



## Active anomaly detection for time-domain discoveries

E. E. O. Ishida, M. V. Kornilov, K. L. Malanchev, M. V. Pruzhinskaya, A. A. Volnova, V. S. Korolev, F. Mondon, S. Sreejith, A. A. Malancheva, S. Das

### ► To cite this version:

E. E. O. Ishida, M. V. Kornilov, K. L. Malanchev, M. V. Pruzhinskaya, A. A. Volnova, et al.. Active anomaly detection for time-domain discoveries. *Astronomy and Astrophysics - A&A*, 2021, 650, pp.A195. 10.1051/0004-6361/202037709 . hal-03296011

**HAL Id: hal-03296011**

**<https://hal.science/hal-03296011>**

Submitted on 22 Jul 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Active anomaly detection for time-domain discoveries

E. E. O. Ishida<sup>1</sup>, M. V. Kornilov<sup>2,3</sup>, K. L. Malanchev<sup>2,4</sup>, M. V. Pruzhinskaya<sup>2</sup>, A. A. Volnova<sup>5</sup>, V. S. Korolev<sup>6,7</sup>,  
F. Mondon<sup>1</sup>, S. Sreejith<sup>1,8</sup>, A. A. Malancheva<sup>9</sup>, and S. Das<sup>10</sup>

<sup>1</sup> Université Clermont Auvergne, CNRS/IN2P3, LPC, 63000 Clermont-Ferrand, France  
e-mail: [emille.ishida@clermont.in2p3.fr](mailto:emille.ishida@clermont.in2p3.fr)

<sup>2</sup> Lomonosov Moscow State University, Sternberg Astronomical Institute, Universitetsky pr. 13, Moscow 119234, Russia

<sup>3</sup> National Research University Higher School of Economics, 21/4 Staraya Basmannaya Ulitsa, Moscow 105066, Russia

<sup>4</sup> Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 West Green Street, Urbana, IL 61801, USA

<sup>5</sup> Space Research Institute of the Russian Academy of Sciences (IKI), 84/32 Profsoyuznaya Street, Moscow 117997, Russia

<sup>6</sup> Central Aerohydrodynamic Institute, 1 Zhukovsky st, Zhukovsky, Moscow Region 140180, Russia

<sup>7</sup> Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region 141701, Russia

<sup>8</sup> Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA

<sup>9</sup> Cinimex, Bolshaya Tatarskaya street 35 bld. 3, Moscow 115184, Russia

<sup>10</sup> Washington State University, Pullman, WA 99163, USA

Received 11 February 2020 / Accepted 9 March 2021

## ABSTRACT

**Aims.** We present the first piece of evidence that adaptive learning techniques can boost the discovery of unusual objects within astronomical light curve data sets.

**Methods.** Our method follows an active learning strategy where the learning algorithm chooses objects that can potentially improve the learner if additional information about them is provided. This new information is subsequently used to update the machine learning model, allowing its accuracy to evolve with each new piece of information. For the case of anomaly detection, the algorithm aims to maximize the number of scientifically interesting anomalies presented to the expert by slightly modifying the weights of a traditional isolation forest (IF) at each iteration. In order to demonstrate the potential of such techniques, we apply the Active Anomaly Discovery algorithm to two data sets: simulated light curves from the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC) and real light curves from the Open Supernova Catalog. We compare the Active Anomaly Discovery results to those of a static IF. For both methods, we performed a detailed analysis for all objects with the ~2% highest anomaly scores.

**Results.** We show that, in the real data scenario, Active Anomaly Discovery was able to identify ~80% more true anomalies than the IF. This result is the first piece of evidence that active anomaly detection algorithms can play a central role in the search for new physics in the era of large-scale sky surveys.

**Key words.** methods: data analysis – supernovae: general – stars: variables: general

## 1. Introduction

The detection of new astronomical sources is one of the most anticipated outcomes of the next generation of large-scale sky surveys. Experiments such as the *Vera Rubin* Observatory Legacy Survey of Space and Time<sup>1</sup> (LSST) are expected to continuously monitor large areas of the sky with remarkable deliberation, which will undoubtedly lead to the detection of unforeseen astrophysical phenomena. At the same time, the volume of data gathered every night will also increase to unprecedented levels, rendering serendipitous discoveries unlikely. In the era of big data, most detected sources will never be visually inspected, and the use of automated algorithms is unavoidable.

The task of automatically identifying peculiar objects within a large set of normal instances has been highly explored in many areas of research (Aggarwal 2016). This has led to the development of a number of machine learning (ML) algorithms for anomaly detection (AD) with a large range of applications (Mehrotra et al. 2017). In astronomy, these techniques have largely been applied to areas such as the identification

of anomalous galaxy spectra (Baron & Poznanski 2017), problematic objects in photometric redshift estimation tasks (Hoyle et al. 2015), characterization of light curves (LCs) of transients (Zhang & Zou 2018; Pruzhinskaya et al. 2019), and variable stars (e.g., Rebbapragada et al. 2009; Nun et al. 2014; Giles & Walkowicz 2019; Malanchev et al. 2021), among others.

Despite encouraging results, the application of traditional AD algorithms to astronomical data scenarios is far from simple. Most of these strategies involve constructing a statistical model for the nominal data and identifying objects that significantly deviate from this model as anomalous. Once identified, these sources are subjected to further scrutiny by an expert who confirms (or not) the discovery of a new phenomenon. However, a statistical anomaly is often the result of observational defects or other spurious interference that are not scientifically interesting, leading to a high rate of candidates that turn out to be of a well-known nature despite their high anomaly scores. This incorrect identification results in a proportional fraction of resources, and research time, spent on investigating these non-peculiar objects.

Since measuring the details of a new source often requires the allocation of spectroscopic follow-up resources, the development of AD strategies able to deliver a low rate of

<sup>1</sup> <https://www.lsst.org/>

objects from scientifically well-known categories is an exceedingly important task. This task will be made more crucial in the light of the upcoming generation of telescopes, which will drastically increase the volume of nominal data and, in the process, engender a challenging AD task. In ML jargon, this would require an adaptive recommendation system that is able to optimally exploit a given ML model by carefully choosing objects that can significantly influence the results if more information about them is provided.

Active learning (AL) is a subclass of ML algorithms designed to guide such an optimal allocation of labeling resources in situations where labels are expensive and/or time consuming (Settles 2012). It has been widely applied in many real world situations and research fields, for example, natural language processing (Thompson et al. 1999), spam classification (DeBarr & Wechsler 2009), cancer detection (Liu 2004), and sentiment analysis (Kranjc et al. 2015). In the context of large-scale photometric surveys, this translates into a recommendation system for planning the distribution of follow-up resources – given a particular scientific goal. Prototypes using this underlying philosophy for supervised learning tasks were applied to the determination of stellar population parameters (Solorio et al. 2005), the supervised classification of variable stars (Richards et al. 2012), microlensing events (Xia et al. 2016), photometric redshift estimation (Vilalta et al. 2017), supernova (SN) photometric classification (Ishida et al. 2019; Kennamer et al. 2020), and the determination of galaxy morphology (Walmsley et al. 2020).

