1 2 3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

Inferring the redshift of more than 150 GRBs with a Machine Learning Ensemble model

MARIA GIOVANNA DAINOTTI,^{1, 2, 3, 4, 5} ELIAS TAIRA,⁶ ERIC WANG,⁷ ELIAS LEHMAN,⁸
 ADITYA NARENDRA^{*},^{9, 10} AGNIESZKA POLLO,^{11, 10} GRZEGORZ M. MADEJSKI,¹² VAHE PETROSIAN,¹²
 MALGORZATA BOGDAN,^{13, 14} APRATIM DEY,¹⁵ AND SHUBHAM BHARDWAJ^{1, 2}
 ¹Division of Science, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
 ²The Graduate University for Advanced Studies (SOKENDAI), Shonankokusaimura, Hayama, Miura District, Kanagawa 240-0115
 ³Space Science Institute, 4765 Walnut St Ste B, Boulder, CO 80301, USA.
 ⁴Nevada Center for Astrophysics, University of Nevada, 4505 Maryland Parkway, Las Vegas, NV 89154, USA

⁵Bay Environmental Institute, P.O. Box 25 Moffett Field, CA, California

⁶Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48825, USA

⁷Department of Computer Science, Yale University, New Haven, CT 06511-8937

⁸Department of Physics, University of California at Berkeley, Berkeley, CA 94720, USA

⁹ Jagiellonian University, Doctoral School of Exact and Natural Sciences, Krakow, Poland

¹⁰Astronomical Observatory of Jagiellonian University, Krakow, Poland*

¹¹National Center for Nuclear Physics (NCB), Warsaw

¹²Department of Physics and SLAC National Accelerator Laboratory, Stanford University, Stanford, CA 94305, USA

¹³Department of Mathematics, University of Wroclaw, 50-384, Poland

¹⁴Department of Statistics, Lund University, SE-221 00 Lund, Sweden

¹⁵Department of Statistics, Stanford University, Stanford, CA, 94305, USA

(Dated: July 2023)

ABSTRACT

Gamma-Ray Bursts (GRBs), due to their high luminosities are detected up to redshift 10, and thus have the potential to be vital cosmological probes of early processes in the universe. Fulfilling this potential requires a large sample of GRBs with known redshifts, but due to observational limitations, only 11% have known redshifts (z). There have been numerous attempts to estimate redshifts via correlation studies, most of which have led to inaccurate predictions. To overcome this, we estimated GRB redshift via an ensemble supervised machine learning model that uses X-ray afterglows of long-duration GRBs observed by the Neil Gehrels Swift Observatory. The estimated redshifts are strongly correlated (a Pearson coefficient of 0.93) and have a root mean square error, namely the square root of the average squared error $\langle \Delta z^2 \rangle$, of 0.46 with the observed redshifts showing the reliability of this method. The addition of GRB afterglow parameters improves the predictions considerably by 63% compared to previous results in peer-reviewed literature. Finally, we use our machine learning model to infer the redshifts of 154 GRBs, which increase the known redshifts of long GRBs with plateaus by 94%, a significant milestone for enhancing GRB population studies that require large samples with redshift.

1. INTRODUCTION

Gamma-Ray Bursts (GRBs) are the most luminous events after the Big Bang. Due to their high 39 luminosities, they are detected up to redshift z = 9.4 (Cucchiara et al. 2011), and thus have the 40 potential to be vital cosmological probes of processes in the early universe. Studying GRBs enables 41 us to deepen our knowledge about the early universe and track how the universe evolves over time 42 since the GRB redshift range goes from 0.0085 (Galama et al. 1998)) to the highest redshifts observed 43 (between 8 and 9.4, (Cucchiara et al. 2011; Tanvir et al. 2008). GRBs (observed mainly in γ -rays, 44 X-rays, and sometimes in optical) are traditionally classified in short duration GRBs (SGRBs), with 45 $T_{90} < 2$ s, where T_{90} is the time interval during which the GRB emits 90% of its total observed fluence 46 (energy emitted in γ -rays) and long duration GRBs (LGRBs) where $T_{90} > 2s$. Observationally, GRBs 47 are characterized by prompt emission, the main emission observed from hard X-rays to high-energy 48 -rays and sometimes in optical (Vestrand et al. 2005; Beskin et al. 2010; Gorbovskoy et al. 2012; 49 Vestrand et al. 2014), and the afterglow emission (Costa et al. 1997; van Paradijs et al. 1997; Piro 50 et al. 1998), a long-lasting multi-wavelength emission, following the prompt, observed in X-rays, 51 optical, and sometimes radio. The afterglow sometimes contains the plateau emission feature where 52 the flux during the plateau remains constant (Nousek et al. 2006; Rowlinson et al. 2014; Zhang et al. 53 2006; Dainotti et al. 2008; Sakamoto et al. 2007; O'Brien et al. 2006; Zaninoni et al. 2013; Liang 54 et al. 2007). Plateaus are observed in 42% of X-ray afterglows (Evans et al. 2009; Li et al. 2018) and 55 in 30% of optical afterglows (Vestrand et al. 2005; Kann et al. 2006; Zeh et al. 2006; Dainotti et al. 56 2020; Panaitescu & Vestrand 2008, 2011; Oates et al. 2012). 57

Currently, the main issue in population studies is the lack of LGRB samples with known redshifts. 58 The direct determination of the redshift of a GRB requires rapid localization and spectral information. 59 One of the most powerful observatories that enable rapid detection and follow-up in multiwavelengths 60 is the Neil Gehrels Swift observatory (hereafter Swift) (Gehrels et al. 2004). Swift uses an X-ray 61 instrument for localization and can obtain spectra and sometimes redshift with the onboard UVOT 62 instrument. The Swift satellite consists of three main instruments: the Burst Alert Telescope (BAT) 63 (Burrows et al. 2005), the X-ray Telescope (XRT) (Barthelmy et al. 2005), and the Ultraviolet/Optical 64 Telescope (UVOT) (Roming et al. 2005). These instruments work together to detect, localize, and 65 collect data on GRBs and their afterglows across various wavelengths, including γ -rays, X-rays, and 66 ultraviolet/optical. 67

Swift, with its localization capabilities, has paved the way to the high-z Universe. Despite all the advantages of localization provided by Swift, still, only 26.2% (423) of Swift's 1615 GRBs have known spectroscopic redshifts up to today's date (7th of December, 2023). Redshift measurements, particularly high-z ones, are challenging due to limited telescope time and the paucity of active GRB follow-up programs.

The machine learning (ML) is applied to the data, which is influenced by the observed sensitivity of the Swift satellite, e.g. flux limit, limiting energy band, occultation by the Earth (data unavailability), etc. The data in our sample are all taken from Swift. Thus, the biases are all related to fluxes, the limiting energy band, and the presence

* E-mail: aditya.narendra@doctoral.uj.edu.pl

of lightcurve gaps in relation to data observed by the same satellite. Therefore, we 77 have a uniform bias, and since our generalization sample is also taken entirely from the 78 Swift satellite, the impact of this bias is reflected in the same way both in the training 79 set and in the generalization sample. This means that the impact of this bias in the 80 prediction is similar to the biases we have in the observed sample. If one would like 81 to apply a bias correction, for example, consider the Malmquist bias effect that allows 82 only the brightest GRBs, with larger fluxes, to be observed, then one needs to employ 83 the redshift. However, we cannot use the information of the redshift in training the 84 ML models since the redshift is the variable we would need to determine. This, indeed, 85 would induce a circularity argument. Thus, beyond the analysis of a posteriori bias 86 correction, we cannot apply any further correction to the initial data. If we had applied 87 any such treatment to the data, we would have induced a data bleeding problem in our 88 training set. So, in summary, the impact on the training sample is negligible, considering 89 that here we use a uniform data set both in the training and the generalization. 90

