Real-time, Value-driven Data Augmentation in the Era of LSST

Niharika Sravan, Dan Milisavljevic [ORCID], Jack M. Reynolds, Geoffrey Lentner, and Mark Linvill
Department of Physics and Astronomy, Purdue University, 525 Northwestern Ave., West Lafayette, IN 47907, USA

Abstract

The deluge of data from time-domain surveys is rendering traditional human-guided data collection and inference techniques impractical. We propose a novel approach for conducting data collection for science inference in the era of massive large-scale surveys that uses value-based metrics to autonomously strategize and coordinate follow-up in real time. We demonstrate the underlying principles in the Recommender Engine For Intelligent Transient Tracking (REFITT) that ingests live alerts from surveys and value-added inputs from data brokers to predict the future behavior of transients and design optimal data augmentation strategies given a set of scientific objectives. The prototype presented in this paper is tested to work given simulated Rubin Observatory Legacy Survey of Space and Time (LSST) core-collapse supernova (CC SN) light curves from the Photometric LSST Astronomical Time-series Classification Challenge data set. CC SNe were selected for the initial development phase as they are known to be difficult to classify, with the expectation that any learning techniques for them should be at least as effective for other transients. We demonstrate the behavior of REFITT on a random LSST night given ∼32,000 live CC SNe of interest. The system makes good predictions for the photometric behavior of the events and uses them to plan follow-up using a simple data-driven metric. We argue that machine-directed follow-up maximizes the scientific potential of surveys and follow-up resources by reducing downtime and bias in data collection.

Unified Astronomy Thesaurus concepts: Astronomy data acquisition (1860); Surveys (1671); Supernovae (1668); Core-collapse supernovae (304)

1. Introduction

A new paradigm of astronomical data collection has been realized over the past decade. Dozens of projects[1] probing the time-variable universe have led to new discoveries (e.g., Arcavi et al. 2016; Lunnan et al. 2017; Margutti et al. 2019; Perley et al. 2019), higher quality data sets and statistics (e.g., Graur et al. 2017; Brout et al. 2019) and improvements in our understanding of the physics governing transient phenomena. The advent of the Rubin Observatory Legacy Survey of Space and Time (LSST) (Tyson 2002; Ivezić et al. 2019) in the upcoming decade, delivering an order of magnitude higher volume of alerts, will be transformative. However, several limitations prevent us from fully exploiting the science potential of even current survey outputs.

A major obstacle is the real-time filtration, interpretation, and prioritization of survey alert streams to identify candidates for follow-up. Several data brokers (e.g., ALeRCE[2]; AMPEL (Nordin et al. 2019); Antares (Saha et al. 2014); Lasair (Smith et al. 2019); MARS[3]; Pitt-Google Broker[4]) have emerged recently with the goal of helping sort, cross-reference, and value-add survey alert streams (e.g., Narayan et al. 2018), with each broker having different scientific priorities and scope of what they will provide to their subscribers. Broker services are also supported by the recent development of software infrastructure, known as Target and Observation Managers or

marshals (e.g., Street et al. 2018; Kasliwal et al. 2019), that permit observers to sort through raw and broker alert streams to plan and trigger follow-up. Although they represent important components of the overall cyberinfrastructure required to support survey science (e.g., Huerta et al. 2019), these services alone will be unable to truly maximize the science potential of the full LSST survey alert stream as they still largely rely on humans to make decisions.

A second limitation is availability of follow-up resources. Full characterization of transients often requires spectroscopy and/or multiwavelength and multimessenger follow-up, but the availability of facilities remains relatively low, time-sensitive, and time-consuming. ZTF's Bright Transient Survey partly addresses this issue by using a dedicated low-resolution integral field unit spectrograph to obtain classification spectra for most ZTF transients <18.5 mag (Fremling et al. 2019; Graham et al. 2019). However, LSST will demand efficient value-based follow-up decisions that are adaptable to the diverse needs of potentially competing science goals (e.g., obtaining flash spectroscopy or planning UV/radio follow-up) for a much more massive alert stream.

The third limitation is that LSST's light-curve (LC) sampling will be insufficient to a priori be interpreted with theoretical models for many transients. LSST will conduct two surveys, the Wide Fast Deep (WFD) and the Deep Drilling Field (DDF), in six filters (bands): *u, g, r, i, z,* and *y*. However, ∼90% of LSST's time will be spent on the WFD survey, in which LSST is nominally expected to revisit the same field, in any band in ∼3 days and in the same band in ∼3 weeks (LSST Science Collaboration et al. 2017). Algorithms developed in response to the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC, The PLAsTiCC team et al. 2018; Kessler et al. 2019) have made significant advances in developing classification models for LSST LCs (e.g., Boone 2019; Gabruseva et al. 2019). Early-time classification

---

using partial LCs is also possible (Carrasco-Davis et al. 2019; Möller & de Boissière 2020; Muthukrishna et al. 2019). However, it is unclear to what extent LSST's LCs can be used to derive other types of science (e.g., explosion physics, stellar progenitor information).

We propose augmenting LSST's coverage with supporting facilities in real time using a value-based metric for constraining the science of interest. The challenge is to shift the burden of follow-up decisions from observers to machines that are both scalable and not subject to human error, fatigue, or bias. As a step toward this goal, in this paper we introduce the Recommender Engine For Intelligent Transient Tracking (REFITT), an autonomous system that ingests live alerts from surveys and value-added inputs from brokers, predicts their future behavior, and makes recommendations for the most valuable events and epochs for obtaining additional photometry, given the goal of most optimally augmenting the alert stream for science inference.

This paper is organized as follows. In Section 2 we provide an overview of the key elements needed for conducting value-driven data augmentation. In Section 3 we present our implementation of these elements in REFITT. In Section 4 we simulate the performance of REFITT for LSST, illustrating its predictions and follow-up recommendations. We discuss steps to improve REFITT and conclude in Section 5.

2. Machine-driven Data Augmentation

We designate the class of autonomous systems that strategize and coordinate value-driven follow-up in real time to be ORACLEs (Object Recommender for Augmentation and Coordinating Liaison Engine). In this section, we identify three essential components of any ORACLE.

1. *Signal prediction*: decisions about the utility of a measurement require value estimates. Given a time-series signal, future values can be estimated using theoretical or regression models. Regression models have the dual advantage of being flexible and not requiring an understanding of the signal generating processes. Machine learning regression models capitalize on past data for this purpose and have been demonstrated to be very successful at a wide variety of forecasting problems (e.g., Wang et al. 2015; Florios et al. 2018; Mukund et al. 2019).

2. *Utility estimation*: next is translation of signal expectation to the utility of obtaining a measurement. While simple and intuitive utilities can be chosen (e.g., utility proportional to prediction error to favor measurements at epochs with most uncertainty), in some cases it might be more appropriate to define measurement utility in relation to a target scientific model. Optimal design of experiments offers several utility functions for the goal of optimally constraining scientific models with respect to various statistical criteria (e.g., A- or D-optimality). They can be used to design follow-up (experiments) that most optimally constrains theoretical models of interest (see, e.g., Yang et al. 2015).

3. *Data acquisition*: finally, data must be collected in order of high expected utility up to a collection budget. While a fixed cost of measurement can be assumed, more sophisticated implementations require consideration of variable measurement cost (e.g., time for acquisition). For the former, given a set of observing agents (OAs), data collection tasks can either be allocated randomly or on

first-come first-served basis. However, given strong differences between OA abilities to acquire data allocated, even elementary matching to drive task allocation (e.g., considering weather at an OA site) would lead to a significant improvement in the quality of data obtained.

REFITT is a prototype of an ORACLE. It is a demonstration of the implementation of the above elements and seeks to provide a complete solution to the challenges of conducting real-time value-driven augmentation.