In this work, we present the first application of AL for AD in astronomical data. Similar strategies have already been reported, with encouraging results, in the identification of anomalous behavior dangerous to web services (Fan 2012), intrusion identification in cloud systems (Ibrahim & Zainal 2019), and the detection of anomalous features in building construction (Wu & Ortiz 2019) – to cite a few. Despite this successful track record, the particular characteristics of astronomical data, more specifically that of astronomical transients (errors in measurements, influence of observation conditions, sparse non-periodic and non-homogeneous time series, etc.) make this demonstration an important milestone in the exploitation of such techniques by the astronomical community. As a proof of concept, we applied the active anomaly detection (AAD) strategy proposed by Das et al. (2017) to two different data sets: simulated LCs from the Photometric LSST Astronomical Time-series Classification Challenge<sup>2</sup> (PLAsTiCC; The PLAsTiCC team 2018) and real LCs from the Open Supernova Catalog<sup>3</sup> (OSC). Our goal with the real data analysis is to lower the burden inflicted by the ML algorithms on domain experts and propose a strategy that would improve the results presented in Pruzhinskaya et al. (2019) while requiring the expert to confirm a lower number of sources. Used in combination with a traditional isolation forest (IF) algorithm, the method allows an increasingly large incidence of true positives (scientifically interesting anomalies) among the objects presented to the expert, in turn enabling a better allocation of resources with the evolution of a given survey.

This paper is organized as follows. We present the data Sect. 2 and the preprocessing analysis in Sect. 3. Section 4 describes the AAD algorithm and its implementation, and the results are presented in Sect. 5. Finally, we present our

conclusions and discuss implications for future large-scale astronomical surveys in Sect. 6.

## 2. Data

This work focuses on finding anomalies within transient LC data sets. Our experiments were performed in both simulated and real data sets.

Our real data sample comes from the OSC. This is a public repository containing SN LCs, spectra, and metadata from a range of sources. The catalog is built using data from different sources whose labels can sometimes be contradictory. This includes preliminary or “fast” classifications that need further confirmation. The catalog is constantly evolving, but at any moment in time it is also known to contain some percentage of non-SN contaminants (Guillochon et al. 2017; Pruzhinskaya et al. 2019), which makes it well suited for our purposes. The real data analysis is based on the data set<sup>4</sup> first presented in Pruzhinskaya et al. (2019). Therefore, the detailed description of quality cuts, the data selection process, and the preprocessing pipeline are given there. For clarity, we describe the main steps of the data preparation below.

From the OSC catalog, we extracted objects with LCs in *BRI* (Bessell 1990), *g'r'i'*, or *gri* filters. We assumed that *g'r'i'* filters are very similar to *gri* and that the coefficients of their transformation equations are quite small (Fukugita et al. 1996; Tucker et al. 2006; Smith et al. 2007). Light curves originally observed in *BRI* filters were converted to *gri* using Lupton’s transformation equations<sup>5</sup>.

The simulated data used in this work are a subsample of the LCs prepared for the PLAsTiCC data challenge, which was constructed to mimic the data scenario that we will encounter after 3 years of LSST observations. In order to build a data environment similar to the one we found in the OSC, we restricted our sample to six classes – SN Ia, SN II, SN Ibc, SN Ia-91bg, binary microlensing, and pair-instability SN (PISN)<sup>6</sup>. The entire PLAsTiCC test set was subjected to the LC fitting procedure described in Sect. 3.

## 3. Light curve fit

In order to obtain a homogeneous input data matrix for the ML algorithms, all LCs were submitted to a MULTIVARIATE GAUSSIAN PROCESS<sup>7</sup> pipeline. Instead of approximating the LCs in different filters independently, MULTIVARIATE GAUSSIAN PROCESS takes the correlation between different bands into account, approximating the data via a Gaussian process (GP) in all filters with one global fit. The kernel used in our implementation is composed of three radial-basis functions,

$$k_i(t_1, t_2) = \exp\left(-\frac{(t_2 - t_1)^2}{2, l_i^2}\right),$$

where  $i$  denotes the photometric band and  $l_i$  are the parameters of the GP to be found from the LC approximation. In addition, MULTIVARIATE GAUSSIAN PROCESS includes six constants,

<sup>4</sup> Data and the preprocessing pipeline for OSC are available at <http://snad.space/osc/>

<sup>5</sup> <http://www.sdss3.org/dr8/algorithms/sdssUBVRITransform.php>

<sup>6</sup> A detailed description of the astrophysical models is given in Kessler et al. (2019).

<sup>7</sup> <http://gp.snad.space>

<sup>2</sup> <https://www.kaggle.com/c/PLAsTiCC-2018>

<sup>3</sup> <https://sne.space/> (Guillochon et al. 2017).

three of which are unit variances of the basis processes and the other three describe their pairwise correlations. In total, the model has nine parameters to be fitted.

The approximation procedure was done in flux space. For each object, we only took those epochs that lie within the interval  $[-240, +240]$  days since the maximum in the  $r$ -band, averaging measurements within a 1-day time bin. Each object was characterized by 374 features. The feature set included ten parameters of MULTIVARIATE GAUSSIAN PROCESS (nine fitted parameters of the kernel and the final log-likelihood), the LC maximum flux, and normalized GP results within  $[-20, +100]$  days since maximum brightness in the  $r$ -band in steps of 1 day, concatenated according to their effective wavelength<sup>8</sup>. After applying these steps to the OSC, we visually inspected the results and eliminated bad fits, obtaining a final set of 1999 objects<sup>9</sup>.

For the PLAsTiCC data, we automatically removed all objects for which the GP fit was unsuccessful (i.e., the likelihood maximization procedure was unable to converge). A total of 7223 objects survived this preprocessing pipeline.

The two approaches described above illustrate the flexibility demanded from any feature extraction procedure aimed at preparing astronomical data to be used in standard ML environments (with the exception of a few deep learning techniques). For future large-scale surveys, such as LSST, a numerical criterion such as the one we employ for PLAsTiCC is advised since visual inspection will not be possible. Regardless of the feature extraction choice, our goal is to highlight how AAD can improve upon results from IF, given that a small fraction of scientifically interesting anomalies are present in the final data set.

## 4. Methodology

In order to compare the AAD results with those obtained with a traditional AD method and with a blind search, we performed a detailed analysis of all instances within  $\sim 2\%$  of the highest anomaly scores (145 objects for the simulated data and 40 objects for the real data). In the simulated data, this process was automatic. Once we selected the classes that represented anomalies (see Sect. 5), the algorithm was able to read the labels directly from the data file. For the OSC data, we recruited a team of two human experts (MP and AV), with extensive experience in observational and theoretical aspects of SN science, to carefully analyze each of the 40 candidates. These specialists performed a thorough investigation of each candidate – including consultation of external literature – and were not involved in the development or implementation of the AAD strategy. In what follows, we only considered the objects flagged as scientifically interesting by both experts as confirmed anomalies.

Anomaly scores were obtained according to three different strategies: random sampling (RS), IF (Sect. 4.1), and AAD (Sect. 4.2). The screening described above allowed us to coherently estimate the rate of scientifically interesting candidates for all these strategies. Each candidate was considered anomalous or nominal according to the guidelines described in Sect. 4.3.

<sup>8</sup> All the feature extraction scenarios reported in Sect. 4 of Pruzhinskaya et al. (2019) were tested. The one described here corresponds to the concatenation of the two feature extraction scenarios that presented a larger spread in anomaly scores (Pruzhinskaya et al. 2019, Fig. 8).

<sup>9</sup> The quality cuts described in Pruzhinskaya et al. (2019) aim to ensure the best behavior of the GP regression. In future surveys, such as LSST, that have a better quality and homogeneity of data, these can certainly be made less strict.

