of the location of the source is more informative than its sparse, noisy, irregularly sampled, or short light curve, it may present a challenge no classifier can overcome, a fundamental limit on available information about the object.

Distinguishing between subclasses of a single phenomenon is subject to limits not only on the light curves of the unknown targets but also by the availability of adequate training sets. It is nonetheless essential to identify subclasses when they have wholly different science applications. As an example, supernovae (SN) Ia and Ibc are notorious for being difficult to distinguish. In fact, it is more common for SN Ibc to be misclassified as SN Ia than the other way around. This asymmetry is due to systematic underrepresentation of SN Ibc in available training sets. However, SN Ibc contaminants in the traditional cosmology analysis done with SN Ia can bias estimates of the cosmological parameters, so the distinction is critical.

Class imbalance is a ubiquitous problem in astronomy that can severely exacerbate this form of inaccuracy, as the relative rates of various astrophysical events and objects differ by orders of magnitude from one another. For example, RRc and RRd Lyrae stars are challenging to separate despite having different pulsation modes, and RRd stars, due to their rarity, are typically subsumed by RRc labels.

An extreme case of inaccurate classification is to classify all objects as the most common class (in the training or test set), which is of particular concern to PLAsTiCC given non-representative class balance of the training set. Such a *cruise control* classifier (left-center bottom panel of Figure 2) counters PLAsTiCC's goal of identifying objects belonging to extremely rare classes. We would like the PLAsTiCC metric to reward a classifier that successfully avoids this kind of error.

2.2. *Realistic classifications*

In order to understand the performance of classifiers on simulated datasets approximating reality, we calculate the values of our metric candidates on representative classifiers of a precursor light curve classification challenge. The Supernova Photometric Classification Challenge (SNPhotCC) (Kessler et al. 2010a) focused on deterministically classifying a heterogenous population of supernovae into subclasses of SN Ia, SN II, and SN Ibc.

The SNPhotCC attracted diverse classification approaches, encompassing χ^2 fits of the supernova light

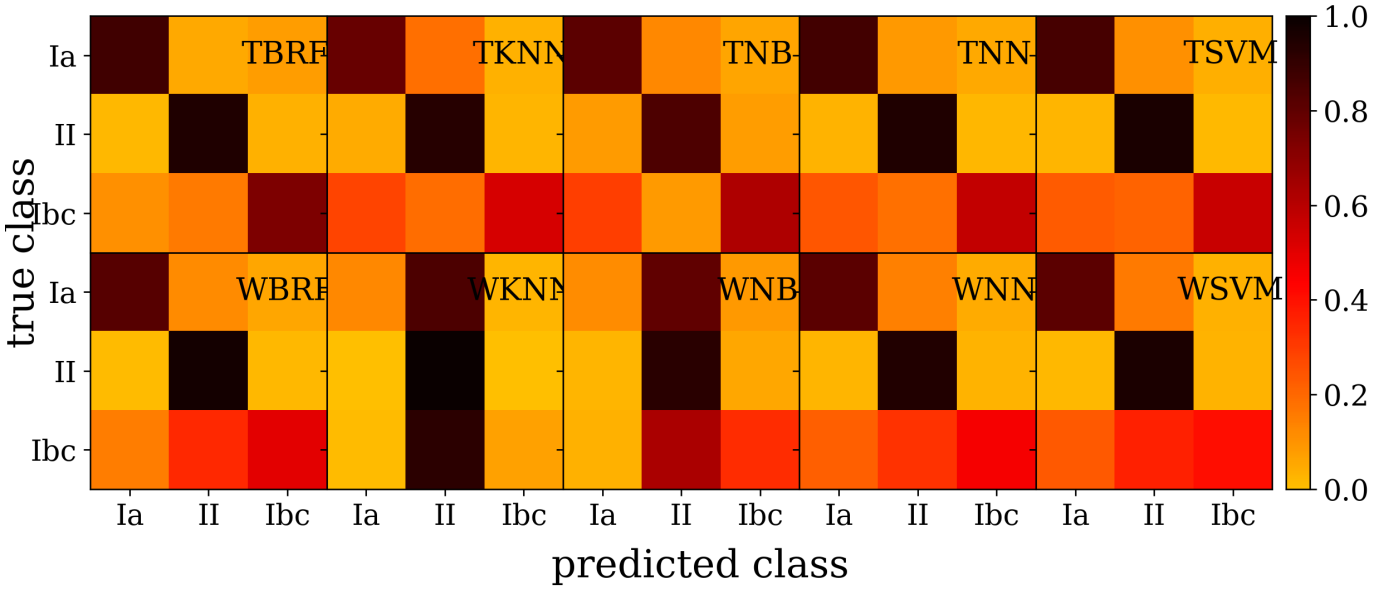

Figure 4. Conditional probability matrices (CPMs) of the Lochner et al. (2016) methods applied to the second post-challenge release of the SNPHOTCC dataset. Columns: the five machine learning methods of Boosted Decision Tree (BDT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Neural Network (NN), and Support Vector Machine (SVM). Top row: five machine learning methods applied to template decompositions as features. Bottom row: the same five machine learning methods applied to wavelet decompositions as features. These CPMs derived from the dataset of a precursor light curve classification challenge by modern methods exhibit some of the systematics identified in Section 2.1 and Figure 2, particularly cruise control (WKNN, WNB), noisy (class Ibc in all but TBDT and WKNN), and perfect (class II in all). It is worth noting that Lochner et al. (2016) applies their classification to a representative sub-sample of the SNPHOTCC data selected once the challenge was complete, circumventing some of the issues of non-representativity present in the original submissions to the SNPHOTCC.

curves to publicly available templates (Nugent et al. 2002), empirical models (Conley et al. 2008), as well as alternatives to curve-fitting such as outlier identification on the training set Hubble diagram, dimensionality reduction, and clustering. Machine learning was also employed, using features such as the light-curve slopes to produce a predictive model for the training data.

Since the conclusion of the SNPHOTCC, the light curves became a testbed for a suite of machine learning classifiers. We consider a collection of probabilistic classification methods, as presented in Lochner et al. (2016), whose CPMs[12] are shown in Figure 4.