Thus, efforts to determine the redshift of GRBs are of paramount importance. One of the key 91 goals of increasing the sample of GRBs with known redshift is to determine an accurate measure 92 of the luminosity function (LF), which provides the number of bursts per unit luminosity, key to 93 understanding the properties of GRB luminosities as a population, the energy release and emission 94 mechanism of GRBs. Another relevant goal is the determination of the cosmic GRB formation rate 95 (GRBFR), which provides the number of events per comoving volume and time, pivotal to under-96 standing the production of GRBs at various stages of the universe. As highlighted in Petrosian et al. 97 (2015), there exists a discrepancy between LGRBs compared to the overall rates of star formation 98 within the lower redshift range (0 < z < 1). This result has been found by several groups with 99 differences, and thus, the debate is still open. Obtaining more redshifts becomes crucial for settling 100 such a debate. 101

Another great advantage of having more GRBs with redshift is the possibility of using GRBs as 102 standardized candles with empirical relations between distance-dependent and intrinsic properties of 103 GRBs. Amongst the earliest of these efforts, is the Dainotti Relation (Dainotti et al. 2008, 2011, 104 2015; Dainotti et al. 2017), a roughly inversely proportional relationship between the rest-frame time 105 at the end of the plateau phase $(T_a/(1+z))$ and its corresponding luminosity (L_a) . Later Dainotti 106 et al. (2013) showed that via the use of the Efron and Petrosian method (Efron & Petrosian (1992)), 107 this relation is intrinsic and not due to selection biases. This relation has also been discovered in 108 the optical and radio emission (Dainotti et al. 2020; Levine et al. 2022). It has also been extended 109 in three dimensions in X-rays, where the peak prompt luminosity L_{peak} has been added to the two-110 dimensional Dainotti relation (Dainotti et al. 2016, 2017; Dainotti et al. 2020). In addition, GRBs 111 observed by Fermi-LAT and detailed in the Second Fermi GRB Catalog (Ajello et al. 2019), which 112 show the existence of the plateau in γ -rays, obey this correlation as well (Dainotti et al. 2021). 113 Continuing on the extension of this relation in other wavelengths, this three-dimensional relation has 114 been found in optical too (Dainotti et al. 2022a). 115

Both the two and three-dimensional relations have been used as a valuable cosmological tool (Cardone et al. 2009, 2010; Dainotti et al. 2013; Postnikov et al. 2014; Cao et al. 2021, 2022a,b; Dainotti et al. 2022b; Dainotti et al. 2023; Bargiacchi et al. 2023; Dainotti et al. 2023). Dainotti et al. (2022b), showed how the Dainotti relation used in combination with SNe Ia is able to obtain consistent results on matter density, Ω_M to SNe Ia in the Λ CDM model with the added benefit of extending the distance ladder up to z = 5, a redshift far greater than the farthest observed SNe Ia observed up until z = 2.26 (Rodney & Riess 2015). Another large part of this paper discussed the prediction of the number of GRB observations that we would need to obtain the same precision as SNe Ia on the matter density, Ω_M in the Λ CDM model as in Conley et al. (2011); Betoule et al. (2014); Scolnic et al. (2018).

Indeed, it has been discussed in Dainotti et al. (2022c) that we need 789 GRBs to reach the same precision of SNe Ia in Conley et al. (2011). Thus, we need to add 567 more GRBs to our current sample of 222 GRBs with X-ray plateaus and known redshift.

To increase the number of GRBs with redshifts, there have been several attempts in this direction 129 by finding correlations between distance-independent quantities (peak flux, duration of afterglow 130 plateau, as the already mentioned Dainotti et al. (2008); Dainotti et al. (2013), etc) and distance-131 dependent GRB properties (prompt emission peak luminosity) to find pseudo-redshifts for the GRBs 132 with unknown redshift (Atteia (2003), Yonetoku et al. (2004), Dainotti et al. (2011)). The results of 133 these analyses are all reliant on the luminosity distance (D_L) , a quantity that, by definition, not only 134 depends on cosmology but small variations of the D_L at high redshift are subject to large variations 135 of the redshift. Thus, these results are inherently subject to inaccuracy. To avoid the issues caused 136 by including D_L in determining pseudo-redshifts, we undertake a new approach that relies on the use 137 of ML algorithms to create our redshift predictions. 138

The paper is structured as follows: in Sec. 1 we detail the problem of paucity of the redshifts, in Sec. 2 we describe the dataset. In Sec. 3, we describe our pipeline from how we process our data to select the variables to be used to how we build and test our models on our data. In Sec. 4 we discuss the performance of our model as well the predictions on the generalization set and we compare those results to the distribution of the existing set of redshifts. Finally in Sec. 5 we summarize and discuss the implications of these results.

2. THE DATA SAMPLE

In this study, we focus on GRBs observed in γ -rays and X-rays detected by the BAT and the XRT 146 telescopes onboard the Swift. We used the data stored in the NASA Swift GRB Search Tool, and the 147 Third Swift-BAT GRB Catalogue (Lien et al. 2016). Our initial step involves preprocessing the raw 148 GRB data to ensure its quality and suitability for further analysis. Considering the distinct nature of 149 various GRB classes, such as LGRBs and SGRBs, which can originate from different progenitors or 150 the same progenitors in diverse environments, it is crucial to avoid blending the characteristics of these 151 diverse classes. For this reason, our study focuses only on LGRBs. Thus, we exclude from our sample, 152 taken from Dainotti et al. (2020) and Srinivasaragavan et al. (2020), SGRBs, SGRBs with extended 153 emission (Norris & Bonnell 2006) and the intrinsically SGRBs (IS) which have $T_{90}/(1+z) < 2s$. The 154 initially available features are the following: 155

- 156 1. T_{90} the time interval during which the GRB emits 90% of its total observed fluence (energy 157 emitted in γ -rays).
- 158 2. F_a the flux at the end of the plateau emission.
- T_{a} the time at the end of the plateau emission.
- 4. α the temporal power-law index after the end of the plateau emission.

- ¹⁶¹ 5. β the spectral index assuming a power-law for the spectral energy distribution in the range ¹⁶² of the plateau emission.
- 6. γ the spectral index obtained as the time-averaged spectral fit from the Swift XRT Photon Counting mode data.
- ¹⁶⁵ 7. Fluence the energy fluence over T_{90} of the prompt emission in units of erg cm⁻².
- 8. PhotonIndex the prompt photon index from the BAT Telescope of the photon energy distribution modeled with a power law.
- 9. NH The column density of neutral hydrogen along the line of sight.
- 169 10. Peak Flux The prompt peak photon flux in units of with unit of (number of photons) cm⁻² 170 s^{-1} .

The dataset with which we train and test our ML models contains 197 LGRBs with all the features listed above, called the predictors, as well as our response variable, the redshift. Furthermore, there are also 221 GRBs without a measured redshift and the same features. This set is called the generalization set. The ML models are used to predict the redshift of these GRBs.

3. METHODOLOGY

Fig. 1 shows the summarized flowchart for our methodology. The individual parts are expanded upon in the following subsections.