In essence, we followed a methodological strategy similar to that used in internet search engines, where the relevance of a document is judged with respect to the information the user needs, that is, the capability of solving the user's real world problem, not just the presence of the queried words (see, e.g., Manning et al. 2008). Similarly, when evaluating different algorithms, we started from a statistically identified anomaly candidate but left the final judgment to the experts – allowing the system to learn the connection between the data and the user-specific interests.

### 4.1. Isolation forest

Anomalies are identified as patterns or individual objects within a data set that do not conform to a well-defined notion of “normal” (Chandola et al. 2009). Starting from this definition, popular AD techniques begin by modeling the nominal data (defining what is normal) and subsequently identifying anomalies as samples that are unlikely to be generated by the determined model. In real data problems, this task is non-trivial since the underlying statistical distribution guiding the data generation process can be quite complex. It is possible to avoid the need for modeling the nominal data by using distance-based techniques. In this paradigm, one starts with the hypothesis that anomalous instances are likely to be far from the normal ones in the input feature space. Thus, by calculating the distance between every possible pair of objects in the data set, it is possible to select samples that, on average, are farther from the bulk of the data set. Such a strategy avoids the need for defining a complete statistical model for the normal data, but it can still be computationally very expensive for large data volumes (Taha & Hadi 2019).

The IF method is a tree-based ensemble<sup>10</sup> method first proposed by Liu et al. (2008). It was inspired by distance-based techniques and thus considers anomalies as data instances that are isolated from the bulk of the data set in a given feature space. However, this isolation is determined locally by training a randomized decision tree (Louppe 2015). In a sequence of steps, the algorithm randomly selects a subset of the data, input features, and split points (decision boundaries or nodes). The feature space is then sequentially subdivided into cells, with the number of sequential cuts determining the path length from the initially large feature space (root) to each final cell (leaf or external node). In this context, anomalies are identified as objects with the smallest path length between the root and an external node. In other words, anomalies are identified as objects that become isolated in a cell more quickly. The combination of results from a number of trees built with different subsamples makes it robust to over-fitting. By exploiting the fact that anomalies are, by definition, rare and prone to isolation, this method avoids the need for expensive distance calculations or statistical modeling of the normal instances.

### 4.2. Active anomaly detection

Active learning algorithms allow expert feedback to be incorporated into the learning model in an iterative manner and, consequently, improve the accuracy of the predicted results. As such, they work in conjunction with a traditional ML strategy, which must either be sensitive to small changes in input information (adding or subtracting a small number of objects from the training set) or allow the incorporation of such knowledge in the

<sup>10</sup> Ensemble methods are those that use a collection of learners in a synergistic manner in the formulation of the final prediction.



subsequent fine tuning of the model. Decision trees fulfill these requirements (see, e.g., Loh 2014). Moreover, for the specific case of AL for AD tasks, ensemble methods are especially significant.

Ensemble methods for AD rely upon the assumption that anomalies will have a higher anomaly score across the entire ensemble, while nominal samples will be assigned lower ones – despite values of the scores themselves being different among ensemble members. This allows us to define a weight vector,  $\mathbf{w}$ , whose elements denote the impact of different members of the ensemble in the final anomaly score. In the case of  $N$  members with perfect predictions, this will be a uniform vector,  $w_i = 1/\sqrt{N}$  for  $i \in [1, N]$ . In a more realistic scenario, certain members will be better predictors than others, and we can translate this behavior by assigning larger weights to more accurate predictors and lower ones to noisier members of the ensemble (see Fig. 1 of Das et al. 2018).

Active Anomaly Discovery<sup>11</sup> (an AAD algorithm proposed by Das et al. 2017) exploits this adaptability in order to fine tune the ensemble according to a specific definition of an anomaly, as pointed out by the expert through a series of labeled examples. The algorithm starts by training a traditional IF and then presents the candidate with the highest anomaly score to a human annotator for classification. If the expert judges the candidate to be an anomaly, the state of the model does not change and the candidate with the next highest score is presented. Whenever a given candidate is flagged as nominal, the model is updated by rescaling the contribution of each leaf node (changes in  $\mathbf{w}$ ) to the final anomaly score. This slight modification preserves the structure of the original forest while adapting the weights to ensure that labeled anomalies are assigned higher anomaly scores than labeled nominal instances.

In summary, scores from AAD have two biases: bias from the unsupervised IF, which increases scores for objects coming from isolated regions of the parameter space; and bias from previously known labels, which increases (or decreases) scores for candidates coming from sparse (or crowded) regions of the parameter space. The strength of AAD is that it is able to help discover both unknown-unknowns through the first bias and known-unknowns through the second bias. Further details about the algorithm are given in Appendix A and in Das et al. (2017, 2018).

### 4.3. Defining anomalies

The definition of an anomaly strongly depends upon the goals and objectives of the researcher. In this work, we are mainly interested in identifying non-SN contamination and/or SNe with unusual properties (Milisavljevic & Margutti 2018). Non-SN objects can be divided into cases of misclassification (quasars, binary microlensing events, novae, etc.) or completely new classes of objects. We did not consider as anomalies cases of possible misclassification that were due to signals that were too weak to allow a confident conclusion regarding the nature of the transient. These cases cannot be carefully studied due to low signal-to-noise ratios and, therefore, are not astrophysically interesting.

We consider as unusual SNe those objects that were proved to be peculiar by previous studies. These could be any kind of peculiarities: a signature of interaction with the circumstellar medium (CSM), an unusual rise or decline in the LC rate, or any other features that are not representative of the corresponding SN type.

The anomalous cases included in our simulated data were chosen to represent different classes of anomalies: SN Ia-91bg as an example of a rare type of SN (47 objects), binary microlensing events as examples of misclassifications (45 objects), and PISNe as a representative of “new physics” (184 objects). In summary, the simulated data contains ~4% (275) anomalies and ~96% (6958) nominal objects<sup>12</sup>.

For data from the OSC, we consider SLSNe and SNe of rare types as anomalous. Super-luminous SNe (SLSNe; Gal-Yam 2012) have an absolute peak magnitude of  $M < -21$  mag, which is 10–100 times brighter than standard SNe. They are sometimes divided into three broad classes: SLSNe-I without hydrogen in their spectra, hydrogen-rich SLSNe-II that often show signs of interaction with the CSM, and, finally, SLSNe-R, a rare class of hydrogen-poor events with slowly evolving LCs, which are powered by the radioactive decay of  $^{56}\text{Ni}$ . Due to their anomalous luminosity, SLSNe are becoming important probes of massive star formation in the high-redshift Universe and may be important cosmological probes, similar to Type Ia SNe (Inserra & Smartt 2014) – although only a couple dozen events have been observed so far (Moriya et al. 2018). The physics that drives this diverse class of SNe is not clearly understood, making it paramount to increase the number of observations.

As examples of SN types, we considered: Ibn (Pastorello et al. 2008), II-pec (Langer 1991), broad-lined Ic SNe associated with gamma-ray bursts (Cano et al. 2017), and low-luminosity IIP SNe (Lisakov et al. 2018). We also added 91T-like, 91bg-like Type Ia, and extreme thermonuclear SNe (e.g., Type Iax SNe; Foley et al. 2013; Taubenberger 2017) to this category. Type 1991bg SNe are characterized by a red color at maximum light and low luminosity. Type 1991T SNe, on the other hand, show a slow decline after maximum light and high-peak luminosity. The contamination due to the presence of 1991bg-like and 1991T-like SNe in cosmological samples can affect the measurements of dark energy parameters. This is extremely important for large surveys such as the LSST, which aims to constrain cosmological parameters using the bulk of normal Type Ia SNe. No non-physical effects (e.g., artifacts of interpolation) were considered as anomalies.