The set of classification algorithms includes template-based classification procedures, denoted as T, (Sako et al. (2011), top row) and a wavelet decomposition, denoted as W, of the light curves to construct the features over which to classify (Newling et al. (2011), bottom row), each paired with Boosted Decision Tree (BDT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Neural Network (NN), and Support Vector Machine (SVM) machine learning algorithms (columns). While the complexity of entries to the SNPHOTCC was greater than this subset, we use these examples to establish the behavior of our metrics on realistic classification submissions.

We draw attention to the marked presence of the systematics introduced in Section 2.1 in the CPMs of Figure 4. Note that the WNN and WNB methods both suffer from the cruise control systematic on SN II, which were the most prevalent in the SNPHOTCC dataset. Nearly all the other CPMs exhibit classifications that are almost perfect for SN Ia, perfect for SN II, and noisy for SN Ibc. A likely cause for this effect is that SN Ibc are poorly represented in training and template sets.

3. METHODS

To optimally discriminate between classification techniques, there must be a performance metric, a single scalar value quantifying how appropriate a classifier is

[12] The classifiers of Lochner et al. (2016) are indeed probabilistic but are reduced to confusion matrices via deterministic labels (by assigning a label of the class achieving the highest probability) for this visualization and the science-motivated metric of Section 3.1. In all other instances, the classification posteriors are used directly.

for the task at hand. Choosing a metric for PLAsTiCC therefore is logically entwined with the challenge goals.

In Section 3.1, we review a familiar binary, deterministic metric of light curve classification in astronomy. In Section 3.2, we introduce metrics appropriate for multi-class probabilistic classification. We take weighted averages of the per-object metrics with per-class weights described in Section 3.3.

3.1. *Science-motivated deterministic metric*

We begin with a presentation of a classification metric that has been used in the evaluation of astronomical light curve classifiers in the recent past. The metric we highlight makes use of the notions of true positive, false positive, and false negative counts from binary deterministic classification. We briefly define the *efficiency* $\epsilon \equiv \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and *purity* $\pi \equiv \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$.

The goal of the SNPhotCC was to identify one particular type of astrophysical source, SN Ia, for a single scientific application, cosmology. As the SNPhotCC was only concerned with SN Ia cosmology, it was effectively binary, in that the metric did not distinguish between non-Ia classes. Since the only SN Ia that would be considered for a cosmology analysis at the time were those with spectroscopic redshifts, the classification was not only binary but also deterministic. The SNPhotCC metric FoM $\equiv \epsilon \cdot \tilde{\pi}$ is the product of the efficiency of SN Ia classification and a modification $\tilde{\pi} \equiv \mathrm{TP}/(\mathrm{TP}+r\mathrm{FP})$ of the purity in terms of a penalty factor r. The inclusion of this second term was motivated by the potential impact on cosmological parameter constraints due to contamination of the SN Ia sample by non-Ia classes. The pseudo-purity can be interpreted as the traditional purity when $r = 1$ as it is related to the size of the spectroscopic sample; for the SNPhotCC, $r = 3$ was used.

3.2. *Probabilistic metrics*

In contrast to SNPhotCC's sole goal of optimal deterministic classification of a single class, PLAsTiCC seeks to identify classifiers that produce multi-class classification posteriors. We consider two metrics of classification probabilities that avoid reducing probabilities to deterministic labels.

Our probabilistic metrics are composed of quantities defined for each possible class m among M potential classes available to light curve n, which is a true member of the set $\mathbb{S}_{m'}$ of astrophysical sources of class m'. The metric value $Q_n = \sum_{m=1}^{M} Q_{n,m}$ for a single light curve n is a sum of the per-class per-light curve metric values $Q_{n,m}$. The metric value $Q_{m'} = \sum_{n\in\mathbb{S}_{m'}} Q_n$ for an entire class m' is the sum of the per-light curve metrics. Section 3.3 discusses how the global metrics are derived from the per-class metrics $Q_{m'}$.

As part of the derivation of the per-class per-light curve metrics, we also define the indicator variable

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

that indicates if an object has been correctly classified as its true type.

3.2.1. *Log-loss*

The log-loss is a quantity borrowed from information theory and is related to a notion of *entropy* $H_n = -\sum_{m=1}^{M} p(m \mid d_n) \ln[p(m \mid d_n)]$, a measure of the space of possible states a system can have, which is in this case the class of which a light curve can be a member. A classification posterior $p(m \mid d_n)$ has minimal entropy if it takes a value of 1 at some class and values of 0 at all others, i.e. if it can trivially be reduced to a deterministic classification, because this is the scenario in which there is only one possible state, that the light curve has a true class m. This definition of entropy, however, is a property of the probability $p(m \mid d_n)$ and has no relation with any concept of the true class of the light curve m'.

To reconcile the classification posterior with the true class known by those running a challenge, we define the cross-entropy

$$L_n \equiv Q_n^L = -\sum_{m=1}^{M} \tau_{n,m} \ln[p(m \mid d_n)], \tag{3}$$

which can be interpreted as the spuriously oversized space of possible states (an increase in disorder) due to using the classification posterior in place of the indicator variable. Whereas H_n is minimized to a value of 0 by any deterministic classification, L_n is minimized to a value of 0 only if τ_n and $p(m \mid d_n)$ are equal to one another. It can also be proven that the uncertain classifier of Section 2.1.1 maximizes L_n (Murphy 2012). As an aside, a difference between L_n and H_n evaluated at $\tau_{n,m}$ would be the information lost to disorder in using $p(m \mid d_n)$ in place of $\tau_{n,m}$, also known as the Kullback-Leibler Divergence (KLD); see Malz et al. (2018) for a comprehensive exploration of the KLD for a continuous 1-dimensional probability space.

The log-loss has only recently established a presence in the astronomy literature (Hon et al. 2017, 2018a). Its greatest strength is that it is straightforwardly interpretable, enabling the metric itself to contribute to uncertainty propagation in an inference problem using the probability densities provided by the classifier.