3. Recommender Engine for Intelligent Transient Tracking

3.1. System Overview

Figure 1 shows a schematic of the REFITT system. It ingests data streams from surveys and brokers and narrows down ongoing events within a chosen window of action. In this paper, the window of action is set at 30 days since the LSST trigger. Core-collapse supernova (CC SN) events (SNe II and SNe Ibc) from the PLAsTiCC database have been used to simulate the LSST alert stream. CC SN types are known to be difficult to characterize due to their fast temporal evolution and, for some subtypes, similarity to SNe Ia. The choice keeps the scope of the initial development focused but ensures that solutions can be generalized to other transient types.

Candidates within the window of action are passed to REFITT's AI for signal prediction. At its core, REFITT relies on a convolutional autoencoder (CAE) to match live events against a library of simulated events. For every live event, "k" most similar events in the reference library are chosen to predict future photometric behavior. If available, classification information from brokers is used to update the system's choice for similar library events. The "k" reference events are then used to estimate future photometric evolution of each event using a combination of dynamic time warping (DTW) alignment and Gaussian process (GP) regression. We discuss each of the above components in detail in Section 3.2.

Next, all events within the window of action are assigned a quality score based on prediction confidence. In this paper, we use a simple data-driven metric to determine if follow-up is needed at the next observing epoch and prioritize observations by prediction quality scores. The follow-up recommendation strategy is discussed in detail in Section 3.4.

Finally, a web application programming interface distributes a list of targets to OAs supporting REFITT according to expected photometric behavior of the target and individual OA's suitability for obtaining the observation (location, weather, aperture, instrumentation, etc.). Observations returned by OAs are fed back into the alert stream. Events whose behavior differs significantly from the system's expectations are audited and assimilated into its reference library. We leave an exploration of strategies for optimal training set assembly for future work (but see Ishida et al. 2019).

3.2. REFITT AI

3.2.1. Reference Library

REFITT uses a library of events with full photometric information within the window of action (30 days since the LSST trigger for this paper) to predict the behavior of live events. The library is built from a subset of CC SNe from the full PLAsTiCC data set (training and testing) that satisfy each of the following criteria:
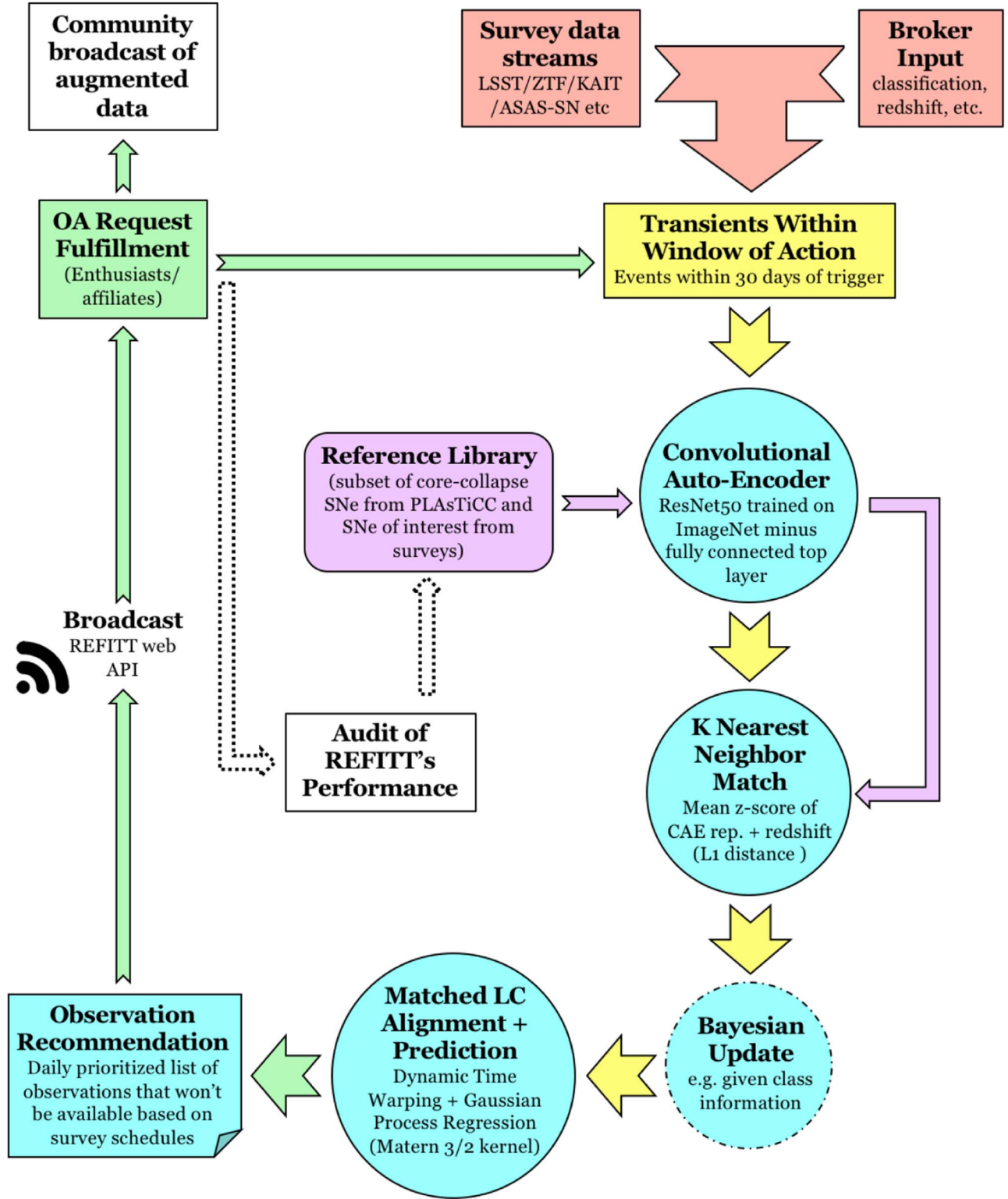
Figure 1. Overview of the REFITT system. It ingests data streams from surveys and inputs from data brokers and designs a follow-up strategy for optimally augmenting transient LCs for science inference. Blue circles indicate REFITT's AI components (discussed in 3.2). Size of arrows indicate volume of data flow while their colors (red to green) represent their information entropy (high to low). OAs (observing agents) collect data recommended by the system which are then fed back into the alert stream. OA submissions that deviate significantly from the system's expectations are assimilated into libraries it uses to make predictions.

1. In a given LSST band there exists at least one photometric measurement before and after peak flux with a mean flux value between one third to two thirds of the mean peak flux value;
2. The above is true in at least three bands; and