178

175

3.1. Data Cleaning and Transformation

Due to the wide range encompassed by specific variables, namely Peak Flux, T_a , Fluence, NH, 179 T_{90} , and F_a , we transformed these variables into log base-10 with the aim of enhancing prediction 180 accuracy. The variables, α , β , γ , and PhotonIndex remain in the linear scale since the range in which 181 they vary is of the order of unity. We then proceed to clean our data such that we exclude 182 any non-physical values from our analysis or values which are unusual for the majority 183 of GRBs such as $\alpha > 3$ (2.5%), $\beta > 3$ (1%), $\gamma > 3$ (0.5%), Photon Index < 0 (0.5%) and 184 $\log(NH) < 20$ (8.6%) so that we can capture the average features of the GRBs. All these 185 values are set to NA and then imputed (see Sec. 3.2). We exclude the values of α , β , 186 and $\gamma > 3$, because these belong to the tail of their respective distributions, as can be 187 seen in Fig. 11. We also perform a similar transformation to our redshifts in which we create the 188 new response variable $\log(1+z)$. This results in a Gaussian distribution with a mean=0.48 and a 189 standard deviation=0.128 for this response variable, rather than a distribution with tails as shown 190 for the distribution of the redshift in the scatter matrix plot of Fig. 2. This new response variable 191 is chosen similarly as in previous literature (Dainotti et al. (2021), Gibson et al. (2022), Narendra 192 et al. (2022)), and it is a natural choice since it mimics the evolution of the variables. In addition, 193 z+1 is a more natural parametrization of the cosmological variable z. We show the scatter matrix 194 plot after the data cleaning and transformation, see Fig. 2. 195

197

3.2. Data Imputation: MICE

¹⁹⁸ Multivariate Imputation by Chained Equations (MICE, Van Buuren & Groothuis-Oudshoorn ¹⁹⁹ (2011)) can impute missing values for multiple variables using variables from the data set that are



Figure 1. The flowchart detailing each step of the pipeline. All yellow boxes represent indicate the number of GRBs in the training set at each step of the process. Grey boxes are pre-processing steps. Orange boxes indicate all steps involving model construction. Green boxes show all steps involving the generalization set predictions. Blue boxes show all the post-processing steps with the training set.



Figure 2. Scatter matrix plot of the data after the cleaning of the sample and variable transformation.

200 201 202

203

204

complete. MICE has the ability to create imputed values in R with a variety of different methods. Here, we use the predictive mean-matching method known as "midastouch" to create our model. We employ the "midastouch" approach, a predictive mean matching (PMM) technique introduced by Little & Rubin (2019). This method begins by populating missing values in a feature with its mean and subsequently estimating these values by training a model on the available complete data. For each prediction, a probability is assigned based on its distance from the value imputed for the missing variable. The missing entry is imputed by randomly drawing from the observed values of the respective predictor, weighted according to the probability defined previously.

This process is then repeated N times, after which the final substituted quantity for each missing value is determined by taking an average over the prediction of the value in each iteration.

In previous literature (Gibson et al. 2022), similar methodologies have been applied to Active Galactic Nuclei data from the Fermi Fourth LAT catalog with no noticeable addition to the uncertainty of the resulting data distribution. In fact, the constructed ML model strictly benefits from its application due to the increased size of the dataset. Given that GRBs also exhibit similarly nonlinear trends within their features, we expect to see similar results in our own study.

Here, we show the missing data in Fig. 3. The bottom x-axis shows the number of missing GRBs corresponding to the variable presented in the upper x-axis. The pink boxes show the missing GRB variables, while the blue boxes indicate GRBs with no missing data for given variables. We now detail the missing data points in our variables: 1 GRB has missing data in γ , 1 has missing data in PhotonIndex, 2 have missing data in β , 4 have missing data in log(Peak Flux), 5 have missing data in α , and 17 have missing data in log(NH). While log(T_{90}), log(F_a), log(T_a), and log(Fluence) have no missing data points.

3.2.1. Nested 10-fold Cross Validation and the Extensive Search

Here, we describe a nested 100-iteration 10-fold cross-validation (10fCV) procedure that will be 223 utilized in both the outlier removal and model construction stages. This procedure is called nested 224 since it requires an external layer of cross-validation. 10 fCV involves dividing our data set into 10 225 distinct subsets, each containing 10 parts. We then iteratively train the model using 9 out of the 10 226 subsets as training data and evaluate its performance on the remaining subset as a testing set. This 227 procedure is repeated for each subset, allowing each subset to serve as a testing set, while the others 228 are used for training. We average the prediction results across all the 100 iterations to obtain the 229 mean prediction. The standard deviation of this distribution is the prediction error. 230

The extensive search we perform uses the nested 100-iteration 10 fCV. The procedure preparatory 231 to the extensive search requires first building all possible formula candidates. These formulas are 232 meant for two purposes. The first purpose is generating multiple models for the robust linear model 233 (RLM) using the M-estimator procedure, as explained in Sec. 3.2.2. The second purpose is to 234 create multiple models for both the Generalized Additive Model (GAM) (see in Sec. 3.3.2) and the 235 Generalized Linear Model (GLM) (see in Sec. 3.3.1), both collectively used to construct the final 236 ensemble method. We build the formulas described above with a generator function that employs 237 the first-order features, which are the observed GRB variables, and the second-order variables, which 238 are the multiplicative terms among the first-order variables. Then, the extensive search allows us to 239 find the best formula among all formulas tested based on the correlation between the predicted and 240 observed redshift and the root mean square error (RMSE), a measure of deviation from the model's 241 fit written for N data points as 242

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}},$$
(1)

where x_i is the true response value for data point *i*, and \hat{x}_i is the respective predicted value.

222



Figure 3. The missing data in our sample. The red boxes show the missing GRB data points, while the blue boxes indicate GRBs with no missing data for a given GRB variable presented in the top axis. The bottom axis shows the number of missing GRBs per variable. The left axis represents the number of observations that have missing data for a specific set of features. For example: there are 172 GRBs with no missing data, 13 GRBs with missing data in log(NH) data, 3 GRBs with missing data in α , and so on. The right axis represents the number of features that are missing for that row.

For GAM, RLM, and GLM, each has its own individual cut-off for the correlation and RMSE. This cut allows us to choose the best formulas. Following this, we use these formulas to predict the

10

redshift of the test set. Next, we identify the formula corresponding to the highest correlation, the
lowest RMSE, and the lowest median absolute deviation (MAD), which is the median of the absolute
difference between each data point and the mean of the dataset given by

250

253

 $MAD = Median(|x_i - \bar{x}|), \qquad (2)$

where x_i are the data points and \bar{x} is the mean of the dataset x. Finally, we select the best formula from these based on the highest weight in the SuperLearner (detailed in Sec 3.3.3)

3.2.2. Outlier Removal

In order to remove outliers we use a preliminary robust regression method, M-estimation, which minimizes the residuals in a given model. The application of the M-estimator enables the fit of an RLM on the imputed data. We conduct an extensive search (see Sec. 3.2.1) to find an optimal formula for the RLM that best fits the data.

To this end, we include square terms of one or multiple features of the data to capture potential non-linear relationships between our predictors and the response variable. The chosen model reads as follows:

273

 $log(1 + z) = (log(NH)^{2} + log(T_{90})^{2} + log(T_{a})^{2} + log(NH)$ $+ PhotonIndex + log(T_{90}) + log(T_{a}) + log(F_{a}))^{2}$ $+ log(PeakFlux) + log(F_{a})^{2} + PhotonIndex^{2} + log(PeakFlux)^{2}.$ (3)

M-estimator is an alternative technique to the ordinary least squares method, which fits the function 262 mentioned above to our data. The ordinary least squares method attempts to minimize the square 263 of the residuals (called the L_2 norm regression) by giving outliers of the data set a higher weight. 264 This significantly affects the results of the regression fit. In contrast, the M-estimator attempts 265 to minimize the sum of a function of residuals. The function chosen for our analysis is the Huber 266 Function (Huber 1964). RLM is used for the detection of highly influential observations. We are using 267 the implementation of RLM as described in the MASS package of R (Venables & Ripley 2002). Data 268 points with weights falling below 0.5 undergo exclusion, a crucial step taken to counterbalance the 269 influence that potentially problematic data points may exert on the model's effectiveness. Following 270 this outlier removal procedure, we eliminate 6 outliers (GRB050826, GRB051109B, GRB080916A, 271 GRB111008A, GRB151112A, and GRB160327A), reducing the size of our data set to 191 GRBs. 272

3.2.3. Feature Selection

The preprocessed data is now divided into two sets: an 80% training set for model training and a 274 20% test set for performance evaluation which is never used for the best model selection. We decide 275 to reduce the number of variables to be investigated, and thus we select the most predictive features 276 given the small data sample. To identify the most important features, we use the Least Absolute 277 Shrinkage and Selection Operator (LASSO) method exclusively on the training set (Tibshirani 1996). 278 To ensure the stability of the results, we perform LASSO for 100 iterations and obtain as a result the 279 averaged weights for each predictor. To reduce the number of features we chose only those 280 that have a non-zero LASSO weight. We extract the following features: $\log(T_{90})$, $\log(F_a)$, 281 $\log(T_a)$, PhotonIndex, $\log(NH)$, and $\log(Peak Flux)$. These features will be used in all the successive 282 steps of our analysis to find the most predictive model. 283



Figure 4. The weights assigned to the features by LASSO.