3.2.2. *Brier score*

The Brier score (Brier 1950), given as

$$B_n \equiv Q_n^B = \sum_{m=1}^{M} (\tau_{n,m} - p(m \mid d_n))^2, \tag{4}$$

is a mean square error calculated between the indicator variable and the classification posterior. Unlike the log-loss, the Brier score has been used extensively in solar flare forecasting (Crown 2012; Mays et al. 2015; Florios et al. 2018), stellar variability identification (Richards et al. 2012; Armstrong et al. 2016), and star-galaxy separation (Kim et al. 2015).

As with the log-loss, the Brier score is minimized to 0 only for a perfect classifier. The Brier score is an attractive option because it both rewards classifiers for assigning more probability to the true class and penalizes classifiers for assigning any probability to classes other than the true class, in contrast to the log-loss, which only accounts for probability assigned to the true class. We expect this difference to significantly distinguish the Brier score from the log-loss.

The interpretation of the Brier score is less obvious than that of the log-loss, as its dimensions depend on those of the probability space upon which the classification posteriors are defined. In addition, modifying it with weights requires choosing whether to weight only per-object values B_n or also the individual terms $B_{n,m}$ contributing to it. We leave to future work the thorough investigation of a nontrivial weighting scheme on the Brier metric, however, opting to treat both metrics the same, according to the weighting scheme of Section 3.3, in our implementation.

3.3. *Weights*

The most concerning systematics discussed in Section 2.1 are those of tunnel vision and cruise control. The actual light curve data stream of LSST will be particularly vulnerable to both due to extreme class imbalance and class hierarchy (for example different subtypes of a single transient or variable class). This susceptibility is compounded by the nonrepresentativity of the PLAsTiCC training set, which is designed to reflect the nonrepresentativity anticipated of LSST. Any metric under equal weight per light curve would incentivize tunnel vision and cruise control focused on the most prevalent class. In order to meet the needs of science cases concerning other, rarer classes, PLAsTiCC's metric will be more nuanced, even if it complicates the interpretability of the metric.

One option is to apply a threshold of classification efficacy on all classes in order to assign an overall winner, though it would require reducing the classification probabilities to deterministic class labels. When doing binary classification with a method that reduces probabilities to deterministic class labels, each light curve is assigned the class of higher probability, even if the two probabilities are quite similar, a situation that is particularly likely if the light curve, in fact, belongs to a third class or if the two classes are subclasses of a single physical phenomenon. A simple reduction to a deterministic label could be made more palatable with a secondary threshold mechanism. For example, requiring a minimum difference in probability density between the maximum probability class and the next highest probability class would help avert this degeneracy.

A simpler alternative that we investigate in this paper is to use a weighted average

$$Q = \frac{1}{\sum_m w_m} \sum_m w_m Q_m \tag{5}$$

of per-class metrics Q_m. (While weights could be assigned to each term $Q_{n,m}$, we do not consider this complexity at this time.) Weights that are not proportional to N^{-1} nor M^{-1} may be chosen to encourage challenge participants to direct more attention to classes with less active classification efforts or those that have been historically more difficult to classify due to observational limitations.

Downweighting the metrics of classes affected by counterproductive systematics could mitigate the impact of the tunnel vision or cruise control classifiers. The weights for the PLAsTiCC metric, however, must be determined before there is knowledge of which systematics affect which classes. Because of this caveat, the choice of weights is isolated to an inherently human problem dictated by the value placed on the scientific merits of knowledge of each class. This paper, on the other hand, can only quantify the impact of weights in relation to the systematics. We thus agnostically test weighting schemes[13] where classes affected by a particular systematic take a given weight $0 \leq w \leq 1$ and all other classes have a weight $(1 - w)/(M - 1)$.

4. RESULTS

[13] The weights considered in this study are more extreme than those ultimately used for PLAsTiCC because the true weights were withheld from some authors prior to the end of the challenge. However, in the Kaggle framework, it is possible to estimate these values by systematically probing the output of the public leader board with entries from the cruise control classifier archetype targeting each class one at a time. Some PLAsTiCC competitors did, in fact, execute this procedure and publicly announced the weights they had discovered, making the information available to all participants.

In the following sections, we explore the response of the log-loss and Brier score metrics to the classifiers of Section 2 and as a function of the weights on affected classes.

4.1. *Mock classifier systematics*

We simulate probabilistic classifications as potential submissions to PLAsTiCC by the methodology of Section 2.1 based on CPMs composed of pairs of the characteristic classifiers shown in Figure 5 under various weightings described below.

The systematics introduced to each baseline are those that we intuitively expect to worsen classification performance of an arbitrary classifier:

- the uncertain, almost perfect, noisy, and subsuming classifiers are anticipated to worsen an otherwise perfect classifier;
- the uncertain, noisy, and subsuming classifiers are anticipated to worsen an otherwise almost perfect classifier;
- the uncertain and subsuming classifiers are anticipated to worsen an otherwise noisy classifier.

In every case, we apply the systematic to one true class, which corresponds to transforming one row of the baseline CPM.

The introduction of weights illustrates the effect each particular systematic has on a given baseline, and more importantly, how up- (or down-) weighting the affected class changes the overall metric value for the mock classifier. Weighting schemes are defined by a weight $0 \leq w \leq 1$ on the affected class, with the remaining baseline classes sharing equal weight $(1-w)/(M-1)$; we test eleven weighting schemes with $w = 0., 0.1, \ldots, 1.$. A higher weight on the systematic corresponds to a lower weight on the more desirable baseline, causing both the log-loss and Brier score to increase. This variation in weights establishes linear relationships between the log-loss and Brier score metrics for each pair of baseline and systematic, but the slope is related to the relative sensitivity of the metrics.

Figure 5 confirms that for all weight on the perfect classifier, the values of both metrics vanish to zero. It is worth noting that the log-loss has more dynamic range than the Brier score overall, and that the log-loss is acutely sensitive to the subsuming systematic on a baseline of a perfect classifier. However, the relative scales of metric values for different baseline-plus-systematic pairs are quite large, requiring three panels, zooming in from left to right.