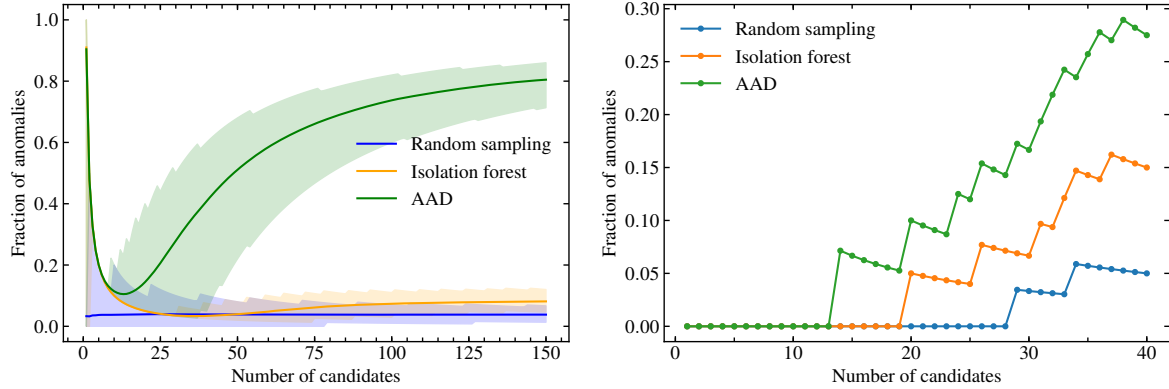
The above criteria were designed to serve as examples of the kinds of requirements one might impose to the AAD algorithm. These will certainly vary depending on the research goal, available labeling resources, and the data at hand. However, for the purposes of this work, the exact anomaly definition serves merely to illustrate the flexibility of our framework. The global behavior of exercises using different anomaly criteria should resemble those presented in Sect. 5.

## 5. Results

We first report the results from applying our method to the subset of the PLAsTiCC data described in Sect. 2. Figure 1 (left panel) shows the fraction of identified anomalies as a function of proposed candidates. This figure was created considering objects in decreasing order of anomaly scores (for IF) and following the order in which they were presented as candidates (for AAD and RS). In order to account for the random nature of the IF algorithm, we performed the experiment 2000 times using different random seeds. The plot shows the mean behavior of all these experiments as solid lines, and the shaded areas mark

<sup>12</sup> We emphasize that we cannot calculate such percentages for the OSC data since it would require our experts to perform a detailed analysis of all 1999 objects.

<sup>11</sup> [https://github.com/shubhomoydas/ad\\_examples](https://github.com/shubhomoydas/ad_examples)



**Fig. 1.** Fraction of anomalies as a function the total number of candidates scrutinized by the expert. The plot shows results obtained with the RS (blue), IF (orange), and Active Anomaly Discovery (green) algorithms. *Left:* results from the simulated PLAsTiCC data set. The solid lines represent the mean, and the shaded regions mark the 5–95 percentiles of results obtained from 2000 realizations with different random seeds. *Right:* results from the real OSC data.

the 5–95 percentiles of all results. After a total of 145 candidates were proposed ( $\sim 2\%$  of the entire data set), we confirmed that, on average, RS found four PISNe, one binary microlensing, and one SN-91bg ( $\sim 4\%$  of all available anomalies), and IF detected eight binary microlensing events and four PISNe ( $\sim 8\%$ ) among the objects with highest anomaly score. Meanwhile, the mean results from AAD after 2000 experiments flagged eight binary microlensing events and 112 PISNe ( $\sim 83\%$  of all available anomalies in the data set). Considering that in the real case the analysis of each anomaly candidate would require the use of expensive spectroscopic telescope time, these results demonstrate how AAD can be a valuable tool in the allocation of such resources.

In order to demonstrate the flexibility of the AAD algorithm to adapt to the anomaly definition set by the expert, as stated in Sect. 4, we also ran the AAD algorithm with a different anomaly definition. In the case where the expert would flag only binary microlensing events as anomalous, the AAD algorithm returned, on average, 15 true positives (in comparison with eight returned using the broader anomaly definition) – almost doubling the success rate of a very narrow search. This confirms that the method is able to adapt to the type of anomaly that is interesting to the expert and increase the fraction of candidates worthy of being investigated further.

The analysis of real data presents a much more complex scenario. In order to confirm if the AAD performance holds when dealing with real observations, we performed the same analysis on data from the OSC. Results are presented in the right panel of Fig. 1. In this scenario, 2% of the entire data set corresponded to  $\sim 40$  objects. Random sampling achieved a maximum AD rate of  $\sim 5\%$  (two objects). The IF method was able to boost this to  $\sim 15\%$  (five objects), while  $\sim 27\%$  of the objects identified by AAD were true positives (11 objects). This represents an increase of  $\sim 80\%$  in the number of true anomalies detected for the same amount of resources spent in scrutinizing candidates<sup>13</sup>. Moreover, similar to what we found in the simulated data, although both strategies require a “burn-in” period to start identifying interesting sources, AAD presented the first anomaly much earlier (14th place, in comparison to 20th place for IF). The full list of identified anomalies is provided in Table 1, and a subset of their LCs is presented in Appendix B.

<sup>13</sup> This percentage corresponds to the point where we exhausted our labeling resources (2% of the initial sample). As we can see in Fig. 1, it depends on the number of candidates analyzed by the expert.

A more detailed comparison between the IF and AAD results is displayed in Fig. 2. The diagram shows the identification of candidates presented to the expert by IF (top) and AAD (bottom). The first two objects are the same for both algorithms, with a discrepancy starting only from the third one. Candidates are ordered by their scores for IF, from left to right. For AAD, they correspond to the highest anomaly score for successive iterations of the AL loop. Anomalies confirmed by the experts are highlighted in yellow. The plot clearly illustrates not only the higher incidence of anomalies for AAD versus IF (11 vs. 6), but also the larger density among the latter candidates. The lines connecting objects that are present in both branches show that the first half of the list contains many objects in common between the two algorithms. On the other hand, the second half of the AAD list contains anomalies that are absent in the upper branch. This demonstrates that the algorithm is also able to adapt to the definition of anomaly according to the feedback received from the expert in a real data scenario. Moreover, one of the most obvious peculiar objects in our sample is a binary microlensing event, Gaia16aye. It was assigned the 33rd highest anomaly score by the IF and was the first real anomaly presented by AAD (in the 14th iteration). These results provide the first pieces of evidence that adaptive learning algorithms can be important tools in planning the optimized distribution of resources in the search for peculiar astronomical objects.