3.3. Model Construction

284

285

286

287

288

289

290

291

292

293

294

Once the data has been preprocessed, we begin constructing an ensemble of supervised ML algorithms to model the relationships between the GRB features and their corresponding redshifts. Supervised ML leverages prior knowledge of the 'training' data on which ML models will be built, and their predictions will be tested on new data (the 'test' set). Parametric models use functions with a set of parameters whose coefficients are fine-tuned to fit the training data. These models, although simpler and faster to train, are however constrained by the functions. Non-parametric models, in contrast, without assuming a predefined function, are thus more flexible and powerful than parametric models. However, they are prone to over-fitting, require large samples, and extensive running time. Semi-parametric models have a parametric and a non-parametric component, thus leveraging the advantages of both models.

We begin our construction by testing a total of 115 different regression methods on 10 iterations of 10fCV. These 115 models include the best of the following ML algorithms, which were tested individually: Random forest models, with differing numbers of trees, ranging from 10 to 500; extreme gradient boosting models using the same tree combinations. Further, 92 models from the caret package (Kuhn & Max 2008) were also tested. And finally, a single support vector machine (SVM) model was picked from a combination of 7921 models. These combinations were obtained by changing
 the various hyperparameters of the SVM model.

Out of these 115 models 25 models were selected for obtaining the highest correlation during the 302 surveys. Namely, these were GAM; GLM, Bayesian GLM, GLM Network, and Interaction GLM; 303 Extreme Gradient Boosted trees; Recursive Partitioning and Regression Trees (RPart) and Random 304 Forest as implemented by the caret package; Random Forest with Conditional Inference, R's na-305 tive RPart, RPart with Pruning, Bagged Trees, Fast Implementation of Random Forest (ranger); 306 Ridge; Stepwise Akaike Information Criterion, Interaction Stepwise Regression, Forward Stepwise 307 Regression, and Classical Stepwise Regression; Feed Forward Neural Network; Regression Towards 308 the Mean; Local Weigthed Regression; Linear Modeling; a tuned Kernel Support Vector Machine; 309 Fast Multivariate Adaptive Regression Splines; and a scalable version of Lasso. 310

Our search for the best ML models consistently exhibited a preference for linear parametric and 311 semi-parametric models. The resilience of these models leads us to believe that standardizing a tuning 312 methodology for the non-parametric models is not required for this work. As a general remark, 313 we would like to stress that, in principle, the fully non-parametric models have the advantage of 314 estimating complicated non-linear relationships between the response variable and the predictors as 315 well as high-order interactions between features. Such non-parametric methods are very powerful 316 when the data set contains many observations. However, they suffer from the so-called "curse of 317 dimensionality", which sets limits on the number of parameters one can efficiently estimate for a 318 given sample size. This "curse" becomes more severe with a greater number of features. As a result, 319 the fully non-parametric ML methods allow the use of only a limited number of features, e.g. when 320 estimating redshifts based on small GRB training sets, as in this case. 321

322

3.3.1. The Generalized Linear Model

GLM is a parametric regression technique that utilizes specialized link functions to relate the distri-323 bution of the response variable to a linear combination of the predictors. As opposed to the standard 324 linear model, GLM excels in its ability to handle various distributions, such as Gaussian, Poisson, and 325 Gamma, by selecting the appropriate link function. The model's parameters are estimated through 326 Maximum Likelihood Estimation (MLE), iteratively refined for optimal fit (Nelder & Wedderburn 327 1972). This allows us to explore different model architectures and identify the most suitable model 328 for our data. We perform the extensive search over 4158 formulae (see the left panel of Fig 5), com-329 posed of first and second-order variables. The cutoffs shown in the right panel Fig 5 include 330 only the formulae with r above the 99.5th percentile (r > 0.587) and RMSE below the 331 10th percentile (RMSE < 0.158). The formula that obtains the best correlation on the test set 332 (see Sec 3.2.1) is selected as the final formula. In our implementation of GLM, we assign the formula 333 below as the desired fitting function. This is based on our results from the extensive search (see Sec. 334 (3.2.1) with a Gaussian link function: 335

336

$$\log(1+z) = (\log(NH)^2 + \log(T_{90})^2 + \log(T_a)^2 + \log(NH) + PhotonIndex + \log(T_{90}) + \log(T_a) + \log(F_a))^2 + \log(PeakFlux) + \log(F_a)^2 + PhotonIndex^2 + \log(PeakFlux)^2.$$

$$(4)$$

Note that this formula is same as the formula obtained for the outlier removal (Eq. 3).

3.3.2. The Generalized Additive Model

In the semi-parametric GAM (Hastie & Tibshirani 1990), the redshift is related to the GRB vari-339 ables via the sum of either parametric or non-parametric functions including smooth functions or a 340 combination of both. The advantage of GAM is that it incorporates smooth functions on specified 341 features to relate nonlinear relationships between the features and the response variable. Each of the 342 smooth functions is represented using a group of basis functions, also known as B-splines, which com-343 pose a piecewise polynomial function to relate the smoothed predictor to the response, constrained 344 by a specified degree of freedom. In our implementation of GAM, however, a penalty term is applied 345 to the B-splines to penalize high complexity, eliminating the need for manually specifying the degrees 346 of freedom. 347

Here, we emphasize the significance of utilizing first-order variables, as opposed to second-order 348 variables. The use of second-order variables possesses the following disadvantages: firstly, the error 349 related to the features would be squared and further propagated onto the redshift. In addition, 350 there would be unnecessary complexity added to the model, especially considering that formulas 351 with first-order variables with similar prediction quality exist. For these reasons, by prioritizing 352 first-order variables, we enhance the precision of our predictions and subsequent error estimation. 353 Thus, our extensive search for GAM is conducted on 141 formulae (see the right panel of Fig 5), 354 composed of both smoothed and unsmoothed first-order variables. We employ cuts to include 355 only the formulae with r above the 97th percentile (r > 0.644) and RMSE below the 356 3rd percentile (RMSE < 0.138) for the final selection based on performance on the test 357 set (see Sec 3.2.1). These cutoffs are different for GLM and GAM because we are 358 conducting the two searches over different numbers of formulae: 4,158 for GLM and 359 141 for GAM. Therefore, it is appropriate to apply different quantile cutoffs. We arrive 360 at the below-mentioned formula for GAM: 361

$$\log (1+z) = s(\log (\text{NH})) + s(\log (T_{90})) + s(\log (T_a)) + \log (F_a) + \text{PhotonIndex} + \log (\text{Peak Flux}).$$

Here s() denotes the smoothing function applied to the parameters in the GAM formula described above.