The left panel of Figure 5 shows the largest variations in metric scores, for the combination of the perfect baseline and a subsuming systematic where one class is given a probability of 1 for being in another particular class and a probability of 0 for being in its true class. This means both metrics are acutely sensitive to the subsuming systematic on a perfect baseline, which can only be overcome by aggressive downweighting. In fact, the log-loss value for a classifier that subsumes a class into one that is classified perfectly should be infinite if the classes unaffected by the systematic have no weight; it is only finite for us because of the limits of numerical precision.

The middle panel of Figure 5 illustrates a narrower range of log-loss and Brier score for the subsuming systematic on the almost perfect and noisy classifier baselines. The subsuming systematic on any baseline besides the perfect classifier defines a new regime of high but not infinite values of the metrics.

The right panel of Figure 5 shows the values for all other systematics on all baselines. Though the slope is lower than in the other panels, the dynamic range of the log-loss remains higher; in other words, the log-loss is in general more sensitive to systematics than the Brier score.

In summary, both the log-loss and Brier score are most sensitive to the subsuming systematic than any other systematic. Tuning the weights can provide an avenue toward imposing a global metric penalty on classifiers exhibiting a systematic on one class.

When all weight is on the class exhibiting the systematic, there is a characteristic limit for each metric's values, shown in Table 1. Because a subsumed class takes the conditional probability vector of the subsuming class, the metric values depend on what systematics may be affecting the subsuming class as well. While the two metrics obviously take different values, in accordance with their slopes given in Table 2, they do agree on the ranking of these classifiers. Though this agreement is not in general guaranteed, it is a desirable behavior, indicating that these metrics would lead to the same conclusion about the severity of each systematic.

The relative sensitivity ratios of the log-loss to the Brier score are the slopes in the trends of Figure 5 and are given in Table 2. The log-loss always has higher sensitivity than the Brier score (i.e. it responds more strongly to up-weighting classes affected by a systematic), particularly to the difference between the perfect classifier and any lesser classifier. A possible implication of this behavior is that the log-loss may have an enhanced ability to distinguish between multiple high-performing classifiers that might not have meaningfully different metric values under the Brier score.

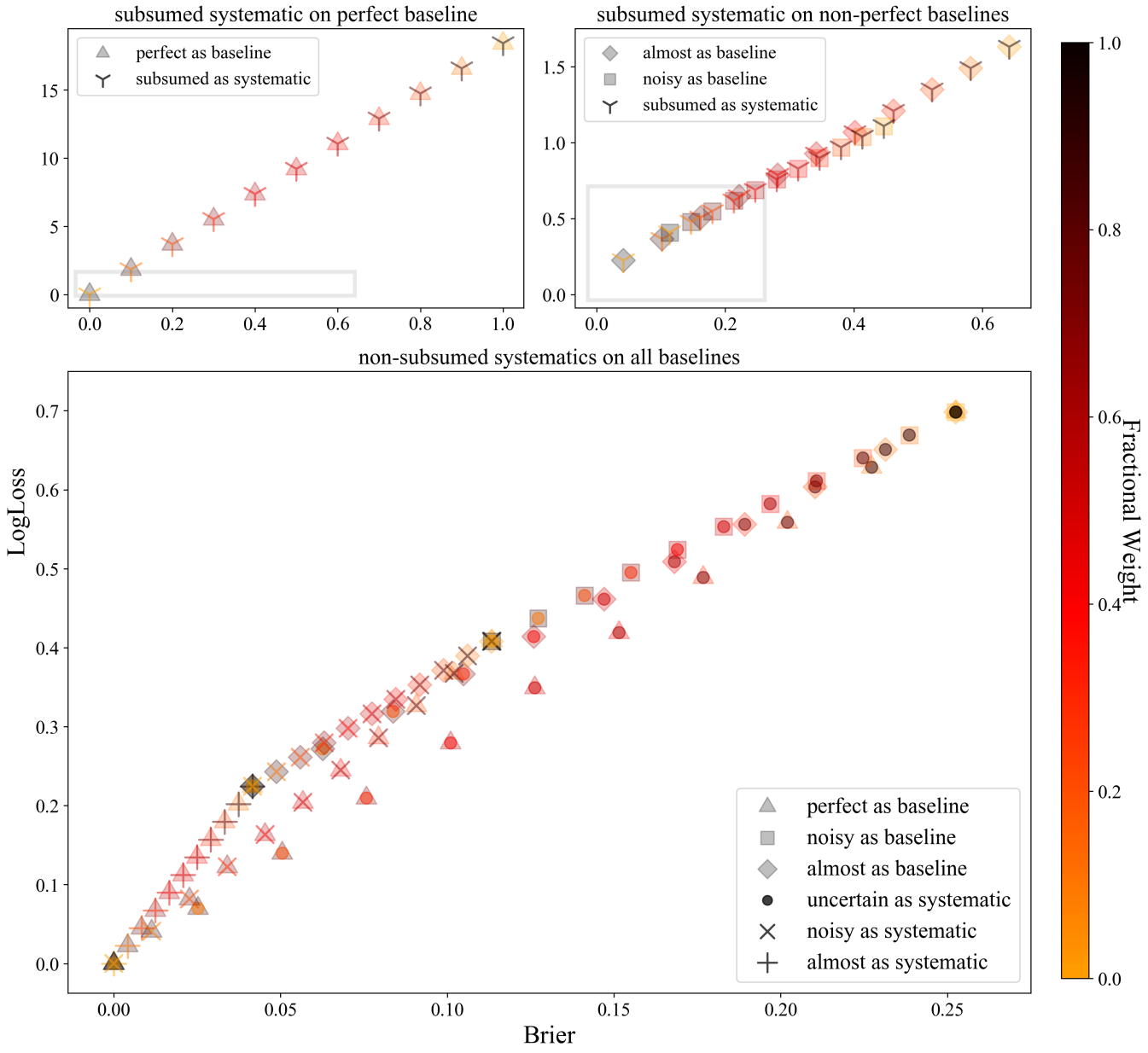

Figure 5. Weighted log-loss and Brier scores for baseline classifiers with combinations of systematics. Each point represents a classifier with a shared baseline behavior (regular polygon marker; triangle for perfect, diamond for almost perfect, square for noisy) for all but one class, which is affected by a particular systematic (asterisk markers; plus for almost perfect, cross for noisy, dot for uncertain, and Y-shape for subsumed). The color of the marker for the systematic effect indicates the weight on the one class affected by that systematic, while the color of the baseline behavior marker indicates the integrated weight evenly distributed over other classes with baseline behavior, where lower weights are greener and higher weights are bluer. From left to right, we zoom in on a particular range of scores, to highlight the scale of the effect of weighted systematics on the metrics for well-behaved methods with low Brier/log-loss values. The ranges of Brier score and log-loss values between the panels are in ratios of approximately 10:7:3 and 100:10:5, respectively, indicating the log-loss's higher sensitivity to the presence of systematics. The metrics are most sensitive to the subsuming systematic on a perfect baseline (triangle with Y-shaped marker), whereas other combinations of baseline and systematic can be grouped with a smaller dynamic range in both metrics.