3. The above are satisfied within a 100 day window, 20 before and 80 days after the LSST trigger.

An LSST trigger is defined as the first photometric measurement in a time series where flux/$\sigma_{\rm flux} > 5$ (Kessler et al. 2019). Photometry with $\sigma_{\rm flux} > 50$ units is removed. These criteria
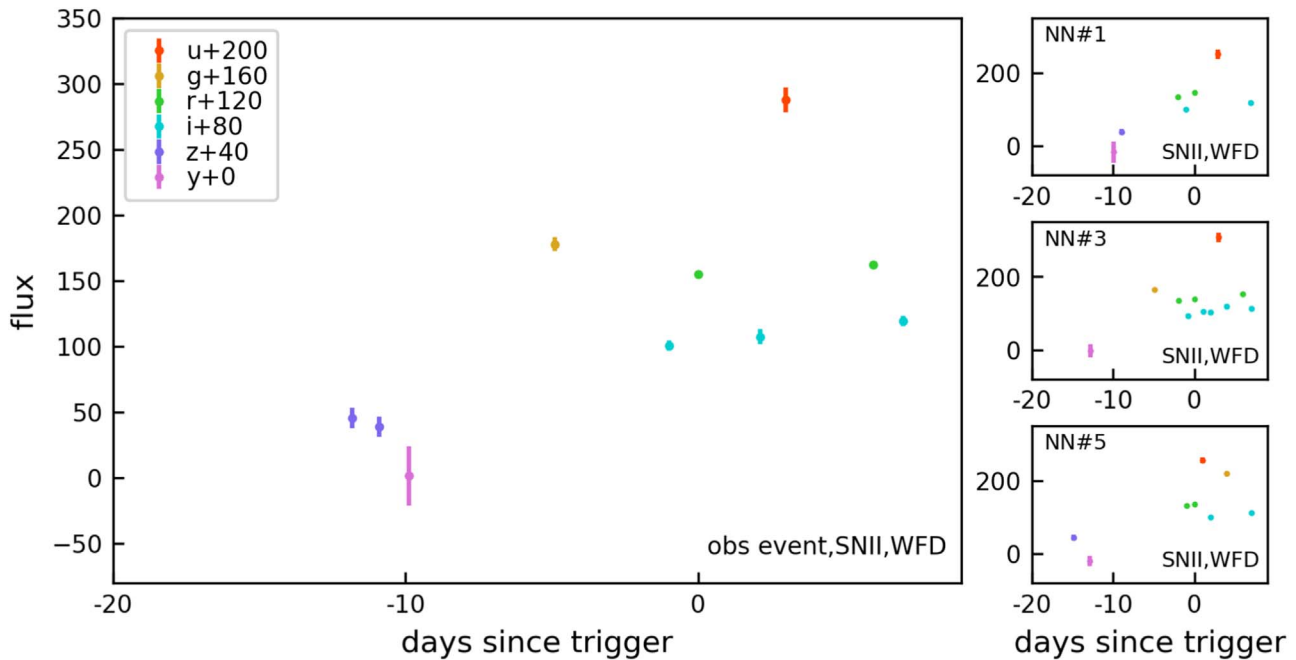
Figure 2. Example of reference LCs matched to a given live event. The main panel shows a SN II event from LSST's WFD survey at 9 days since trigger. Panels to the right show three similar reference events (nearest neighbors 1, 3, and 5 from top to bottom) identified. To select similar events we use the observed LC, event photometric redshift, and classification from a weak hypothetical classifier.

ensure that conducting GP regression on events in the reference library results in clean fits throughout the window of action. Performing the above selection on the PLAsTiCC data set yields 3228 CC SN events, 2312 of Type IIP and 916 of Type Ibc. The asymmetry in training samples for the two classes causes our predictions for SNe Ibc to be worse than for SNe IIP (see Section 4).

The PLAsTiCC data set is not ideal for the eventual goal of strategizing real-time follow-up. As mentioned above, a major reason for this is the imbalance in number of LCs for each transient type. Additionally, the accuracy of some of the LCs is a subject of debate (M. Modjaz 2020, personal communication) and needs to be explored further. PLAsTiCC serves as a convenient starting point for demonstrating essential components of the system and identifying limitations. We leave improving our reference library as an avenue of future development (see also discussion in Section 5).

We use the reference library for training. In this paper we use it to set the value of the hyperparameter "k," which is the number of nearest neighbors to use for predicting the future behavior of an event. Results for testing are shown in Section 4.

3.2.2. Convolutional Autoencoder

Events within the window of action are passed to a CAE to find vector representations for them. CAEs are a subset of dimensionality reduction methods that use convolutional neural networks (CNNs) to find vector representations of input data summarizing relational information. For example, given a red–green–blue image, layers in a well-trained CNN find edges or color blobs. CNNs are equivariant to translation (but not scale and rotation, Goodfellow et al. 2016), i.e., they are agnostic to the location of an object in an image. This is especially useful because we would like to characterize an event without

knowledge of its evolutionary phase (position on time axis) and true brightness (position on flux axis).

We adopt ResNet50 (He et al. 2015) minus its fully-connected top layer trained on ImageNet (Russakovsky et al. 2015) as our CAE. Recent investigations have shown that features learned by layers in trained deep neural networks generalize well to even distantly related tasks (Yosinski et al. 2014). Ideally, a dedicated CAE would be designed, as our current CAE is sensitive to the number and spacing of data points in a LC. Training pretrained ResNet50 plus a deconvolution block with LC pairs having the same shape but different photometric sampling should desensitize the CAE to survey cadence. However, for the purpose of this paper our current CAE is sufficient.

In order to supply LCs to our CAE, we first convert each PLAsTiCC event time series in six LSST channels to 20 $3 \times 200 \times 200$ vectors. For this we bin the LC in each LSST channel in 200 bins in both time and flux. We bin flux by evaluating the normal distribution for flux mean (with value 1) and uncertainties at bin centers. The flux range is fixed from -100 to 500 flux units. This choice spans the flux range for 98% of CC SNe in the PLAsTiCC data set. The range in time is fixed from 20 days before a trigger to the nearest integer day of the time since a trigger for that event. We then pass the 200×200 vectors in three LSST channel combinations to our CAE. The outputs from our CAE are stacked (always in the same order) and passed to system's matching component which is described next.

3.2.3. Light-curve Matching

REFITT uses CAE representations of ongoing events to find similar events in its reference library. It always computes L1 distances between CAE representations of live events and representations of library events at the same epoch. If photometric redshift is available from brokers, it additionally

computes L1 distances between scalar redshift values for live and library events. Therefore, given a live event, similarity scores for every library event are computed using the mean z-score of CAE representation and redshift distances. We find that L1 performs better than L2 as a similarity metric. This is expected since L1 is known to perform better than L2 in high dimensions (Aggarwal et al. 2001). We leave an exploration of the use of more sophisticated similarity metrics (e.g., peak signal-to-noise ratio, feature similarity, structural similarity) to future work.

REFITT also incorporates classification for events from brokers by performing a Bayesian update to its belief of the most similar library events for that event. Specifically, given classification c_κ for an event from classifier κ, it performs an update to its prior belief about the probability that the event actually belongs to class C to:

$$P(C|c_\kappa) = \frac{P(c_\kappa|C)P(C)}{P(c_\kappa)}, \qquad (1)$$

where

$$P(c_\kappa) = P(c_\kappa|C)P(C) + P(c_\kappa|\bar{C})P(\bar{C}) \\ + P(c_\kappa|\overline{C,\bar{C}})P(\overline{C,\bar{C}}). \qquad (2)$$

Here $P(C)$ is the REFITT's prior probability that an event belongs to class C, $P(\bar{C})$ that it belongs to class $\bar{C}$, and $P(\overline{C,\bar{C}})$ that it belongs neither to class C nor class $\bar{C}$. Similarly, $P(c_\kappa|C)$ is the likelihood that an event actually belongs to class C, given this classification by classifier κ, and so on.

For a given event, we define $P(C)$ as the fraction of events of class C among "k" most similar library events. $P(\bar{C})$ is $1-P(C)$ and $P(\overline{C,\bar{C}})$ is 0. For this paper, this means that we always assign an event to either one of two CC SN types. We note that such an assumption is very common in classifiers, including for most submitted in response to PLAsTiCC, i.e., they always assign an event to labels they train for.

In this paper, we assume classification information at every epoch is available via brokers from a hypothetical classifier (κ) with the following performance (metrics motivated by Boone 2019): $P(II_\kappa|II) = 0.6$, $P(II_\kappa|Ibc) = 0.1$, $P(Ibc_\kappa|Ibc) = 0.5$, and $P(Ibc_\kappa|II) = 0.05$. We note that currently these metrics represent the most optimistic scenario since they correspond to a classifier looking at complete LC information in many cases. In addition, in a real classifier these metrics would change as a function of time.

Figure 2 shows an example of similar library events matched to a live SN II event from LSST's WFD survey at 9 days since a trigger. The system was also provided the event redshift and class (from classifier κ). The system successfully exploits even sparse LC information to select similar library events. The live event is a library event and chosen to demonstrate REFITT's training in this and the next section.

3.2.4. Light-curve Prediction: Alignment and Regression

We use similar events in the reference library to predict the future LC for every live event within the window of action. There are two steps to do this: alignment and regression.