3.3.3. SuperLearner

The SuperLearner is an ensemble of ML models that has the advantage of combining several ML methods into a single model and leveraging the predictive power of each singular model. It is also able to use the same model with varying configurations and assess how each of these models performs. However, to perform this assessment, the models must be constructed in such a way as to minimize the anticipated risk, which quantifies the model accuracy, by reducing the RMSE.

Outside of the algorithm, we employ a nested 10 fCV 100 times (as mentioned in Sec 3.2.1) to gauge 371 the accuracy of each individual ML model. By analyzing the ensemble's behavior across different runs, 372 we aim to identify the best-performing set of models and their corresponding weights. Subsequently, 373 SuperLearner creates an optimal weighted combination of these models, providing an ensemble based 374 on the performance on the test data. SuperLearner provides coefficients that indicate the weight (A_i) 375 or significance of each individual learner within the collective ensemble. By default, these weights 376 are non-negative and sum to 1. This approach has been demonstrated to achieve asymptotically the 377 same accuracy as the most effective prediction algorithm in the ensemble. 378

(5)

338

362



Figure 5. The plot of the cross-validation results of RMSE and the r in the $\log(z + 1)$ of GLM (right) and GAM (left) formulae. Each dot on either plot represents a formula performance within the 10fCV. The red dots represent the formulae that were above the chosen RMSE and Correlation.

We survey 115 models individually, as described in Sec. 3.3, and we employ 25. Then, we use weights assigned by SuperLearner as a discriminator among the most predictive models, and we remove models weighted < 0.25 after having performed a 100 10fCV to ensure the stability of the most performative models.

3.4. The relative importance

To assess the contribution of each predictor, we use the relative importance of features which is an 384 average of local linear approximation of prediction. For each observation, synthetic data is generated 385 by adding Gaussian noise. Subsequently, we construct an approximate change in prediction through 386 a linear model prediction, $P = X \times B$ where B are fit coefficients. The local relative importance of 387 feature i is defined by $R_i = |Bi|/P|Bi|$ for all sample points. In Fig. 6, the bars show the relative 388 influence of the variables, where each bar contains the sum of the relative influence of the first-order 389 and second-order variables. As we can see, the second most important variable is T_a , highlighting 390 the importance of adding the plateau emission among the features. Indeed, all the variables related 391 to the plateau emission survive the trimming performed by LASSO, and these variables, T_a , Peak 392 Flux, and F_a , naturally recover the Dainotti correlation discussed in the Sec. 1. 393

394

379

380

381

382

383

3.4.1. Error prediction and the Catastrophic Outlier Removal

The GRBs in our sample have error measurements for $\log(F_a)$, $\log(T_a)$, $\log(T_{90})$, $\log(\text{Peak Flux})$ and PhotonIndex. The uncertainties on $\log(NH)$ are difficult to gather since they are not present in the BAT catalog and for many GRBs such uncertainties are lacking. Thus, in the measurement errors we do not include the $\log(NH)$ uncertainties. In order to account for these observational uncertainties, we perform a Monte-Carlo Markov Chain (MCMC) approach, making the assumption



Figure 6. Plot of relative influence of each predictor including in each predictor also the second-order variables.

that our uncertainties are Gaussian, because the uncertainties on the variables are independent between measurements of the same variable, and are random.

The error bars on z_{pred} are generated by MCMC simulations with Gaussian distributions centered around the central value of the observed variables and as standard deviation their measurement uncertainties. This procedure is repeated 100 times for each GRB in the 10fCV algorithm and allows us to obtain a redshift distribution whose minimum and maximum represent the error bars on z_{pred} (see the errorbars in the Fig. 7).

We obtain the 1σ and 2σ cones in Fig. 7 between z_{pred} and z_{obs} . The 1σ and 2σ are defined as:

$$1\sigma = 10^{\sigma^*} z + (10^{\sigma^*} - 1)$$
$$2\sigma = 10^{2\sigma^*} z + (10^{2\sigma^*} - 1)$$

407 where σ^* is the standard deviation in the log(z+1) scale.

It is at this stage that we remove additional outliers deemed to be 'catastrophic'. These catastrophic outliers are defined in Jones & Singal (2020) as the GRBs which $|\Delta z| > 2\sigma$. In our case, 8 GRBs 16

are catastrophic outliers, and as shown in Fig. 7, they fall outside of the blue cone. Following
this additional outlier removal procedure, our sample of 151 GRBs is reduced to 143 GRBs. It is
important to note that we do not retrain our models following the catastrophic outlier removal.

3.4.2. Bias Correction

Following the removal of catastrophic outliers, we perform bias correction on our predictions. Bias is 414 defined as the mean of the difference between predicted and observed values of the response variable. 415 When training models on an imbalanced sample or a sample that has some level of discrimination 416 against a set of random variable outcomes, the model's predictions can have significant bias. With 417 an imbalanced sample, the model becomes more skilled at predicting the redshift range with more 418 observations. To correct for this bias, we use the Optimal Transport bias correction technique. This 419 involves sorting in ascending order the predicted and observed values and fitting a linear model 420 between them. The bias is corrected using the slope and intercept of this fit following this formula: 421

$$Y_{pred} = \beta_0 + \beta_1 Y_{SI}$$

where Y_{preds} is the corrected predictions, Y_{SL} is the SuperLearner predictions, and β_0 and β_1 are the intercept and slope of the linear fits, respectively. Solving for the fitted Y_{pred} values provides the bias-corrected redshift estimates. We apply this technique separately to four independent regions, namely, $z_{obs} < 2$, $2 < z_{obs} < 3.5$, $3.5 < z_{obs} < 5.0$, and $z_{obs} > 5.0$. This allows us to adapt the bias correction to different ranges of z_{obs} , providing more accurate results.

427

3.4.3. Predicting the redshift of the Generalization Set

Given that our model has been fully validated using the 10 fCV, we may now begin making predic-428 tions for our generalization set. However, before we use the generalization data, we must 429 first perform data cleaning. This process is similar to the data cleaning of the training 430 set. However, the main difference is that we now remove GRBs outside of the parameter 431 space of our training set, in addition to the cuts performed to ensure a representative 432 set of the parameters. This results in a total of 67 GRBs (29.8% decrease) being re-433 moved out of 221 GRBs. Thus, our generalization set consists of 154 GRBs. When 434 predicting the generalization set, it is important to make sure that the GRBs are within 435 the parameter space of the trained SuperLearner model. Otherwise, the model will ex-436 trapolate the redshift predictions for those GRBs, leading to lower confidence in their 437 accuracy. With the best model obtained with the SuperLearner, we predict the redshift of these 438 154 GRBs using the same predictors we have used in the training set. Since there is no observed 439 redshift data for this dataset, we checked that our predicted redshifts come from the same parent 440 population of the observed redshifts. We perform the Kolmogrov-Smirnov test (Karson 1968) and 441 Anderson-Darling test (Stephens 1974) to verify if the two sets of data share the same underlying 442 distribution. 443

444

4. RESULTS

In this section, we present the results of our analysis, including the performance of GAM, GLM, and the SuperLearner ensemble. We also discuss the implications of our findings and the application of these models in estimating the redshift in the generalization set.



Figure 7. The scatter plot between z_{obs} and z_{pred} . Upper panel: Predictions before removal of catastrophic outliers and bias correction. Middle panel: Predictions after the removal of catastrophic outliers. Lower panel: Predictions after the removal catastrophic outliers and application of bias correction.



Figure 8. Histograms of our chosen metrics on testing sets. Both correlation and RMSE are computed in the $\log(z + 1)$ scale. Left panel: correlation between predicted and observed redshift. Right panel: RMSE of the predicted redshifts.