Classifier characteristic	Brier score	Log-loss
Perfect	0.0	0.0
Almost perfect	0.042	0.225
Noisy	0.113	0.408
Uncertain	0.253	0.699
Subsumed from Noisy	0.447	1.109
Subsumed from Almost	0.641	1.629
Subsumed from Perfect	1.0	18.421[a]

[a]The entry for the log-loss of a classifier that subsumes a class into one that is otherwise perfectly classified should be infinite but is bounded by the numerical precision of our calculations.

Table 1. ~~The value of each metric when the weight is entirely on the class with the indicated characteristic. Weighting changes the metric performance: the value of each metric when the weight is entirely on the class with the indicated characteristic (corresponding to a $w = 1$ case in Figure 5). The log-loss is more sensitive than the Brier score, with larger values of the score (indicating poor classification performance), particularly for the subsuming systematic.~~ Metric values computed using Equation 5 with unit weights for the mock data produced by mock classification schemes described in Sec. 2.1. While the log-loss metric has a larger dynamic range than the Brier score for poor classification, the toy classifiers would be ranked the same way by either metric.

	Systematics			
Baselines	Subsumed	Uncertain	Noisy	Almost
Perfect	18.421	2.763	3.601	5.387
Almost perfect	2.343	2.246	2.556	
Noisy	2.102	2.085		

Table 2. The slopes for each baseline-plus-systematic pair in the space of log-loss versus Brier score. A higher slope corresponds to increased sensitivity of the log-loss over the Brier score. The contrast between log-loss and Brier score is highest on a baseline of the perfect classifier, meaning the log-loss may be more appropriate for discriminating between classifiers that are already extremely good.

On the other hand, the log-loss can be seen as more susceptible to the tunnel vision classifier because its value improves sharply with any move toward perfection. If the subsumed class has little weight, the metric values are quite low, moreso for the log-loss than the Brier score. This means that under a population-proportional weighting scheme, it would not be penalized for subsuming an uncommon class if it performed well for a more common class, a situation that would not serve the needs of the astronomical community.

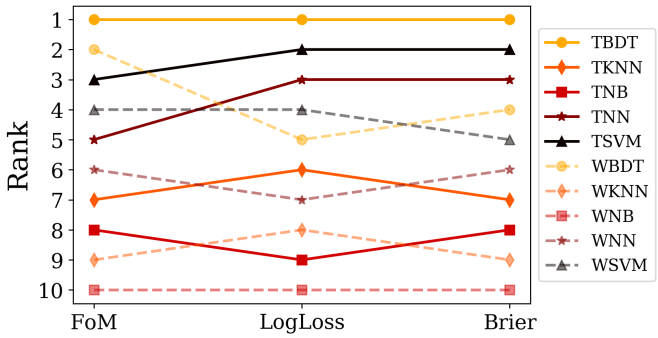

Figure 6. The rankings of each of ten `snmachine` classifiers with equal weight per object under the three metrics. The metrics broadly agree on the ranking of the classifiers, confirming consistency between a conventional metric of classification performance and the metrics of probabilistic classifications presented here. However, there are some differences with pairwise swapping between the log-loss and Brier rankings and some significant reordering of ranks 2 through 5 with the FoM metric relative to the probabilistic metrics.

4.2. *Representative classifications*

We apply the log-loss and Brier metrics to the classification output from `snmachine`. While the classification methods described in Lochner et al. (2016) refer to the idealized subset of the SNPhotCC data, these approaches are the state-of-the-art in classification of extragalactic transients. We present in ~~Table 6~~Figure 6 the rankings under the log-loss and Brier score metrics assuming an equal weight per object. ~~Table 6 also contains the ranking of classifier performance under each metric.~~

We apply our metrics to the classification output from `snmachine` applied to the SNPhotCC dataset as an example of representative light curves and representative classifiers used in extragalactic astronomy. We present in ~~Table 6~~Figure 6 the rankings of each classifier under the log-loss and Brier scores assuming an equal weight per object, as well as the original SNPhotCC metric described in Section 3.1. ~~Table 6 also contains the ranking of classifier performance under each metric.~~

The Brier score, log-loss, and SNPhotCC FoM are in agreement as to the first- and last-ranked classifiers. This consensus indicates that both of the potential PLAsTiCC metrics are roughly consistent with our intuition about what makes a good classifier, providing an anchor between accepted notions of an appropriate metric and the metrics of probabilistic classifications under consideration here. One should be careful not to generalize, however, as the rankings under the three metrics are not identical.

We note that the FoM differs more from the Brier score and log-loss metrics than they do from one an-

other. This is perhaps unsurprising, given that the SNPhotCC was specifically looking to value classification algorithms that were pure (that yielded a large number of SNIa classifications and few interlopers from the other classes), as opposed to metric that rewards good performance across classes.

5. DISCUSSION

The goal of this work is to identify the metric most suited to PLAsTiCC, which seeks classification posteriors of complete light curves similar to those anticipated from LSST, with an emphasis on classification over all types, rewarding a "best in show" classifier rather than focusing on any one class or scientific application.[14] The weighted log-loss is thus the metric most suited to the current PLAsTiCC release.