First, the observed and similar reference events are aligned using DTW. DTW finds the optimal pairing between pairs of points in two time series, such that their Euclidian distance is shortest. DTW alignment is robust to signal stretching. This is
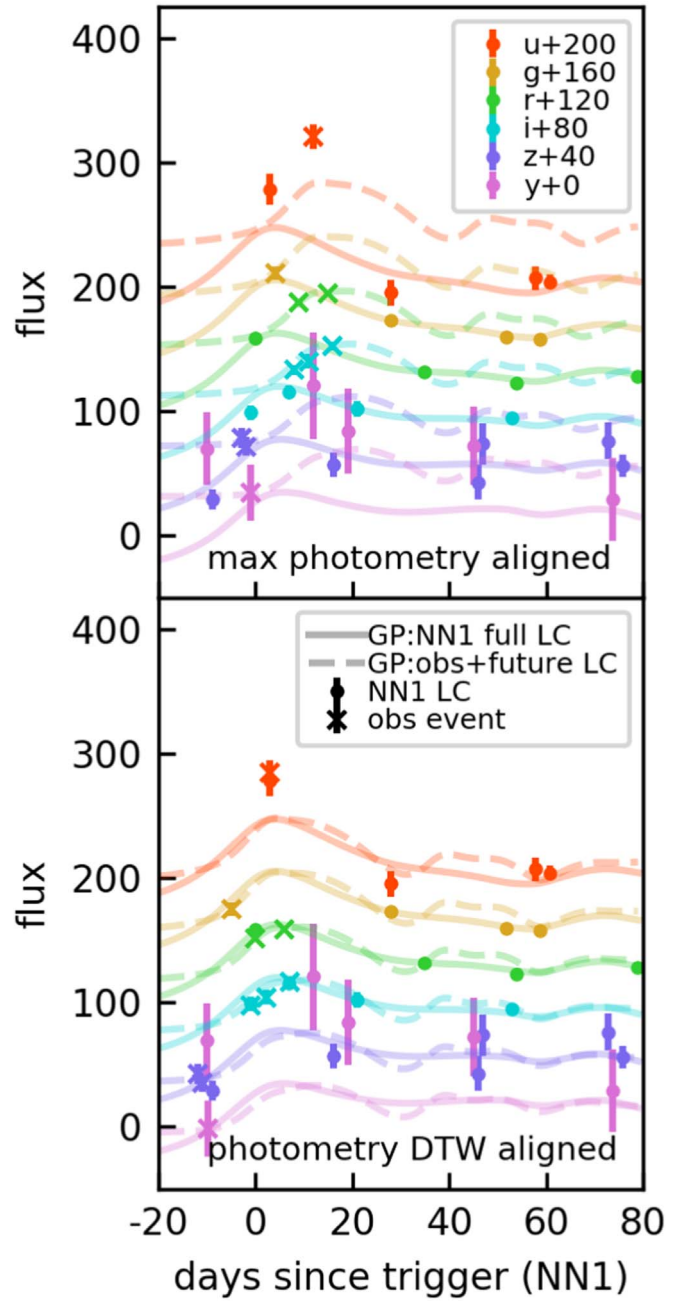

Figure 3. Comparison of the effect of using max photometry anchoring (top) and DTW (bottom) to align live observed and reference LCs. The observed LC (crosses) is for the same event as in the main panel of Figure 2. Dots show the full LC of the most similar reference event (upper-right panel of Figure 2). Note that the underlying fits are only shown to guide the eye and not used for alignment.

particularly useful for aligning LCs that could differ slightly (e.g., due to different progenitor or explosion properties). We use the metrics module of tslearn package (Tavenard 2017) to perform DTW alignment. We exclude noise-level photometry when aligning by excluding flux measurements with mean flux values <10% of maximum observed flux in both live and reference LCs. We then use the median difference in time and flux between pairs of DTW aligned photometry to align the two LCs.

Figure 3 shows the result of using DTW to align sparse LSST LCs. As expected, max photometry alignment can be very poor for sparse signals. DTW achieves good alignment
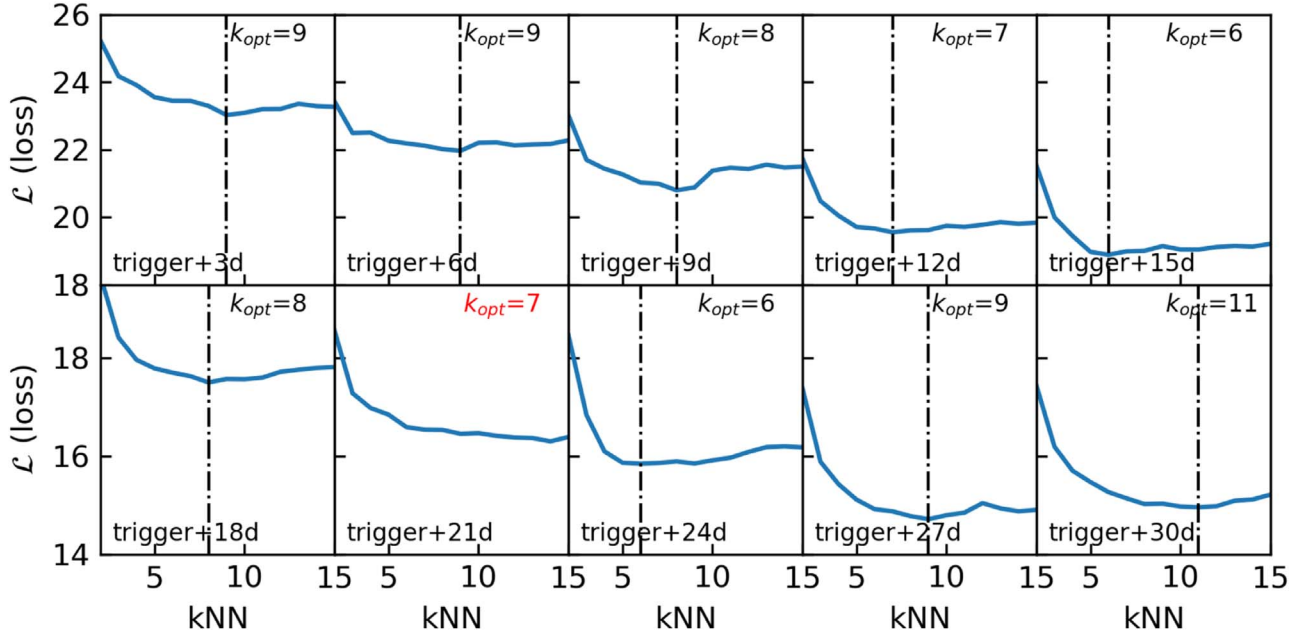
Figure 4. Leave-one-out cross-validation loss as a function of "k" nearest neighbors used for prediction at various integer epochs since an LSST trigger. Optimal values of "k" (k_{opt}) that correspond to minimum loss at every epoch are indicated. Our loss at 21 days since a trigger did not converge so we hard set the k_{opt} at this epoch to 7 assuming physical continuity from before and after.

without first needing to perform regression for resampling as would be needed to perform alignment using cross correlation. We compare the performance of DTW to cross-correlation alignment in the Appendix.

Next, we perform GP regression for reference LCs following the method introduced by Boone (2019). Briefly, we assume that LCs are described by a GP with mean 0 and a two-dimensional Matern-3/2 kernel. As in Boone (2019) we fix the length scale in wavelength to 6000 Å. The length scale in time is estimated via maximum likelihood. In this work, when flux estimated by a GP is <0 we set its value to 0. We also assume a saturating flux at 1000 units. The fits shown in Figure 3 are GP fits.

We use both steps, alignment and regression, for each reference LC to estimate the future LC for an observed event. Specifically, we estimate the mean and standard deviation of predictions from "k" most similar reference events. We ignore uncertainties from GP fits as these are quite small.