4.1. Performance Metrics and Comparison

The benchmark for evaluating the quality of our results revolves around minimizing the RMSE within the SuperLearner algorithm. To further evaluate the efficacy of our results, we also utilize the following metrics in conjunction with SuperLearner RMSE: the Pearson correlation coefficient (r) between z_{obs} and z_{pred} , the normalized median absolute deviation (NMAD), and the bias defined as the mean of $z_{pred} - z_{obs}$. These metrics provide a reliable assessment of each model's ability to predict redshift values accurately.

4.2. SuperLearner Results

The Superlearner identified GAM and GLM as the best predicting models. The ensemble generated with GAM ($A_1 = 0.649$), exhibits the highest predictive capability, followed by GLM ($A_2 = 0.351$). In this context, A_1 and A_2 denote coefficients indicating the optimal model and reflecting the weighted average of multiple models.

The results obtained from SuperLearner are presented in Fig. 7. We obtain a r=0.719, 460 RMSE=0.91, bias=0.11 and NMAD=1.29 between z_{obs} and z_{pred} including all GRBs. The 461 catastrophic outlier percentage is 5.3%. Following the steps mentioned in Sec. 3.4.1 and Sec. 3.4.2, 462 the results after the catastrophic outliers are removed, as shown in the middle panel of Fig. 7. Here, 463 we obtain an improvement of the correlation which reach r=0.762, RMSE=0.83 and bias=0.15. 464 The bias-corrected results are presented in the bottom panel of Fig. 7. Here, we see an improvement 465 compared to the non-biased corrected results. The improvement is visible in all the metrics: r=0.932466 (29% increase), RMSE=0.46 (49% decrease) and NMAD=0.68 (47% decrease) between 467 z_{obs} and z_{pred} . The catastrophic outlier percentage also drops from 5.3% to 4.9%. For high-z GRBs 468 $(z_{obs} \ge 3), r=0.85, RMSE=0.5, and bias=0.17.$ 469

470

448

449

450

451

452

453

454

455

456

457

458

459

4.3. GAM and GLM Performance

We begin by evaluating the predictive performance of the GAM and GLM models individually. We identified the most promising formulas for each model based on cross-validation metrics, specifically correlation, RMSE, and NMAD. We found that the selected formulas for both GAM and GLM models demonstrated substantial correlation (0.65 and 0.66, respectively) in $\log(z + 1)$ scale and exhibited relatively low RMSE (0.99 and 1.20) and NMAD (1.36 and 1.34) values when performed on 100 10fCV. The GAM and the GLM formulae that perform the best in terms of correlation are quoted in Eq. 5 and Eq. 3, respectively.

Results of SuperLearner on the test sets during 10fCV are presented in Fig. 8, which shows the distribution of r (left panel) and the RMSE (right panel) of the 100 runs of the nested-fold 10fCV procedure. Our results show that the accuracy of the prediction is quite stable with a correlation coefficient between the z_{pred} vs. z_{obs} which peaks at 0.75 and an RMSE which peaks at 0.13 for the majority of partitions in the training sets and test sets. However, for a small number of partitions, we observe a small correlation coefficient and a higher RMSE. This is indeed natural due to the large heterogeneity of the data and the relatively small sample size.

4.4. Predicting the Generalized Data Set

The distributions of predicted redshift of the generalization set and z_{obs} of the training set are 486 presented in Fig. 10 with dashed and solid bars, respectively. The generalization set distribution 487 has been obtained using the model after the optimal transport bias correction. We have checked the 488 Anderson Darling for these two distributions, and the hypothesis that they are drawn by the same 489 parent population is rejected. We further investigated the reason why the two distributions are not 490 compatible with each other and we investigated the distribution of each variable that comes into 491 play. We observed that the distributions of two variables $(\log(T_{90}))$ and $\log(NH)$ are also not drawn 492 by the same parent population according to the Anderson Darling Test. Thus, possibly this is the 493 cause of this discrepancy. In addition, in our initial sample we have removed the IS GRBs, but since 494 the IS GRBs need to have a redshift to be defined, we do not have the possibility to highlight them. 495 Another problem is that the classification of the SGRBs with extended emission for GRBs without 496 redshift is not often reported in the literature, and thus, it is hard to classify GRBs appropriately in 497 the generalization set. In order to further investigate this issue we added the IS GRBs to 498 the sample and repeated the same procedure described above. We note that the best 499 GAM and the best GLM are the same as the ones detailed in the previous sections. The 500 number of IS GRBs is 11. This means that our training sample has been increased by 501 6% compared to the total sample of the training and the test set. This increase of the 502 training sample has not changed the formula for the prediction telling us indeed that 503 our prediction is stable at the increase of the sample at least at the level of 6%. 504

We have computed both the prediction errors and the error bars on the predictions following the 505 approach detailed in Sec. 3.4.1. The contribution of these error bars are shown in the boxplot 506 presented in Fig. 9. It is clear for some GRBs the error measurements are large, but overall we can 507 assess that our method is reliable to infer the redshift for GRBs for which the redshift is unknown. 508 We would also like to clarify that because of the size of the image, not all the names 509 of the GRBs are present. We also cut GRBs that have a prediction error greater 510 than the redshift predicted by our model. This results in GRB090518, GRB150817A, 511 and GRB190604B being removed from the dataset, leaving us with 151 GRBs of the 512 generalization set in Fig. 9. 513



Figure 9. The box plot for each GRB is shown for the generalization set. The error bars are shown for each prediction.



Figure 10. Histogram comparing the distributions of the training set z_{obs} and the redshift predictions of the generalization set.

Prior to our investigation, several groups have used linear or non-linear relations between relevant 516 GRB parameters. However, the redshift inferred with these methods has not yet led to accurate 517 measurements (Reichart 2001; Yonetoku et al. 2004; Atteia 2005). Even when the inferred redshift 518 uncertainty is small (5%), these measurements are provided only for a few cases (Guiriec et al. 2016). 519 When the redshift is inferred from the correlation between the peak in the νF_{ν} spectrum, E_{peak} , 520 and the energy emitted isotropically (Amati et al. 2002) during the prompt emission, r obtained 521 between z_{obs} and z_{pred} is 0.67 (Atteia 2003). These z_{pred} estimates, tested on 17 GRBs, are accurate 522 only by a factor of 2 (Amati 2006). When the redshift is inferred from the bi-dimensional X-ray 523 Dainotti correlation between F_a and $T_a/(1+z)$, only 28% of cases have small error bars, namely 524 $(z_{pred} - z_{obs})/z_{obs} < 1$ (Dainotti et al. 2011). All these attempts are parametric. 525

Compared to Ukwatta et al. (2016), parameters like T_{90} , Photon Index, Fluence, and NH remain 526 the same. Thus, it is the plateau variables, used here for the first time, that enhance our results. 527 Ukwatta et al. (2016) employed Random Forest to estimate redshifts using prompt parameters, 528 obtaining a correlation of 0.57 between z_{pred} and z_{obs} . Comparing our results we obtain a 38% 529 improvement in correlation for the non-bias corrected results (top panel of Fig. 7) and 530 63% improvement in correlation for the bias-corrected results (bottom panel of Fig. 7. 531 Another problem, hardly explored in the literature (besides our study), is to account for measurement 532 errors of GRB variables used to train ML models. Further, comparing our results with Racz et al. 533 (2017), who achieved a r=0.67, we see a 22% increase in our correlation in the $\log_{10}(z+1)$ scale when 534 we apply the bias correction. In addition, our methodology is more complete than this work, since 535 we use the LASSO feature selection, the M-estimator, the nested 100 10 fCV, the Superlearner and 536 the bias correction. Furthermore, we performed an extensive search by minimizing the RMSE and 537 maximizing the correlations. All these steps have not been performed in the mentioned proceeding. 538

539

5. SUMMARY, DISCUSSION AND CONCLUSION

In this paper, we have developed a methodology to predict the redshift of GRBs from the Swift catalogue using prompt, plateau and afterglow parameters derived from the BAT+XRT observations. The steps we have followed are:

- 1. Cleaning and imputing the missing variables with MICE;
- ⁵⁴⁴ 2. Removing the outliers using an M-estimator;
- Applying the LASSO method to select the most predictive feature to reduce the number of
 predictors given the small data sample;
- 4. Using the SuperLearner to perform a nested 10fCV method on 25 different models to determine
 which models would be the most successful.
- 5. We determine that our best ML models for the ensemble are GAM and GLM;
- 6. Optimizing each model by performing an extensive search aimed to maximize the Pearson Correlation coefficient and minimize RMSE on a test data;
- ⁵⁵² 7. Creating our final ensemble with Eq. 5 for GAM and Eq. 3 for the GLM.
- ⁵⁵³ 8. Performing a 4-way bias correction on our training set predictions.