~~Future releases of PLAsTiCC will focus on different challenges in transient and variable object classification, with metrics appropriate to identifying methodologies that best enable those goals. We discuss approaches to identifying optimal metrics for these variations, which may be developed further in future work.~~ Transient and variable object classification is crucial for a variety of scientific objectives. The impact of a shared performance metric on this diversity of goals leads to complex and covariant trade-offs, which thus must be evaluated using multiple metrics. While a detailed accounting of these possibilities for future releases of PLAsTiCC and the selection of appropriate metrics for individual science cases are outside the scope of this first investigation, we discuss below some issues concerning the identification of metrics for a few example science cases.

5.1. ~~*Early classification*~~ *Ongoing transient follow-up*

Spectroscopic follow-up is only expected of a small fraction of LSST's detected transients and variable objects due to limited resources for such observations. In addition to optical spectroscopic follow-up, photometric observations in other wavelength bands (near infrared and x-ray from space; microwave and radio from the ground) or at different times will be key to building a physical understanding of the object, particularly as we enter the era of multi-messenger astronomy with the added possibility of optical gravitational wave signatures. Prompt follow-up observations are highly informative for fitting models to the light curves of familiar source classes and to characterizing anomalous light curves that could indicate never-before-seen classes that have eluded identification due to rarity or faintness. As such, decisions about follow-up resource allocation must be made quickly and under the constraint that resources wasted on a misclassification consume the budget remaining for future follow-up attempts. A future version of PLAsTiCC focused on early light curve classification should have a metric that accounts for these limitations and rewards classifiers that perform better even when fewer observations of the lightcurve are available.

We consider the decision of whether to initiate follow-up observations to be binary and deterministic. However, it is possible to conceive of non-binary decisions about follow-up resources; for example, one could choose between dedicating several hours on a spectroscopic instrument following up on one likely candidate or dedicating an hour each on several less likely candidates. Here, we will discuss a metric for an early classification challenge to be focused on deterministic classification because the conversion between classification posteriors and decisions is uncharted territory that we do not explore at this time.

Even within the scope of spectroscopic follow-up as a primary motivation for early light curve classification, the goals of model-fitting to known classes and discovery of new classes would likely not share an optimal metric. The critical question for choosing the most appropriate metric for any specific science goal motivating follow-up observations is to maximize information. We provide two examples of the kind of information one must maximize via early light curve classification and the qualities of a deterministic metric that might enable it.

5.2. *Spectroscopic supernova cosmology*

Supernova cosmology with spectroscopically confirmed light curves benefits from true positives, which contribute to the constraining power of the analysis by including one more data point; when the class in which one is interested is as plentiful as SN Ia and our resources limited a priori, we may not be concerned by a high rate of false negatives. False positives, on the other hand, may not enter the cosmology analysis, but they consume follow-up resources, thereby depriving the endeavor of the constraining power due to a single SN Ia.

A perfect classifier would lead to a maximum amount of information about the cosmological parameters conditioned on the follow-up resource budget. ~~For this scientific application, the metric must be chosen to balance the value of the information forgone by a false positive and the value of information forgone~~

[14] At the conclusion of PLAsTiCC, other metrics specific to scientific uses of one or more particular classes will be used to identify "best in class" classification procedures that will be useful for more targeted science cases.

~~by a false negative, and the value placed on these is effectively weighted by the value we as researchers place on follow-up resources.~~ In this scientific application, a classifier that maximizes true positives and minimizes false positives boosts the constraining power over cosmological parameters. However, it does so at a cost of rising false negatives, which represent constraining power forgone. As this tradeoff is asymmetric, it is insufficient to consider only the true and false positive and negative rates, as the SNPhotCC FoM does, without propagating their impact on the information gained about the cosmological parameters.

5.3. *Anomalous transient and variable detection*

A particularly exciting science case is anomaly detection, the discovery of entirely unknown classes of transient or variable astrophysical sources, or distinguishing some of the rarest types of sources from more abundant types. Like the case of spectroscopic supernova cosmology discussed above, anomaly detection also gains information only from true positives, but the cost function is different in that the potential information gain is unbounded when there is no prior information about undiscovered classes. ~~An example would be the recent detection of a kilonova, flagged initially by the detection of gravitational waves from an object.~~ The discovery of pulsars serves as an example of novelty detection enabled by a human classifier (Hewish et al. 1968; Bell Burnell 1969).

Resource availability for identifying new classes is more flexible, increasing when new predictions or promising preliminary observations attract attention, and decreasing when a discovery is confirmed and the new class is established. In this way, a false positive does not necessarily consume a resource that could otherwise be dedicated to a true positive, and the potential information gain is sufficiently great that additional resources would likely be allocated to observe the potential object. Thus, a metric ~~tuned to~~ for evaluating anomaly detection would aim to *minimize the false negative rate and maximize the true positive rate.*

5.4. *Difficult light curve classification*

Photometric light curve classification may be challenging for a number of reasons, including the sparsity and irregularity of observations, the possible classes and how often they occur, and the distances and brightnesses of the sources of the light curves. These factors may represent limitations on the information content of the light curves, but appropriate classifiers may be able to overcome them to a certain degree.

Though quality cuts can eliminate the most difficult light curves from entering samples used for science applications, such a practice discards information that may be of value under an analysis methodology leveraging the larger number of light curves included in a sample without cuts. Thus, classification methods that perform well on light curves characterized by lower signal-to-noise ratios are specially important for exploiting the full potential of upcoming surveys like LSST.

This version of PLAsTiCC implements quality cuts to homogenize difficulty to some degree, and notions of classification difficulty may depend on information that will not be available until after the challenge concludes. While the groundwork for a metric incorporating data quality has been laid by Wu et al. (2018), we defer to future work an investigation of this possibility.

6. CONCLUSION

As part of the preparation for PLAsTiCC we investigated the properties of metrics suitable for probabilistic light curve classifications in the absence of a single scientific goal. Therefore, we sought a metric that avoids reducing classification probabilities to deterministic labels and is compatible with a multi-class, rather than binary (two-class), setting. In line with the goals of PLAsTiCC, an important desideratum was to have a metric that tends to reward a classifier's performance across all classes over a classifier that performs well on a small subset of the classes and poorly on others. Given the potential of large class imbalance in astronomical datasets, we were also interested in the possibility of up-weighting the importance of certain rarer transient classes if need be; consequently we wanted to understand the way the metric would behave with the use of per-class weights.