3.3. Training

Our architecture requires the setting of the hyperparameter "k," where "k" is the number of most similar reference events to use for predictions. We estimate "k" that minimizes the training loss:

$$\mathcal{L}(k, t) = \frac{1}{n(\mathcal{D})} \sum^{\forall \mathcal{D}} \frac{\sum_{i=1}^{N_{\text{future}}} |\sigma_{i,\text{pred}} - |\mu_{i,\text{pred}} - \mu_{i,\text{true}}||}{N_{\text{future}}}. \quad (3)$$

We compute a dedicated optimal "k" for every integer day since a trigger. Here $\mathcal{D}$ denotes the reference library. The effect of the first summation is performing leave-one-out cross validation (LOOCV). In the second term, for a given library event observed at a given time since a trigger (t), we compute the mean absolute difference between the standard deviation of the signal prediction (σ_{pred}) from the absolute error of signal

prediction at every future epoch with available photometry (N_{future}). The absolute error of the prediction is the absolute difference between the mean of the predicted flux (μ_{pred}) and the mean of the true observed flux (μ_{true}). In other words, $\mathcal{L}$ has the effect of selecting "k" such that, on average, REFITT's predicted signal uncertainty is close to the true error in its prediction. We only compute loss for events where μ_{pred} changes by at least 25 units in at least three LSST bands. This is to exclude low flux events that tend to have small σ_{pred} values and flat loss curves.

Figure 4 shows the LOOCV loss as a function of "k" most similar reference events used for prediction at various integer values of time since a trigger. As expected the overall loss decreases with increasing time since a trigger as more photometry become available and predictions improve. At 21 days since a trigger, our model is underfit, i.e., increasing "k" decreases bias without increasing variance. We hard set the optimal "k" at this epoch to 7, assuming physical continuity from epochs before and after.

To ensure robustness of the system to finding similar library events for prediction given noise, we carried out tests where we artificially increased or decreased flux uncertainties in LCs. The system is robust to this source of noise: for changes within ∼25%, the system selects the same neighbors for prediction. For changes up to ∼50%, at most one neighbor differs.

Finally, we use the optimal "k" values to compute the mode of predicted LC uncertainties expected for the next day (using 1 flux unit bins) in each band. The modal uncertainty is the typical uncertainty due to intrinsic phase space scatter in the library. We use this metric to strategize follow-up (discussed next).

3.4. Recommendations

REFITT makes a prediction for every event within its window of action. It then uses these predictions to design an
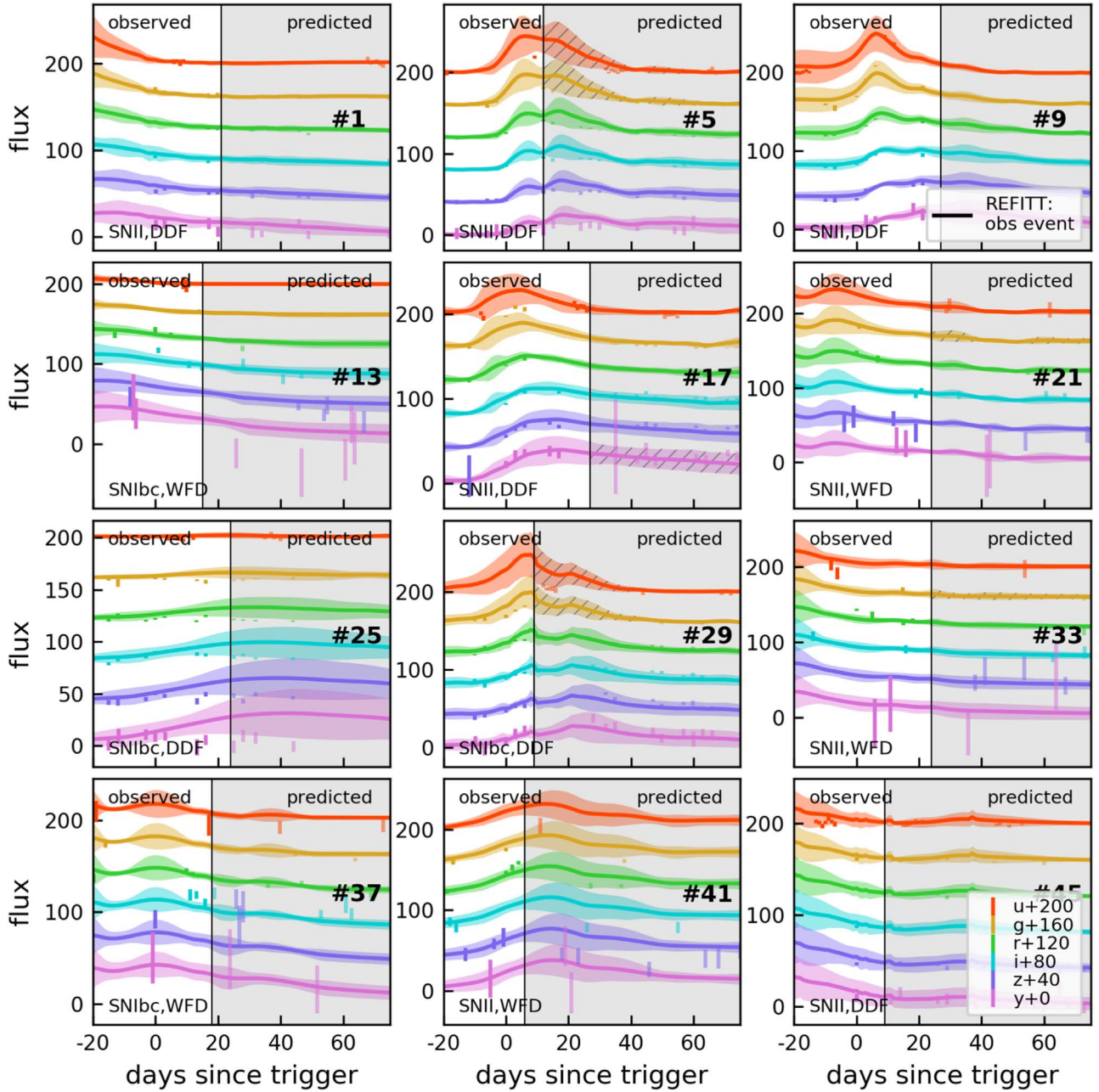
Figure 5. Simulation of REFITT's prediction, prioritization, and follow-up recommendation (hatched regions) on a random night of LSST operations. REFITT received ~32,000 CC SN events within the predefined window of action, and ranked them based on its prediction quality (rank in bold; SN subtype and survey type in the bottom left corners). For a given event, photometry in the white regions is observed and in the gray regions is yet to be observed. Solid lines and shaded regions are mean and 1σ uncertainties of predictions. Prediction uncertainties typically span the true error in predictions. REFITT does not require good LC coverage in order to make good predictions. The system makes good predictions for the event ranked 41 more than 6 days before peak in the *u*-band and for the event ranked 37 even though its peak is missed entirely.

observing strategy aimed at optimally augmenting the LCs. In this paper, we use a data-driven metric for strategizing follow-up (see also Section 5).

Each LC prediction is assigned a quality score based on how well it is known to have performed and expected to perform:

$$\mathcal{S} = \left(\sum_i^{N_{\mathrm{obs}}} \frac{|\mu_{i,\mathrm{pred}} - \mu_{i,\mathrm{obs}}|}{N_{\mathrm{obs}}} \right) + \alpha \left(\sum_{j=-20}^{80} \sigma_{j,\mathrm{pred}} \right), \qquad (4)$$