9. Predicting the redshifts for the generalization set, including their prediction errors and the estimated observational uncertainties on the predicted redshifts. This is achieved using MCMC simulations based on the uncertainties on the variables (see Sec. 4.4).

In comparison to other attempts to infer GRB redshift, our ensemble achieves an increase of 63% and 38% in the correlation between predicted and observed redshift reaching r = 0.93, compared to other works in which only random forest or gradient boosting alone were used. Ukwatta et al. (2016) found r = 0.57 with random forest, while Rácz et al. (2017) found r = 0.67 both with random forest and gradient boosting. The main difference, besides our enhanced prediction and the fact that we use a more complete methodology which has been detailed in Sec. 3, is the use of the plateau properties.

In this work we highlight that the use of ML techniques for redshift prediction in GRBs offers 564 several advantages over the commonly used parametric methods which employs only the use of 565 bidimensional relations. Our individual GAM and GLM models exhibit strong predictive capabilities 566 and demonstrate how to surpass the accuracy of existing approaches. In general, our study has 567 shown that the use of parametric and semi-parametric methods brings an enhanced performance 568 compared to the fully non-parametric approaches, like random forest, which are more prone to over-569 fitting compared to our methods. In addition, the advantage for the parametric and semi-parametric 570 models are more interpretable than the non-parametric models. We have also shown that improved 571 performance are actually happening with a reduced set of variables which contain first-order terms 572 exclusively. 573

Further, our study is a proof of concept for utilizing GRBs as standard candles in cosmology with the newly estimated redshifts from the generalization set. We have indeed increased the sample size of GRBs with known redshift by adding 154 GRBs (we have 423 GRBs with redshift so far) by 50% of the total sample of GRBs with redshift. With this new number we can use the Dainotti relation to serve as reliable cosmological tool. If we consider the increase of the estimates of GRBs with X-ray plateaus, then the increase is 94%.

Looking towards the future, this is a preparatory work which will allows us to find values of Ω_M 580 with a similar precision as Conley et al. (2011). As discussed in Dainotti et al. (2022a), we would 581 require 789 GRBs with X-ray plateaus to reach such a precision if we use GRBs which possess 582 X-ray plateaus (see Table 9 in Dainotti et al. (2022b)). Since in the current analysis we obtain 583 154 GRBs with unknown redshift and X-ray plateaus, we currently have a sample of 222+154=376584 GRBs with redshifts both known and inferred. Since we have a yearly rate of 15 GRBs with X-ray 585 plateaus observed with redshift and 15 observed without redshift, our sample can be incremented from 586 August 2019 until December 2023 by an additional 124 GRBs. This means that once we analyze the 587 available GRBs with plateaus we will have more than 500 GRBs with known and inferred redshift. 588 This will leave us only to wait for 789-500=289 GRBs to be observed, which with a rate of 30 589 GRBs with X-ray plateaus per year, will be reached in roughly 10 years. However, if we apply 590 the lightcurve reconstruction analysis Dainotti et al. (2023) we will have an higher precision on the 591 plateau parameters (on average 37.5%) which will allow us to need less GRBs, namely 37.2% less of 592 the initial total sample as detailed in Dainotti et al. (2022a). Therefore, from the total sample of 593 789 we should remove 37.2% of it which is 293 GRBs. Thus, we would need 496, out of these we 594 have already 489. This leaves us to wait only half a year to then reach the precision of Conley et al. 595 (2011). Thus, with the aid of both ML and lightcurve reconstruction, we can reach a higher precision 596

22

555 556

⁵⁹⁷ in cosmology than the one obtained by Betoule et al. (2014). For this analysis, we would need 987 ⁵⁹⁸ GRBs, as detailed in Dainotti et al. (2022a), so 987-489=498 GRBs more that can be observed in ⁵⁹⁹ roughly 18 years. However, if we consider the uncertainties of the parameters of the plateaus divided ⁶⁰⁰ by half the number of GRBs needed will be almost half of this sample, so the precision of Betoule ⁶⁰¹ et al. (2014) can be reached only in 9 years.

In conclusion, this work successfully predicts the redshifts of GRBs with X-ray plateau emission and introduces a pipeline of data processing techniques for obtaining reproducible and reliable results. Through preprocessing, imputation, model selection, and performance evaluation, we predict redshifts for 154 LGRBs.

Our results are a proof-of-concept showing the potential of ML-based methods to enhance the field of astrophysics and cosmology. By expanding the dataset of GRBs with known redshifts we can successfully tackle population studies such as the more accurate estimate of the luminosity function and density rate evolution, enabling a deeper understanding of the high-*z* universe and its evolution. As further data becomes available and ML techniques continue to evolve, the accuracy and precision of redshift predictions for GRBs are likely to improve, offering exciting prospects for future cosmological research.

6. APPENDIX

⁶¹⁴ Here we present the distributions of α , β , and γ before the cuts mentioned in Sec. 3.1. ⁶¹⁵ Based on the histograms, α , β , and γ values > 3 are clearly outliers, as they belong to the ⁶¹⁶ tail of their respective distributions. Because we wish to have the parameter space of ⁶¹⁷ the training data correspond as much as possible to the overall trend of the distributions ⁶¹⁸ for all the variables at play, thus, we impute them with MICE before performing the ⁶²⁹ prediction analysis.</sup>

621

613

6.1. Results without removing outliers

Here we present the results obtained when we do not eliminate the M-estimator outliers. Namely, here the training set is larger by 2%, from 152 to 155. We applied the exact same methodology described previously. Below we present the 10fCV correlation plots for the SuperLearner. As can be observed in the Fig. 12, we see a 4% reduction in correlation, a 5.5% increase in RMSE, and a 0.73% increase in the MAD.



Figure 11. Histogram distributions for α , β and γ , showing the distribution before the outlier removal.



Figure 12. The scatter plot between z_{obs} and z_{pred} . Predictions from the SuperLearner after 100 10fCV without using M-estimation outlier removal.

Thus, these results conclusively show that removing the outliers with M-estimator does indeed improve the performance of our ML models. However, given the small decrease in correlation and the small increase in RMSE we can observe a stable prediction in terms of these metrics.

631

7. ACKNOWLEDGMENTS

We acknowledge Artem Poliszczuk for the discussion about the extensive search. We would like to acknowledge the United States Department of Energy for its sponsorship of the Summer Undergraduate Laboratory Internship program which supported this work. We acknowledge funding from Swift GI Cycle XIX, grant number 22-SWIFT22-0032. This work was also supported by the Polish National Science Centre grant UMO-2018/30/M/ST9/00757 and by Polish Ministry of Science and Higher Education grant DIR/WK/2018/12.