~~We compared two metrics specific to probabilistic classifications: the Brier score and the log-loss.~~ Our experimental design considers simulated classification submissions from a set of mock classifier archetypes expected of generic transient and variable classifiers. To start with, we identified two metrics of multi-class classification probabilities established in the literature: the Brier score and the log-loss. We left aside popular metrics (such as accuracy, true/false positive/negative rates, and AUC functions thereof) which did not satisfy these criteria, even though it is in principle possible to extend such metrics for these scenarios. The Brier score and the log-loss metrics are structurally and conceptually different, with wholly different interpretations. The Brier score is a sum of square differences between probabilities; the explicit penalty term is an attractive feature, but it treats probabilities as generic scores. The log-loss on the other hand is readily interpretable, meaning the

metric itself could be propagated into forecasting the cosmological constraining power of LSST, affecting the choice of observing strategy.

We evaluated these metrics using the simulated classification probability submissions from the classifier archetypes with unit weights and then by varying the weights in Equation 5. In the absence of per-class weights, both the Brier score and the log-loss metrics are susceptible to rewarding a classifier that performs well on the most prevalent class and poorly on all others, which fails to meet the needs of PLAsTiCC's diverse motivations. On the basis of the mock classifier rankings under equal per-class weights, we found that both metrics reward the classifiers that are better and penalize those that are worse, where better and worse are defined by our common intuition, yielding the same rankings under either metric and demonstrating that both could be appropriate for PLAsTiCC.

Even though the Brier score and log-loss metrics take values consistent with one another, they are structurally and conceptually different, with wholly different interpretations. The Brier score is a sum of square differences between probabilities; the explicit penalty term is an attractive feature, but it treats probabilities as generic scores and is not interpretable in terms of information. The log-loss on the other hand is readily interpretable, meaning the metric itself could be propagated into forecasting the constraining power of LSST, affecting the choice of observing strategy. ~~We discovered that the log-loss is somewhat more sensitive to the systematic errors in classification that we find most concerning for generic scientific applications.~~ While both metrics could be appropriate for PLAsTiCC, the log-loss is preferable due to its interpretability in terms of information. ~~Both metrics are susceptible to rewarding a classifier that performs well on the most prevalent class and poorly on all others, which fails to meet the needs of PLAsTiCC's diverse motivations.~~

Due to our desire to potentially upweight rare classes, we explored a weighted average of the metric values on a per-class basis as a possible mitigation strategy to incentivize classifying uncommon classes, effectively "leveling the playing field" in the presence of highly imbalanced class membership. While modifyinging the log-loss metric to handle weights for different classes diminishes its interpretability, it can still be understood as information gain subject to the value we as scientists place on knowledge stemming from each class.

Given that both log-loss and Brier score passed the basic sanity tests for PLAsTiCC, there was no need to devise new metrics built upon established metrics of binary or deterministic classification. Since both were deemed appropriate, we chose the weighted log-losss metric due to its possibility of interpretation in terms of information theory, at least in the limit of equal weights ~~Although weights do impact the interpretability of the log-loss, we select a per-class weighted log-loss as the optimal choice for PLAsTiCC.~~

We conclude by noting that care should be taken in planning future open challenges to ensure alignment between the challenge goals and the performance metric, so that efforts are best directed to achieve the challenge objectives. It is our hope hope that this study of metric performance across a range of systematic effects and weights may serve as a guide to approaching the problem of identifying optimal probabilistic classifiers for general science applications.

ACKNOWLEDGMENTS

Author contributions are listed below.
A.I. Malz: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, supervision, validation, visualization, writing - editing, writing - original draft
R. Hložek: data curation, formal analysis, funding acquisition, investigation, project administration, software, supervision, validation, visualization, writing - editing, writing - original draft
T. Allam Jr: investigation, software, validation, writing - original draft
A. Bahmanyar: formal analysis, investigation, methodology, software, writing - editing, writing - original draft
R. Biswas: conceptualization, methodology, software, supervision, writing - editing, writing - original draft
M. Dai: writing - editing
L. Galbany: writing - editing
E.E.O. Ishida: conceptualization, project administration, supervision, writing - editing
S.W. Jha: writing - editing
D. Jones: software
R. Kessler: writing - editing
M. Lochner: conceptualizaton, data curation, formal analysis, visualization, writing - editing
A.A. Mahabal: data curation, software, writing - editing, writing - original draft
K.S. Mandel: conceptualization, supervision, writing - editing
J.R. Martínez-Galarza: data curation, software, visualization, writing - original draft
J.D. McEwen: conceptualization, investigation, supervision
D. Muthukrishna: data curation, validation
G. Narayan: data curation, formal analysis
H. Peiris: conceptualization, funding acquisition, super-

vision
C.M. Peters: writing - editing
K. Ponder: visualization, writing - editing
C.N. Setzer: conceptualization, software

This paper has undergone internal review in the LSST Dark Energy Science Collaboration. The authors would like to thank Melissa Graham, Weikang Lin, and Chad Schafer for serving as the LSST-DESC publication review committee, as well as Tom Loredo for other helpful feedback. The authors also express gratitude to the anonymous referee for substantive suggestions that improved the paper.

Software: jupyter (Kluyver et al. 2016), matplotlib (Hunter 2007), numpy (Oliphant 2006, 2007; Walt et al. 2011), proclam (Malz 2018), scikit-learn (Pedregosa et al. 2011), scipy (Jones et al. 2001; Buitinck et al. 2013)

AIM is advised by David W. Hogg and was supported by National Science Foundation grant AST-1517237. TA is supported in part by STFC. RB and CS are supported by the Swedish Research Council (VR) through the Oskar Klein Centre. Their work was further supported by the research environment grant "Gravitational Radiation and Electromagnetic Astrophysical Transients (GREAT)" funded by the Swedish Research council (VR) under Dnr 2016-06012. AAM was supported in part by the NSF grants AST-0909182, AST-1313422, AST-1413600, and AST-1518308, and by the Ajax Foundation

The financial assistance of the National Research Foundation (NRF) towards 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. This work is partially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 306478-CosmicDawn.