where indices j denote integer days since a trigger and α is a weighting factor. This score rewards predictions that have small mean observed errors and small overall uncertainties (from 20 days before up to 80 days after a trigger). In this work we set α to 1, however, it can be tuned to prioritize either metric. We only score events whose mean predicted flux changes by at least 25 flux units in at least one band between 20 before and 80 days after a trigger. This criterion is used to
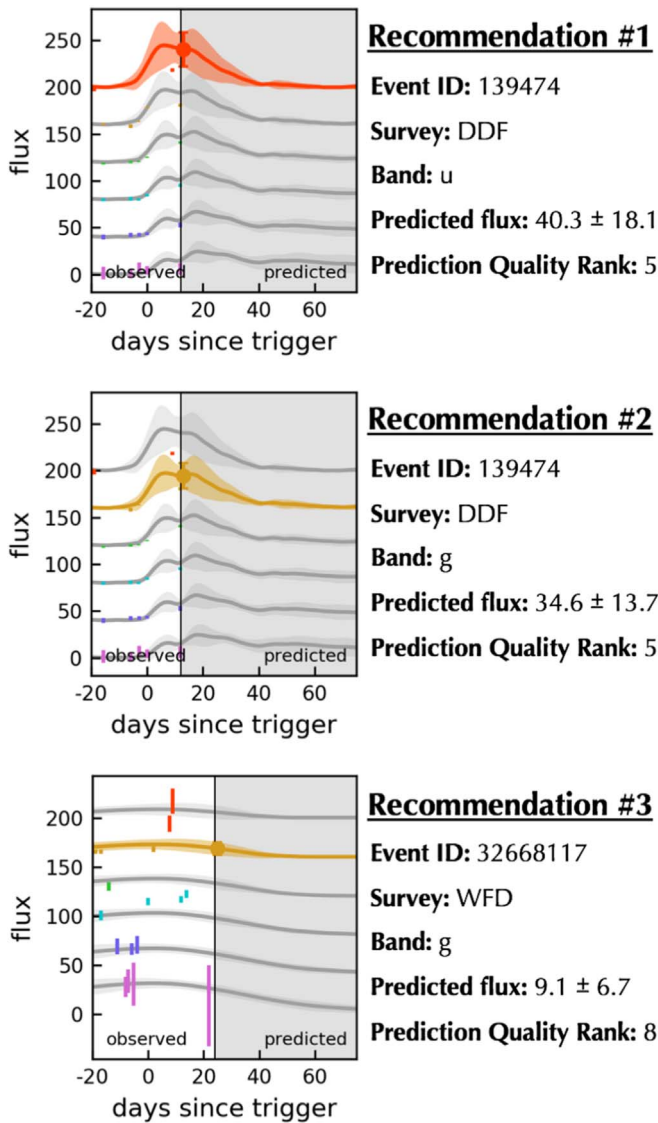
Figure 6. Example of REFITT's follow-up recommendations for the same LSST night as in Figure 5. For this figure we do not show photometry yet to be observed. Follow-up is recommended when the prediction uncertainty is higher than typical prediction uncertainties expected from training. Note that the top two recommendations are for the same event, with priority for the band with larger prediction uncertainty.

exclude events in their late fading phases and is arbitrary. We also only score events with >0.25 mean number of photometry per day to exclude events with several missing observations.

All events in the window of action are ranked according to $\mathcal{S}$. We then search down the ranked list to check whether the predicted uncertainty for the next day in any band is greater than modal uncertainty computed during training. Since the modal uncertainty represents the bulk of the scatter in predictions due to the library, uncertainties larger than it represent variations between neighbors themselves. After confirming that an observation at that epoch and band will not be available given LSST's and other survey schedules (in this paper whether data is available at that epoch in the event simulation), we flag it for follow-up in order of $\mathcal{S}$. If more than one band in an event meets the above criteria we prioritize the band with the larger prediction uncertainty.

4. REFITT in the Era of LSST

Figures 5 and 6 demonstrate REFITT's predictions, prioritization, and follow-up recommendations on a random night of LSST operations. We simulate the alert stream using CC SN events in the PLAsTiCC data set not in the reference library (i.e., test set), assuming they occur randomly over the 3 yr LSST operation simulated in PLAsTiCC. We simulate broker input by supplying event photometric redshift and class (from classifier κ). On a given night, the system considers ~32,000 CC SN events that are within the 30 days window of action.

REFITT is able to make good predictions for both SN II/Ibc subtypes and events from WFD/DDF surveys. The 1σ uncertainties typically span the true error in predictions which indicates successful training. REFITT does not require good LC coverage in order to make good predictions. For example, the system makes good predictions for the event ranked 41 in Figure 5 more than 6 days before peak in the u-band. Similarly, the system makes good predictions for the event ranked 37 even though its peak is missed entirely. For the event ranked 25 in Figure 5, an observation is not recommended despite the uncertainty in several bands being greater than the system's tolerance (modal uncertainty from training) because LSST is expected to revisit the object within the next 24 hr.

Generally, the system's predictions for SNe II are better than Ibc. This is due to the imbalance in the number of SN II and Ibc events in the reference library. Our reference library has more SN II than SN Ibc events by a factor of ~2.5 (see Section 3.2.1). Developing a reference library with uniform distribution for event properties would be crucial for the system to operate live and will be a central focus of future work (see Section 3.2.1 and Section 5).

Figure 7 demonstrates how REFITT updates its predictions and follow-up recommendations given new data. As expected, REFITT's predictions improve over time. The effect of new data can be clearly seen when the event shown is at 15 days since trigger, when the first u-band photometry become available. We note that the uncertainty at 3 days since trigger for this event is not typical for all events and can be better, especially for SN II events. Similarly, the fit achieved by 27 days since trigger is not typical and can be better for SN II (see examples in Figure 5) and worse for SN Ibc events. As discussed earlier, these differences are due to the imbalance between the number of reference events of the two SN subtypes in the reference library.

REFITT also updates its follow-up strategy given new data. In the early phases of the event shown in Figure 7, the system characterizes it as a SN II (bluer photometry) and recommends follow-up in $u, g, r,$ and i bands. Then given new data in g, r, and i bands over the next 6 days REFITT is able to improve its prediction and no longer recommends follow-up in r and i bands, though still characterizing the event as SN II. After 6 more days, given the first u-band data, the system begins characterizing the event more like a SN Ibc (redder photometry) with an increased uncertainty in its predictions. It also recommends follow-up in all bands from g through i. We underscore that the ideal confirmation of REFITT's predictive or learning ability would be with live data. Machine learning models learn to represent properties of training data. If these differ from real data in ways such as to be critical for the
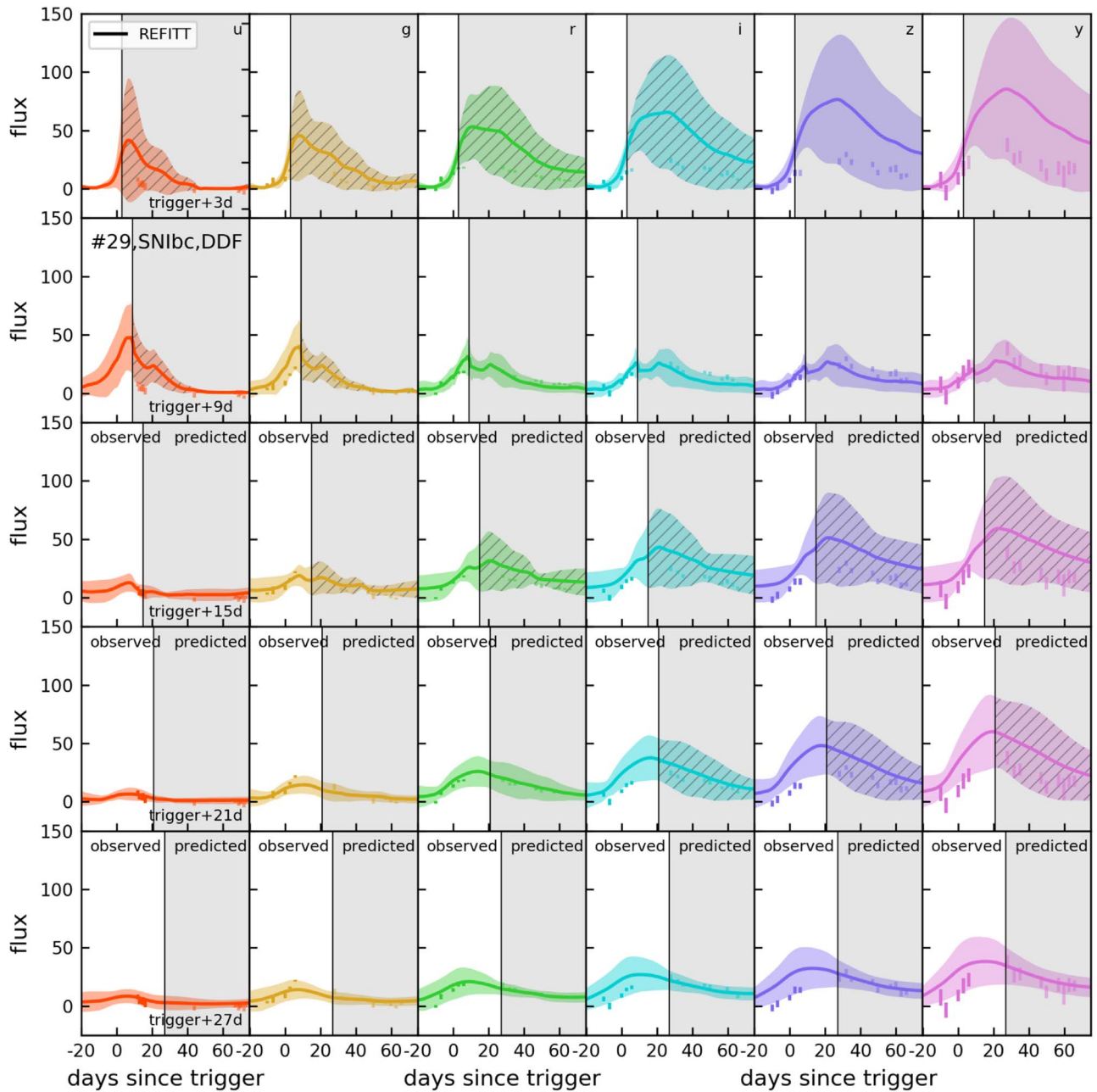
Figure 7. REFITT's prediction for a SN Ibc event from the DDF survey as a function of time since trigger (increasing downwards). As in Figure 5, photometry in white regions are observed and gray regions are yet to be observed. Solid lines, shaded and hatched regions have the same meaning as in Figure 5. Panels from left to right separate observations and predictions by band (*u–y*). This event is ranked 29 based on the prediction quality when at 9 days since trigger in the simulation shown in Figures 5 and 6. REFITT's predictions improve and its uncertainty decreases with time as new photometry is collected.