## 6. Conclusions

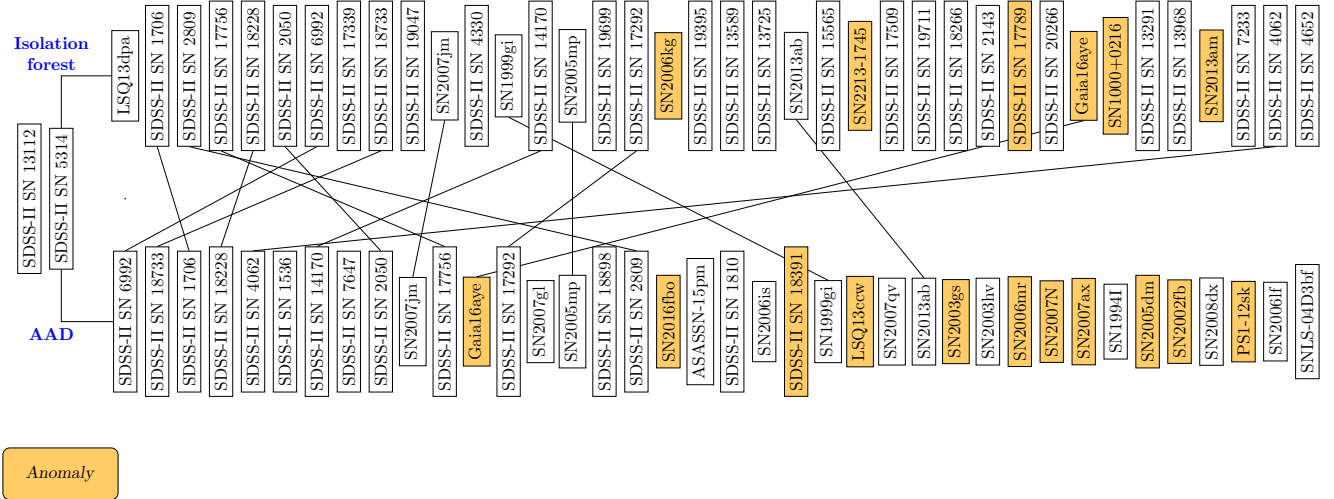
The next generation of large-scale sky surveys will certainly detect a variety of new astrophysical sources. However, since every photometrically observed candidate requires further investigation via spectroscopy, the development of automated AD algorithms with low incidences of false positives is crucial. Moreover, such algorithms must be able to detect scientifically interesting anomalies – as opposed to spurious features due to observing conditions or errors in the data processing pipeline. Active learning methods are known to perform well in such data scenarios. They represent a class of adaptive learning strategies where expert feedback is sequentially incorporated into the ML model, allowing high accuracy in prediction while keeping the distribution of analysis resources under control.

We report results supporting the use of AL algorithms in the allocation of resources for astronomical discovery. We use simulated and real LCs as benchmarks to compare the rate of true

**Table 1.** Anomalies identified by the IF and AAD algorithms.

SN name	$\alpha$	$\delta$	Type	$z$	References
Isolation forest					
SN2006kg	01:04:16.98	+00:46:08.9	AGN	0.230	Bassett et al. (2006), Östman et al. (2011)
SN2213-1745	22:13:39.97	−17:45:24.5	SLSN-R	2.046	Cooke et al. (2012)
SDSS-II SN 17789	01:29:16.13	+00:42:37.9	SLSN		Sako et al. (2018)
Gaia16aye	19:40:01.10	+30:07:53.4	ULENS, CV		Bakis et al. (2016), Wyrzykowski et al. (2016)
SN1000+0216	10:00:05.86	+02 16 23.6	SLSN-II	3.899	Cooke et al. (2012)
SN2013am	11:18:56.94	+13:03:50.0	LL IIP <sup>(a)</sup>	0.003	Nakano et al. (2013), Lisakov et al. (2018)
AAD					
Gaia16aye	19:40:01.10	+30:07:53.4	ULENS, CV		Bakis et al. (2016), Wyrzykowski et al. (2016)
SN2016fbo	01:01:35.54	+17:06:04.3	Ia <sup>(b)</sup>	0.030	Foley et al. (2018)
SDSS-II SN 18391	02:22:42.43	+00:25:05.0	?Unknown/?Star <sup>(c)</sup>		Sako et al. (2018)
LSQ13ccw	21:35:51.64	−18:32:52.0	Ibn	0.060	Pastorello et al. (2015)
SN2003gs	02:27:38.36	−01:09:35.4	Ia Pec	0.005	Krisciunas et al. (2009)
SN2006mr	03:22:42.84	−37:12:28.5	Ia-91bg	0.006	Contreras et al. (2010)
SN2007N	12:49:01.25	−09:27:10.2	Ia-91bg	0.013	Stritzinger et al. (2011), Folatelli et al. (2013)
SN2007ax	08:22:43.26	+22:33:16.9	Ia-91bg	0.007	Stritzinger et al. (2011), Folatelli et al. (2013)
SN2005dm	02:18:39.25	−06:54:10.8	Ia-91bg	0.017	Aldering et al. (2005), González-Gaitán et al. (2014)
SN2002fb	01:57:48.90	+36:20:26.3	Ia-91bg	0.016	Blondin et al. (2012)
PS1-12sk	08:44:54.86	+42:58:16.9	Ibn	0.054	Sanders et al. (2013)

**Notes.** <sup>(a)</sup>LL IIP – low-luminosity IIP SNe. <sup>(b)</sup>The LC in the OSC has a poor quality and contains incorrect photometrical points that make it looks anomalous. <sup>(c)</sup>Unknown object in Sako et al. (2018); host classified as a star by SDSS DR15.



**Fig. 2.** Comparison between the outputs of the IF and Active Anomaly Discovery algorithms when applied to the OSC data. Rectangles contain the object names of selected candidates in the order of their importance. The yellow boxes show anomalies that were visually confirmed. Solid lines indicate the objects in common for both branches.

anomalies discovered by a traditional IF algorithm to those identified by Active Anomaly Discovery (Das et al. 2017).

We show that Active Anomaly Discovery is able to increase the incidence of true anomalies in real data by 80% when compared to static IF. Moreover, the algorithm can adapt to the definition of anomaly imposed by the expert, which leads to a higher density of true positives in later iterations. This not only ensures a larger number of peculiar objects in total, but also guarantees that each new scrutinized source will, in the long run, contribute to the improvement of the learning model. In this context, not even the resources spent in analyzing false positives, in the beginning of the survey, are wasted.

In order to ensure a reliable estimation of true positive rates, we presented a controlled real data scenario in the form of a catalog containing 1999 fully observed SN LCs. This allowed visual confirmation of all the objects within the 2% highest anomaly

scores for all the algorithms. As an example of the potential AL techniques have in extracting useful information from legacy data, we highlight that the discovery of an important astrophysical contaminant (the binary microlensing event Gaia16aye) was presented to the expert much earlier following the active strategy when compared to its static counterpart (14th vs. 33rd highest anomaly score). Moreover, results from simulated data confirmed that the algorithm is flexible enough to allow the adaptation of the anomaly definition according to the interest of the expert – something that is not possible within the traditional AD paradigm. We acknowledge that important issues need to be further addressed (e.g., the variability of results for different feature extraction methods, stream mode learning, and scalability). Nevertheless, results presented here support the hypothesis that adaptive techniques can play important roles in the future of astronomy.

**Acknowledgements.** E. E. O. Ishida and S. Sreejith acknowledge support from CNRS 2017 MOMENTUM grant under project Active Learning for Large Scale Sky Surveys. M. Pruzhinskaya and M. Kornilov are supported by RFBR grant according to the research project 18-32-00426 for anomaly analysis and LCs approximation. K. Malanchev and V. Korolev are supported by RBFR grant 20-02-00779 for preparing the Open Supernova Catalog and PLAsTiCC data. A. Volnova acknowledges support from RSF grant 18-12-00522 for analysis of interpolated LCs. We used the equipment funded by the Lomonosov Moscow State University Program of Development. The authors acknowledge the support by the Interdisciplinary Scientific and Educational School of Moscow University “Fundamental and Applied Space Research”. This research has made use of NASA’s Astrophysics Data System Bibliographic Services and following PYTHON software packages: NUMPY (van der Walt et al. 2011), MATPLOTLIB (Hunter 2007), SCIPLY (Jones et al. 2001), PANDAS (McKinney 2010), and SCIKIT-LEARN (Pedregosa et al. 2011).