Canadian co-authors 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.

We acknowledge the University of Chicago Research Computing Center for support of this work. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This research at Rutgers University is supported by US Department of Energy award DE-SC0011636.

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & 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.

REFERENCES

Abraham, S., Aniyan, A. K., Kembhavi, A. K., Philip, N. S., & Vaghmare, K. 2018, Mon Not R Astron Soc, 477, 894

Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2016, Mon Not R Astron Soc, 456, 2260

Bell Burnell, J. 1969, Thesis, Department of Radio Astronomy, University of Cambridge, Cambridge, UK, doi:10.17863/CAM.4926

Bethapudi, S., & Desai, S. 2018, Astronomy and Computing, 23, 15

Bloom, J. S., Richards, J. W., Nugent, P. E., et al. 2012, PASP, 124, 1175

Brier, G. W. 1950, Mon. Wea. Rev., 78, 1

Buitinck, L., Louppe, G., Blondel, M., et al. 2013, arXiv:1309.0238 [cs]

Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2017, ApJ, 836, 97

Conley, A., Sullivan, M., Hsiao, E. Y., et al. 2008, ApJ, 681, 482

Crown, M. D. 2012, Space Weather, 10, doi:10.1029/2011SW000760
Dieleman, S., Willett, K. W., & Dambre, J. 2015, Mon Not R Astron Soc, 450, 1441
Djorgovski, S. G., Mahabal, A. A., Donalek, C., et al. 2012, in 2012 IEEE 8th International Conference on E-Science, 1–8
Djorgovski, S. G., Donalek, C., Mahabal, A., et al. 2011, arXiv:1110.4655 [astro-ph, physics:physics]
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Sol Phys, 293, 28
George, D., Shen, H., & Huerta, E. A. 2018, Phys. Rev. D, 97, 101501
Gieseke, F., Polsterer, K. L., Thom, A., et al. 2010, in 2010 Ninth International Conference on Machine Learning and Applications, 352–357
Harvey, D., Kitching, T. D., Noah-Vanhoucke, J., Hamner, B., & Salimans, T. 2013, arXiv:1311.0704 [astro-ph, physics:physics]
Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., & Collins, R. A. 1968, Nature, 217, 709
Hon, M., Stello, D., & Yu, J. 2017, Mon Not R Astron Soc, 469, 4578
—. 2018a, Mon Not R Astron Soc, 476, 3233
Hon, M., Stello, D., & Zinn, J. C. 2018b, ApJ, 859, 64
Hoyle, B. 2016, Astronomy and Computing, 16, 34
Hunter, J. D. 2007, Computing in Science Engineering, 9, 90
Ishida, E. E. O., Souza, D., & S, R. 2013, Mon Not R Astron Soc, 430, 509
Jones, D. O., Scolnic, D. M., Riess, A. G., et al. 2018, ApJ, 857, 51
Jones, E., Oliphant, T., & Peterson, P. 2001, SciPy: Open Source Scientific Tools for Python
Kessler, R., Conley, A., Jha, S., & Kuhlmann, S. 2010a, arXiv:1001.5210 [astro-ph]
Kessler, R., Bassett, B., Belov, P., et al. 2010b, PASP, 122, 1415
Kim, E. J., & Brunner, R. J. 2017, Mon Not R Astron Soc, 464, 4463
Kim, E. J., Brunner, R. J., & Carrasco Kind, M. 2015, Mon Not R Astron Soc, 453, 507
Kitching, T., Amara, A., Gill, M., et al. 2011, Ann. Appl. Stat., 5, 2231
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in ELPUB, 87–90
Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31
Mahabal, A., Sheth, K., Gieseke, F., et al. 2017, in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 1–8
Mahabal, A., Djorgovski, S. G., Turmon, M., et al. 2008, Astronomische Nachrichten, 329, 288
Malz, A. I. 2018, ProClaM, doi:10.5281/zenodo.3352639
Malz, A. I., Marshall, P. J., DeRose, J., et al. 2018, AJ, 156, 35
Mandelbaum, R., Rowe, B., Bosch, J., et al. 2014, ApJS, 212, 5
Mays, M. L., Taktakishvili, A., Pulkkinen, A., et al. 2015, Sol Phys, 290, 1775
Möller, A., Ruhlmann-Kleider, V., Leloup, C., et al. 2016, J. Cosmol. Astropart. Phys., 2016, 008
Morii, M., Ikeda, S., Tominaga, N., et al. 2016, Publ Astron Soc Jpn Nihon Tenmon Gakkai, 68, doi:10.1093/pasj/psw096
Murphy, K. P. 2012, Machine learning: a probabilistic perspective (The MIT Press)
Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, The Astrophysical Journal Supplement Series, 236, 9
Newling, J., Varughese, M., Bassett, B., et al. 2011, Mon Not R Astron Soc, 414, 1987
Nugent, P., Kim, A., & Perlmutter, S. 2002, PASP, 114, 803
Oliphant, T. 2007, Python for Scientific Computing, Vol. 9, doi:10.1109/MCSE.2007.58
Oliphant, T. E. 2006, A guide to NumPy (USA: Trelgol Publishing)
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J Machine Learning Res, 12, 2825
Richards, G. T., Myers, A. D., Peters, C. M., et al. 2015, ApJS, 219, 39
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32
Roberts, E., Lochner, M., Fonseca, J., et al. 2017, J. Cosmol. Astropart. Phys., 2017, 036
Rubin, D., Aldering, G., Barbary, K., et al. 2015, ApJ, 813, 137
Sako, M., Bassett, B., Connolly, B., et al. 2011, ApJ, 738, 162
Walt, S. v., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science & Engineering, 13, 22
Wu, C., Wong, O. I., Rudnick, L., et al. 2018, arXiv:1805.12008 [astro-ph]
Zevin, M., Coughlin, S., Bahaadini, S., et al. 2017, Class. Quantum Grav., 34, 064003