task the model is performing, it would respond poorly to real data despite performing remarkably well during training.

Figure 8 shows the distribution of properties of the top 500 events according to REFITT's prediction quality compared to those from the LSST survey. REFITT does not have a strong bias for either SN II/Ibc events or WFD/DDF surveys, The latter is desirable because it means that the system does not require high cadence photometry in order to make good predictions. Focusing on DDF events for augmentation might not be the most productive use of follow-up resources, as they may have enough photometry to allow them to be interpreted with theoretical models. The system also does not have a preference for late phase events (or longer times since a

trigger). This naturally favors data collection at earlier more profitable phases of an event.

5. Conclusions

LSST's unprecedented high-quality but sparsely-cadenced photometry requires real-time follow-up to maximize the science potential of the survey. Given the high cost for obtaining data, follow-up strategies need to be value driven as not all data are equally valuable for constraining science. We call systems that autonomously strategize and coordinate follow-up in real time ORACLEs. REFITT is a prototype of an ORACLE that uses predictive modeling for sparse multichannel LSST LCs and
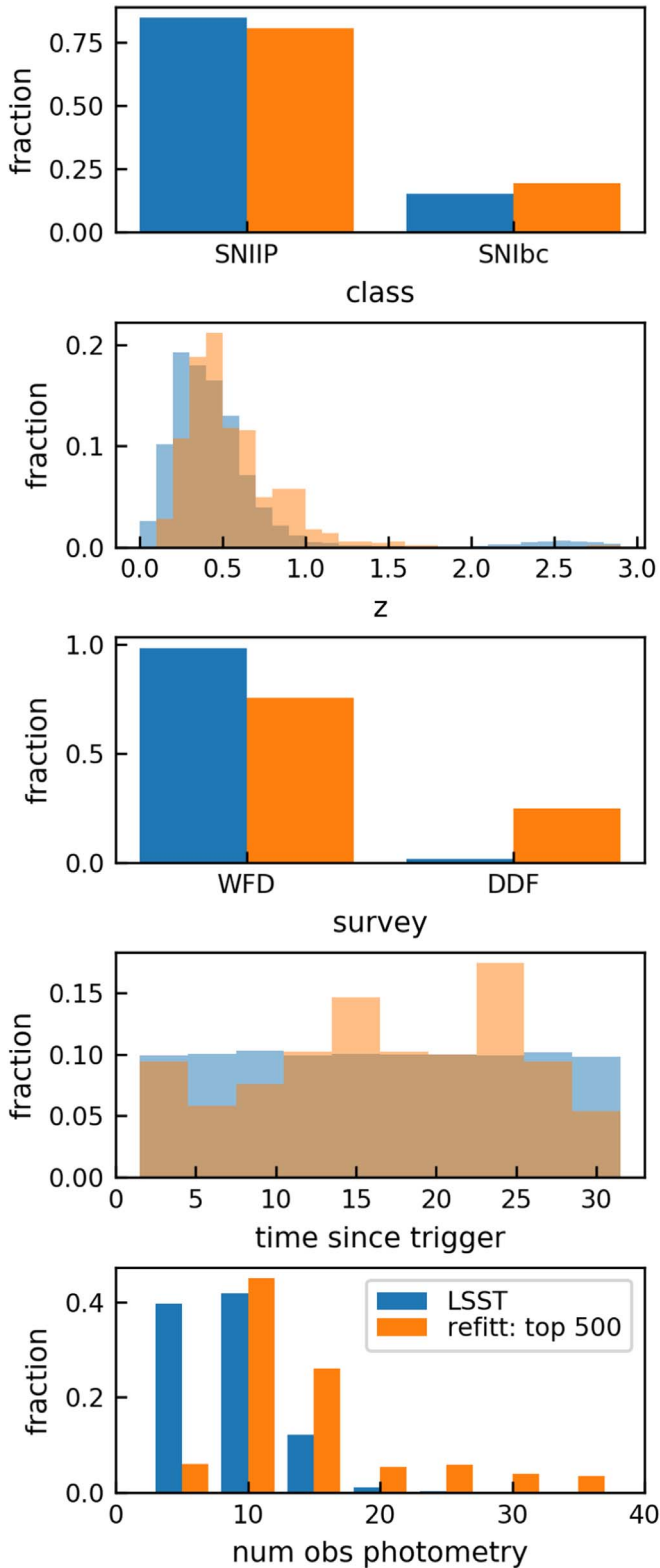
Figure 8. Distribution of properties of top 500 events according to REFITT's prediction quality (orange) compared to those from LSST (blue) on the same night as in Figures 5–7. REFITT does not exhibit a strong bias in the types of events it prioritizes for follow-up, however exhibits mild preference for events from the DDF survey and correspondingly more photometric points in an event. This is mostly due to the overrepresentation of DDF events in its reference library (due to our selection criteria listed in 3.2.1).

value-based metrics to strategize follow-up in real time. The system capitalizes on, without interfering with, survey schedules and is flexible to accommodate any science objective. It also seeks to optimally leverage global networks of telescopes (e.g., Astronomical Event Observatory Network, Las Cumbres Observatory, Brown et al. 2013; Global Relay of Observatories Watching Transients Happen,Kasliwal et al. 2019) for optimally augmenting valuable data.

The current implementation of REFITT can predict sparse LSST CC SN LCs and use them to design follow-up strategies. However, in order to operate on the full LSST survey stream, multiple improvements are desirable. First is the development of a robust reference library with LCs for diverse transient types, along with multiwavelength photometric and spectroscopic templates (e.g., Vincenzi et al. 2019). Second is the development of a dedicated CAE, potentially by fine-tuning the ResNet50 architecture (Zeiler & Fergus 2013; Oquab et al. 2014; Yosinski et al. 2014). Third, the system should ingest additional information from brokers (e.g., Milky Way extinction, galaxy association) to improve its predictions, since it is more likely that brokers will reliably be able to provide contextual information compared to real-time classification. These can be included as additional parameters when scoring reference event matches.