## References

- Aggarwal, C. 2016, *Outlier Analysis* (Springer International Publishing)
- Aldering, G., Bailey, S., Lee, B. C., et al. 2005, *ATel*, 596, 1
- Bakis, V., Burgaz, U., Butterley, T., et al. 2016, *ATel*, 9376
- Baron, D., & Poznanski, D. 2017, *MNRAS*, 465, 4530
- Bassett, B., Becker, A., Brewington, H., et al. 2006, *Cent. Bur. Electron. Telegrams*, 688, 1
- Bessell, M. S. 1990, *PASP*, 102, 1181
- Blondin, S., Matheson, T., Kirshner, R. P., et al. 2012, *AJ*, 143, 126
- Cano, Z., Wang, S.-Q., Dai, Z.-G., & Wu, X.-F. 2017, *Adv. Astron.*, 2017, 8929054
- Chandola, V., Banerjee, A., & Kumar, V. 2009, *ACM Comput. Surv.*, 41
- Contreras, C., Hamuy, M., Phillips, M. M., et al. 2010, *AJ*, 139, 519
- Cooke, J., Sullivan, M., Gal-Yam, A., et al. 2012, *Nature*, 491, 228
- Das, S., Wong, W. K., Fern, A., Dietterich, T. G., & Amran Siddiqui, M. 2017, *Workshop on Interactive Data Exploration and Analytics (IDEA’17), KDD workshop*, [arXiv:1708.09441]
- Das, S., Rakibul Islam, M., Kannappan Jayakodi, N., & Rao Doppa, J. 2018, ArXiv e-prints [arXiv:1809.06477]
- DeBarr, D., & Wechsler, H. 2009, in *Sixth Conference on Email and Anti-Spam. Mountain View, California, Citeseer*, 1
- Fan, W. K. G. 2012, in *2012 7th International Conference on Computer Science Education (ICCSE)*, 690
- Folatelli, G., Morrell, N., Phillips, M. M., et al. 2013, *ApJ*, 773, 53
- Foley, R. J., Challis, P. J., Chornock, R., et al. 2013, *ApJ*, 767, 57
- Foley, R. J., Scolnic, D., Rest, A., et al. 2018, *MNRAS*, 475, 193
- Fukugita, M., Ichikawa, T., Gunn, J. E., et al. 1996, *AJ*, 111, 1748
- Gal-Yam, A. 2012, *Science*, 337, 927
- Giles, D., & Walkowicz, L. 2019, *MNRAS*, 484, 834
- González-Gaitán, S., Hsiao, E. Y., Pignata, G., et al. 2014, *ApJ*, 795, 142
- Guillochon, J., Parrent, J., Kelley, L. Z., & Margutti, R. 2017, *ApJ*, 835, 64
- Hoyle, B., Rau, M. M., Paech, K., et al. 2015, *MNRAS*, 452, 4183
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, 9, 90
- Ibrahim, N. M., & Zainal, A. 2019, *Int. J. Swarm Intell. Res.*, 10, 53
- Insera, C., & Smartt, S. J. 2014, *ApJ*, 796, 87
- Ishida, E. E. O., Beck, R., González-Gaitán, S., et al. 2019, *MNRAS*, 483, 2
- Jones, E., Oliphant, T., Peterson, P., et al. 2001, *SciPy: Open source scientific tools for Python*
- Kennamer, N., Ishida, E. E. O., Gonzalez-Gaitan, S., et al. 2020, *ApJ*, 902, 74
- Kessler, R., Narayan, G., Avelino, A., et al. 2019, *PASP*, 131
- Kranjc, J., Smailović, J., Podpečan, V., et al. 2015, *Inf. Process. Manage.*, 51, 187
- Krisciunas, K., Marion, G. H., Suntzeff, N. B., et al. 2009, *AJ*, 138, 1584
- Langer, N. 1991, *A&A*, 243, 155
- Lisakov, S. M., Dessart, L., Hillier, D. J., Waldman, R., & Livne, E. 2018, *MNRAS*, 473, 3863
- Liu, Y. 2004, *J. Chem. Inf. Comput. Sci.*, 44, 1936
- Liu, F. T., Ting, K. M., & Zhou, Z. H. 2008, in *2008 Eighth IEEE International Conference on Data Mining (IEEE)*, 413
- Loh, W.-Y. 2014, *Int. Stat. Rev.*, 82, 329
- Louppe, G. 2015, PhD Thesis, University of Liège
- Malanchev, K. L., Pruzhinskaya, M. V., Korolev, V. S., et al. 2021, *MNRAS*, 502, 5147
- Manning, C. D., Raghavan, P., & Schütze, H. 2008, *Introduction to Information Retrieval* (Cambridge, UK: Cambridge University Press)
- McKinney, W. 2010, in *Proceedings of the 9th Python in Science Conference*, eds. S. van der Walt, & J. Millman, 56
- Mehrotra, K. G., Mohan, C. K., & Huang, H. 2017, *Anomaly Detection Principles and Algorithms*, 1st edn. (Springer Publishing Company, Incorporated)
- Milisavljevic, D., & Margutti, R. 2018, *Space Sci. Rev.*, 214, 68
- Moriya, T. J., Sorokina, E. I., & Chevalier, R. A. 2018, *Space Sci. Rev.*, 214, 59
- Nakano, S., Sugano, M., Kadota, K., et al. 2013, *Cent. Bur. Electron. Telegrams*, 3440
- Nun, I., Pichara, K., Protopapas, P., & Kim, D.-W. 2014, *ApJ*, 793, 23
- Östman, L., Nordin, J., Goobar, A., et al. 2011, *A&A*, 526, A28
- Pastorello, A., Mattila, S., Zampieri, L., et al. 2008, *MNRAS*, 389, 113
- Pastorello, A., Hadjiyska, E., Rabinowitz, D., et al. 2015, *MNRAS*, 449, 1954
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, 12, 2825
- Pruzhinskaya, M. V., Malanchev, K. L., Kornilov, M. V., et al. 2019, *MNRAS*, 489, 3591
- Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. 2009, ArXiv e-prints [arXiv:0905.3428]
- Richards, J. W., Starr, D. L., Brink, H., et al. 2012, *ApJ*, 744, 192
- Sako, M., Bassett, B., Becker, A. C., et al. 2018, *PASP*, 130
- Sanders, N. E., Soderberg, A. M., Foley, R. J., et al. 2013, *ApJ*, 769, 39
- Settles, B. 2012, *Active Learning* (Morgan & Claypool Publishers)
- Smith, J. A., Tucker, D. L., Allam, S. S., et al. 2007, in *The Future of Photometric, Spectrophotometric and Polarimetric Standardization*, ed. C. Sterken, et al., *ASP Conf. Ser.*, 364, 91
- Solorio, T., Fuentes, O., Terlevich, R., & Terlevich, E. 2005, *MNRAS*, 363, 543
- Stritzinger, M. D., Phillips, M. M., Boldt, L. N., et al. 2011, *AJ*, 142, 156
- Taha, A., & Hadi, A. S. 2019, *ACM Comput. Surv.*, 52
- Taubenberger, S. 2017, in *The Extremes of Thermonuclear Supernovae*, eds. A. W. Alsabti, & P. Murdin, 317
- The PLAsTiCC team (Allam, T. J., et al.) 2018, ArXiv e-prints [arXiv:1810.00001]
- Thompson, C. A., Califf, M. E., & Mooney, R. J. 1999, in *ICML*, Citeseer, 406
- Tucker, D. L., Kent, S., Richmond, M. W., et al. 2006, *Astron. Nachr.*, 327, 821
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *Comput. Sci. Eng.*, 13, 22
- Vilalta, R., Ishida, E. E. O., Beck, R., et al. 2017, in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*
- Walmsley, M., Smith, L., Lintott, C., et al. 2020, *MNRAS*, 491, 1554
- Wu, T., & Ortiz, J. 2019, in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys ’19* (New York, NY, USA: Association for Computing Machinery), 380
- Wyrzykowski, L., Leto, G., Altavilla, G., et al. 2016, *ATel*, 9507
- Xia, X., Protopapas, P., & Doshi-Velez, F. 2016, *Cost-Sensitive Batch Mode Active Learning: Designing Astronomical Observation by Optimizing Telescope Time and Telescope Choice*, 477
- Zhang, R., & Zou, Q. 2018, *J. Phys.: Conf. Ser.*, 1061