Finally, it may be beneficial to strategize follow-up to optimally constrain theoretical models of interest. One option would be to prioritize follow-up in order of increasing D-optimality (Kiefer & Wolfowitz 1959) of obtaining an additional observation at the next available epoch. Such an approach would maximize the Shannon information (Shannon 1948) about parameters describing a chosen theoretical model for the full alert stream.

Appendix
Comparison of Cross Correlation and DTW LC Alignment

A common technique used to align signals is cross correlation. The method requires both signals to have measurements at all comparison timestamps. For irregularly sampled signals, this can be accomplished by estimating the signals at regular time intervals (resampling) using a fitting procedure.

Figure A1 shows the result of using cross correlation to align LCs. We use our GP regression to resample both LCs in 0.1 day intervals. Then the median optimal timestamp estimated from cross-correlating the resampled LCs in each band was used to align them along the time axis. Similarly, the median maximum flux from the resampled LCs in all bands
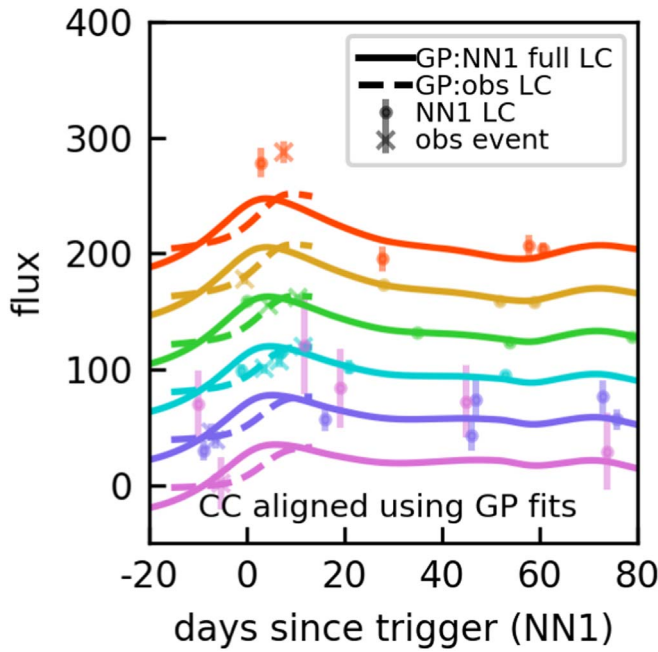
Figure A1. Similar to Figure 3 but using cross correlation to align live observed and reference LCs. Photometry shapes (crosses and dots) have the same meaning as in Figure 3. Differently from Figure 3, the GP fits shown in opaque solid and dashed lines were used to resample the signals for alignment. Cross correlation performs worse than DTW in achieving alignment.

was used to align them along the flux axis. The alignment obtained via this method is worse than from DTW (compare to the lower panel of Figure 3). It also takes six to seven times longer to compute than DTW. For partial LSST LCs, DTW yields superior performance, achieving faster and more accurate alignment without requiring LC resampling.

ORCID iDs

Dan Milisavljevic ⓘ https://orcid.org/0000-0002-0763-3885

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. 2001, in Database Theory—ICDT 2001, ed. J. van den Bussche & V. Vianu (Berlin: Springer), 420

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2015, ITPAM, 38, 252

Arcavi, I., Wolf, W. M., Howell, D. A., et al. 2016, ApJ, 819, 35

Bellm, E. 2014, in The Third Hot-wiring the Transient Universe Workshop, ed. P. R. Wozniak (Stanford, CA: Stanford Univ. Press), 27, http://www.slac.stanford.edu/econf/C131113.1/proceedings.html

Bellm, E., & Kulkarni, S. 2017, NatAs, 1, 71

Boone, K. 2019, AJ, 158, 257

Brout, D., Sako, M., Scolnic, D., et al. 2019, ApJ, 874, 106

Brown, T. M., Baliber, N., Bianco, F. B., et al. 2013, PASP, 125, 1031

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., et al. 2019, PASP, 131, 108006

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560

Djorgovski, S. G., Drake, A. J., Mahabal, A. A., et al. 2011, arXiv:1102.5004

Filippenko, A. V., Li, W. D., Treffers, R. R., & Modjaz, M. 2001, in ASP Conf. Ser. 246, IAU Coll. 183: Small Telescope Astronomy on Global Scales, ed. B. Paczynski, W.-P. Chen, & C. Lemme (San Francisco, CA: ASP), 121

Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28

Fremling, U. C., Miller, A. A., Sharma, Y., et al. 2019, arXiv:1910.12973

Gabruseva, T., Zlobin, S., & Wang, P. 2019, arXiv:1909.05032

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA: MIT Press)

Graham, M. J., Kulkarni, S. R., Bellm, E. C., et al. 2019, PASP, 131, 078001

Graur, O., Bianco, F. B., Modjaz, M., et al. 2017, ApJ, 837, 121

He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv:1512.03385

Huerta, E. A., Allen, G., Andreoni, I., et al. 2019, NatRP, 1, 600

Ishida, E. E. O., Beck, R., González-Gaitán, S., et al. 2019, MNRAS, 483, 2

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

Kasliwal, M. M., Cannella, C., Bagdasaryan, A., et al. 2019, PASP, 131, 038003

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

Kiefer, J., & Wolfowitz, J. 1959, Ann. Math. Stat., 30, 271

Kochanek, C. S., Shappee, B. J., Stanek, K. Z., et al. 2017, PASP, 129, 104502

Lunnan, R., Kasliwal, M. M., Cao, Y., et al. 2017, ApJ, 836, 60

Margutti, R., Metzger, B. D., Chornock, R., et al. 2019, ApJ, 872, 18

LSST Science Collaboration, Marshall, P., Anguita, T., et al. 2017, LSST Science Collaborations Observing Strategy White Paper: "Science-driven Optimization of the LSST Observing Strategy," v1.0, Zenodo, doi:10.5281/zenodo.842713

Mckinney, W. 2011, PoS, 445, 51

Möller, A., & de Boissière, T. 2020, MNRAS, 491, 4277

Mukund, N., Coughlin, M., Harms, J., et al. 2019, CQGra, 36, 085005

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002

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

Nordin, J., Brinnel, V., van Santen, J., et al. 2019, A&A, 631, 147

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. 2014, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR '14 (Washington, DC: IEEE Computer Society), 1717

Perley, D. A., Mazzali, P. A., Yan, L., et al. 2019, MNRAS, 484, 1031

Russakovsky, O., Deng, J., Su, H., et al. 2015, Int. J. Comput. Vis., 115, 211

Saha, A., Matheson, T., Snodgrass, R., et al. 2014, Proc. SPIE, 9149, 914908

Shannon, C. E. 1948, BSTJ, 27, 379

Smith, K. W., Williams, R. D., Young, D. R., et al. 2019, RNAAS, 3, 26

Street, R. A., Bowman, M., Saunders, E. S., & Boroson, T. 2018, Proc. SPIE, 10707, 1070711

Tavenard, R., Faouzi, J., Vandewiele, G., et al. 2017, Tslearn: A Machine Learning Toolkit Dedicated to Time-series Data, https://github.com/rtavenar/tslearn

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

Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018, PASP, 130, 064505

Tyson, J. A. 2002, Proc. SPIE, 4836, 10

Valenti, S., Sand, D. J, Yang, S., et al. 2017, ApJL, 848, L24

Vincenzi, M., Sullivan, M., Firth, R. E., et al. 2019, MNRAS, 489, 5802

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nat. Methods, 17, 261

Wang, W., Rothschild, D., Goel, S., & Gelman, A. 2015, Int. J. Forecast., 31, 980

Yang, J. J., Wang, X., Protopapas, P., & Bornn, L. 2015, arXiv:1501.02467

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. 2014, arXiv:1411.1792

Zeiler, M. D., & Fergus, R. 2013, arXiv:1311.2901