## Appendix A: Active anomaly detection algorithm

Below we give a brief description of how the weights are updated in each iteration of the learning loop. Further details are available in [Das et al. \(2018\)](#).

The algorithm starts by training a traditional IF ([Liu et al. 2008](#)), which requires the user to determine a contamination level,  $\tau \in [0, 1]$ , a percentile used to separate normal objects from anomalies. Once the forest is trained, we denote as  $q_\tau$  the anomaly score corresponding to the chosen contamination level. Each “leaf node” in the forest is subsequently assigned a uniform weight,  $w_i = 1/\sqrt{N_{\text{nodes}}}$ . Supposing the average number of leaf nodes per tree is  $N_{\text{avt}}$ , the dimension of the weight vector will be equal to the total number of nodes,  $\dim(\mathbf{w}) = N_{\text{trees}} \times N_{\text{avt}} = N_{\text{nodes}}$ . We also define a vector,  $\mathbf{z}$ , for each object in the data set that also has dimension  $N_{\text{nodes}}$ . Considering the entire set of leaf nodes as a spatial feature space, each element of  $\mathbf{z}$  marks the final positions occupied by a given object throughout the forest. In this context, for each object,  $\mathbf{z}$  is a sparse vector, with 0 in all elements corresponding to unoccupied leaf nodes. The anomaly score of the  $i$ th object is denoted as  $q_i = \mathbf{z}_i \cdot \mathbf{w}$ .

Given a data set  $\mathbf{H}$ , we call  $\mathbf{H}_F \subseteq \mathbf{H}$  the subset of objects that were already analyzed by the expert,  $\mathbf{H}_A \subseteq \mathbf{H}_F$  the set of labeled anomalies, and  $\mathbf{H}_N \subseteq \mathbf{H}_F$  the set of labeled normal objects. Let  $y_i \in [\text{anomaly}, \text{normal}]$  be the label given by the expert to the  $i$ th object. Our goal is to learn the weight vector,  $\mathbf{w}$ , which will allow the labeled anomalies to have a score higher than the score threshold corresponding to the user choice of  $\tau$ ,  $\mathbf{w} : q_{\mathbf{H}_A} \geq q_\tau$ .

Using a hinge loss defined as:

$$l(q, \mathbf{w}; \mathbf{z}_i, y_i) = \begin{cases} 0 & \text{if } \mathbf{w} \cdot \mathbf{z}_i \geq q \text{ and } y_i = \text{anomaly} \\ 0 & \text{if } \mathbf{w} \cdot \mathbf{z}_i < q \text{ and } y_i = \text{normal} \\ q - \mathbf{w} \cdot \mathbf{z}_i & \text{if } \mathbf{w} \cdot \mathbf{z}_i < q \text{ and } y_i = \text{anomaly} \\ \mathbf{w} \cdot \mathbf{z}_i - q & \text{if } \mathbf{w} \cdot \mathbf{z}_i \geq q \text{ and } y_i = \text{normal}, \end{cases} \quad (\text{A.1})$$

the weights for each  $t$  iteration of the AL loop can be found by solving

$$\mathbf{w}^{(t)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{1}{|\mathbf{H}_A|} \sum_{\mathbf{z}_i \in \mathbf{H}_A} l_{q_\tau^{(t-1)}}^i + \frac{1}{|\mathbf{H}_N|} \sum_{\mathbf{z}_i \in \mathbf{H}_N} l_{q_\tau^{(t-1)}}^i + \frac{1}{|\mathbf{H}_A|} \sum_{\mathbf{z}_i \in \mathbf{H}_A} l_{z_\tau^{(t-1)}}^i + \frac{1}{|\mathbf{H}_N|} \sum_{\mathbf{z}_i \in \mathbf{H}_N} l_{z_\tau^{(t-1)}}^i + \|\mathbf{w} - \mathbf{w}_p\|^2 \right\}, \quad (\text{A.2})$$

where

$$l_{q_\tau^{(t-1)}}^i \equiv l(\hat{q}_\tau(\mathbf{w}^{(t-1)}), \mathbf{w}; \mathbf{z}_i, y_i), \quad (\text{A.3})$$

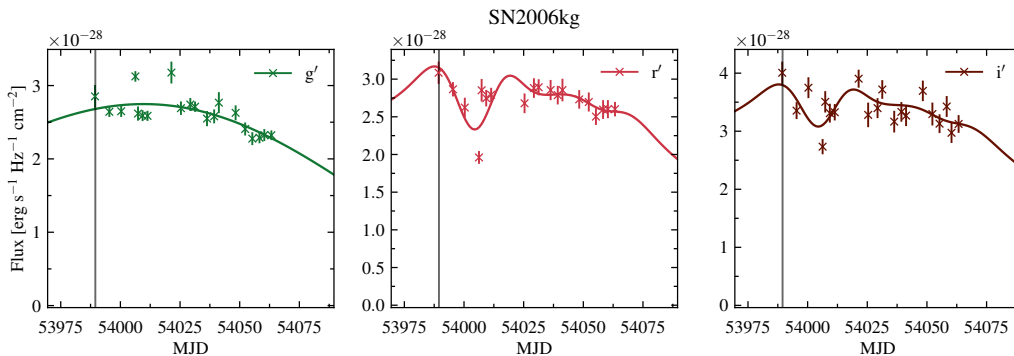
$$l_{z_\tau^{(t-1)}}^i \equiv l(\mathbf{z}_\tau^{(t-1)} \cdot \mathbf{w}, \mathbf{w}; \mathbf{z}_i, y_i), \quad (\text{A.4})$$

$$\mathbf{w}_p \equiv \left[ \frac{1}{\sqrt{N_{\text{nodes}}}}, \dots, \frac{1}{\sqrt{N_{\text{nodes}}}} \right]^T. \quad (\text{A.5})$$

Here,  $\mathbf{z}_\tau^{(t-1)}$  marks the final leaf position for the object at the quantile anomaly score threshold for iteration  $t-1$ , and  $\hat{q}_\tau(\mathbf{w}^{(t-1)})$  denotes its anomaly score<sup>14</sup>. Equation (A.2) was solved using an RMSProp algorithm, a linear loss function, and its corresponding gradient<sup>15</sup>.

## Appendix B: Visualization of selected anomalies

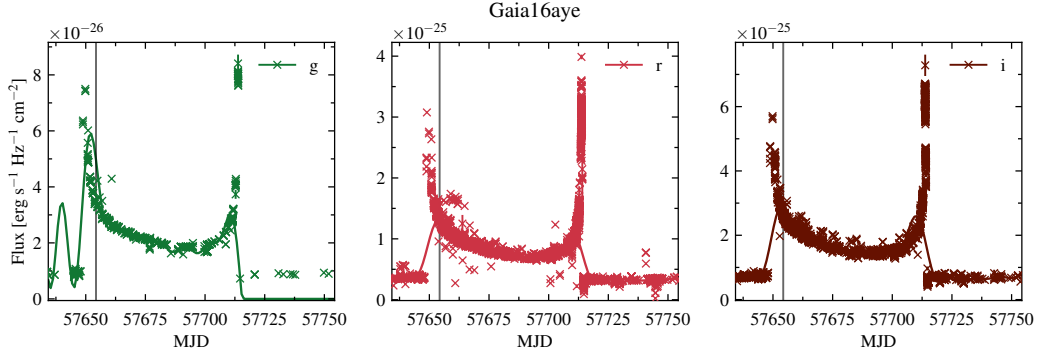
For illustrative purposes, here we show the LCs of five identified anomalies that are potentially interesting for the observer (three from the OSC data and two from the PLAsTiCC data). Two of them – SN 2006kg (Fig. B.1) and Gaia16aye (Fig. B.2) – are cases of misclassification, which the OSC partly suffers. SN2213-1745 (Fig. B.3) is an example of an SLSN, the rare class of SNe that has a huge and unexplained luminosity ([Moriya et al. 2018](#)). 78063034 (Fig. B.4) belongs to the rare class of microlensing events found in the test PLAsTiCC sample. Finally, 104498 (Fig. B.5) is an example of a 91bg-like Type Ia SN found in the training set of PLAsTiCC.



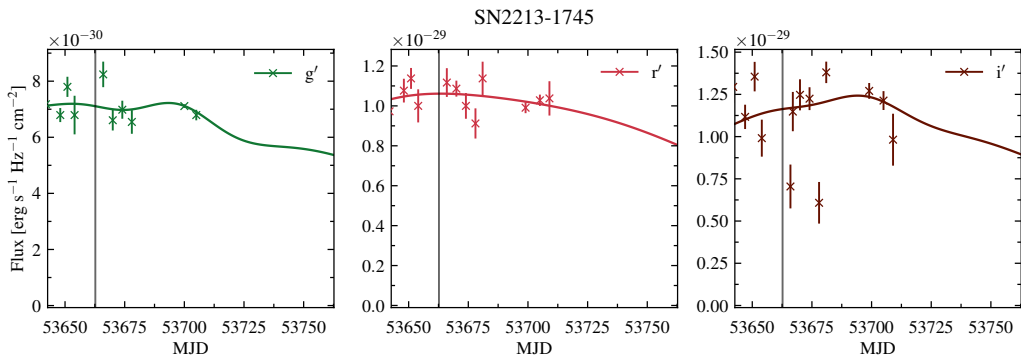
**Fig. B.1.** Light curves in  $g'r'i'$  filters of the active galactic nucleus SN2006kg ([Sako et al. 2018](#)). Solid lines are the results of our approximation using MULTIVARIATE GAUSSIAN PROCESS. The vertical line denotes the moment of maximum in the  $r'$  filter.

<sup>14</sup> By definition, both quantities were calculated with  $\mathbf{w} = \mathbf{w}^{(t-1)}$ .

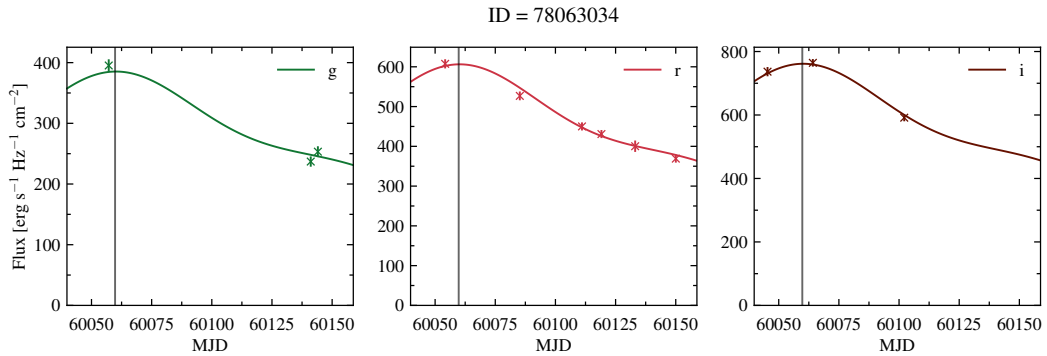
<sup>15</sup> [https://github.com/shubhomoydas/ad\\_examples/blob/master/ad\\_examples/aad/aad\\_loss.py](https://github.com/shubhomoydas/ad_examples/blob/master/ad_examples/aad/aad_loss.py)



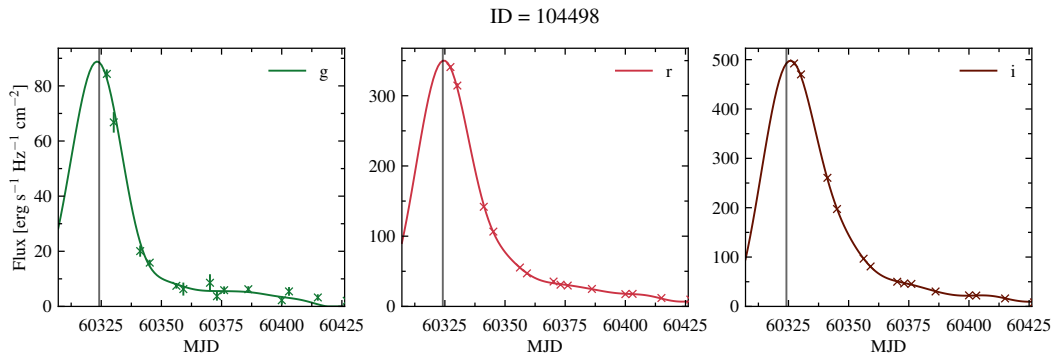
**Fig. B.2.** Light curves in *gri* filters of the binary microlensing event Gaia16aye (<http://gsaweb.ast.cam.ac.uk/alerts/alert/Gaia16aye/followup>). Solid lines are the results of our approximation using MULTIVARIATE GAUSSIAN PROCESS. The vertical line denotes the moment of maximum in the *r* filter.



**Fig. B.3.** Light curves in *g'r'i'* filters of the SLSN 2213-1745 (Cooke et al. 2012). Solid lines are the results of our approximation using MULTIVARIATE GAUSSIAN PROCESS. The vertical line denotes the moment of maximum in the *r'* filter.



**Fig. B.4.** Light curves in *gri* filters of a microlensing event (ID = 78063034) from the PLAsTiCC models. Solid lines are the results of our approximation by MULTIVARIATE GAUSSIAN PROCESS. The vertical line denotes the moment of maximum in the *r* filter.



**Fig. B.5.** Light curves in *gri* filters of the SN Ia-92bg (ID = 104498) from the PLAsTiCC models. Solid lines are the results of our approximation using MULTIVARIATE GAUSSIAN PROCESS. The vertical line denotes the moment of maximum in the *r* filter.