REFERENCES

638	Ajello, M., Arimoto, M., Axelsson, M., et al. 2019,	648	—. 2005, 2
639	ApJ, 878, 52, doi: $10.3847/1538-4357/ab1d4e$	649	doi: 10.1
640	Amati, L. 2006, Monthly Notices of the Royal	650	Bargiacch
641	Astronomical Society, 372, 233,	651	2023, M
642	doi: 10.1111/j.1365-2966.2006.10840.x	652	Barthelmy
643	Amati, L., Frontera, F., Tavani, M., et al. 2002,	652	I R of
644	Astronomy and Astrophysics, 390, 81,	053	J. IL., E
645	doi: 10.1051/0004-6361:20020722	654	Beskin, G
646	Atteia, JL. 2003, Astronomy & Astrophysics,	655	The Ast
647	407, L1, doi: 10.1051/0004-6361:20030958	656	doi: 10.1

- . 2005, Il Nuovo Cimento C, 28, 647–652,
- doi: 10.1393/ncc/i2005-10120-7
- Bargiacchi, M., Dainotti, M., & Capozziello, S.
 2023, MNRAS
- ⁵² Barthelmy, S. D., Barbier, L. M., Cummings,
- J. R., et al. 2005, Space Science Reviews
- Beskin, G., Karpov, S., Bondar, S., et al. 2010,
 The Astrophysical Journal Letters, 719, L10,
- doi: 10.1088/2041-8205/719/1/L10

- Betoule, M., Kessler, R., Guy, J., et al. 2014, 657 Astronomy & amp Astrophysics, 568, A22, 658 doi: 10.1051/0004-6361/201423413 659 Burrows, D. N., Hill, J. E., Nousek, J. A., et al. 660 2005, Space Science Reviews, 120, 165, 661 doi: 10.1007/s11214-005-5097-2 662 Cao, S., Dainotti, M., & Ratra, B. 2022a, Monthly 663 Notices of the Royal Astronomical Society, 512, 664 439, doi: 10.1093/mnras/stac517 665 —. 2022b, Monthly Notices of the Royal 666 Astronomical Society, 516, 1386, 667 doi: 10.1093/mnras/stac2170 668 Cao, S., Khadka, N., & Ratra, B. 2021, Monthly 669 Notices of the Royal Astronomical Society, 510, 670 2928, doi: 10.1093/mnras/stab3559 671 Cardone, V. F., Capozziello, S., & Dainotti, M. G. 672 2009, MNRAS, 400, 775, 673 doi: 10.1111/j.1365-2966.2009.15456.x 674 Cardone, V. F., Dainotti, M. G., Capozziello, S., 675 & Willingale, R. 2010, MNRAS, 408, 1181, 676 doi: 10.1111/j.1365-2966.2010.17197.x 677 Conley, A., Guy, J., Sullivan, M., et al. 2011, ApJS 678 Costa, E., Frontera, F., Heise, J., et al. 1997, 679 Nature, 387, 783, doi: 10.1038/42885 680 Cucchiara, A., Levan, A. J., Fox, D. B., et al. 681 2011, The Astrophysical Journal 682 Dainotti, M., Levine, D., Fraija, N., & Chandra, 683 P. 2021, Galaxies, 9, 95, 684 doi: 10.3390/galaxies9040095 685 Dainotti, M. G., Bargiacchi, G., Bogdan, M., 686 Capozziello, S., & Nagataki, S. 2023, Reduced 687 uncertainties up to 43% on the Hubble constant 688 and the matter density with the SNe Ia with a 689 new statistical analysis, arXiv, 690 doi: 10.48550/ARXIV.2303.06974 691 Dainotti, M. G., Bargiacchi, G., Łukasz Lenart, 692 A., et al. 2022a, The Astrophysical Journal, 693 931, 106, doi: 10.3847/1538-4357/ac6593 694 Dainotti, M. G., Cardone, V. F., & Capozziello, S. 695 2008, Monthly Notices of the Royal 696 Astronomical Society, 391, L79, 697 doi: 10.1111/j.1745-3933.2008.00560.x 698 Dainotti, M. G., Del Vecchio, R., Shigehiro, N., & 699 Capozziello, S. 2015, ApJ, 800, 31, 700 doi: 10.1088/0004-637X/800/1/31 701 Dainotti, M. G., Fabrizio Cardone, V., 702 Capozziello, S., Ostrowski, M., & Willingale, R. 703 2011, The Astrophysical Journal, 730, 135, 752 704
- ⁷⁰⁵ doi: 10.1088/0004-637X/730/2/135
- Dainotti, M. G., Hernandez, X., Postnikov, S., 706 et al. 2017, The Astrophysical Journal, 848, 88, 707 doi: 10.3847/1538-4357/aa8a6b 708 Dainotti, M. G., Hernandez, X., Postnikov, S., 709 et al. 2017, The Astrophysical Journal, 848, 88 710 Dainotti, M. G., Lenart, A., Chraya, A., et al. 711 2022b, APJ. 712 https://arxiv.org/pdf/2209.08675.pdf 713 Dainotti, M. G., Ostrowski, M., & Willingale, R. 714 2011, Monthly Notices of the Royal 715 Astronomical Society, 418, 2202, 716 doi: 10.1111/j.1365-2966.2011.19433.x 717 Dainotti, M. G., Petrosian, V., Singal, J., & 718 Ostrowski, M. 2013, The Astrophysical Journal, 719 774, 157, doi: 10.1088/0004-637x/774/2/157 720 Dainotti, M. G., Petrosian, V., Singal, J., & 721 Ostrowski, M. 2013, ApJ, 774, 157, 722 doi: 10.1088/0004-637X/774/2/157 723 Dainotti, M. G., Postnikov, S., Hernandez, X., & 724 Ostrowski, M. 2016, The Astrophysical Journal 725 Letters, 825, L20, 726 doi: 10.3847/2041-8205/825/2/L20 727 Dainotti, M. G., Sharma, R., Narendra, A., et al. 728 2023, ApJS, 267, 42, 729 doi: 10.3847/1538-4365/acdd07 730 Dainotti, M. G., Via Nielson, G. S., Rinaldi, E., 731 et al. 2022c, Galaxies, 514, 1828, 732 doi: 10.3390/galaxies10010024 733 Dainotti, M. G., Livermore, S., Kann, D. A., et al. 734 2020, ApJL, 905, L26, 735 doi: 10.3847/2041-8213/abcda9 736 Dainotti, M. G., Livermore, S., Kann, D. A., et al. 737 2020, The Astrophysical Journal Letters, 905, 738 L26, doi: 10.3847/2041-8213/abcda9 739 Dainotti, M. G., Omodei, N., Srinivasaragavan, 740 G. P., et al. 2021, ApJS, 255, 13, 741 doi: 10.3847/1538-4365/abfe17 742 Efron, B., & Petrosian, V. 1992, ApJ, 399, 345, 743 doi: 10.1086/171931 744 Evans, P. A., Beardmore, A. P., Page, K. L., et al. 745 2009, Monthly Notices of the Royal 746 Astronomical Society, 397, 1177, 747 doi: 10.1111/j.1365-2966.2009.14913.x 748 Galama, T. J., Vreeswijk, P. M., & Paradijs, J. V. 749 1998, Nature 750 Gehrels, N., Chincarini, G., Giommi, P., et al. 751
- ⁷⁵² 2004, The Astrophysical Journal, 611, 1005,
 ⁷⁵³ doi: 10.1086/422091

754	Gibson, S. J., Narendra, A., Dainotti, M. G., et al.	804
755	2022, Frontiers in Astronomy and Space	805
756	Sciences, 9, doi: 10.3389/fspas.2022.836215	806
757	Gorbovskoy, E. S., Lipunova, G. V., Lipunov,	807
758	V. M., et al. 2012, Monthly Notices of the Royal	808
759	Astronomical Society, 421, 1874,	809
760	doi: 10.1111/j.1365-2966.2012.20195.x	810
761	Guiriec, S., Gonzalez, M. M., Sacahui, J. R., et al.	811
762	2016, The Astrophysical Journal, 819, 79,	812
763	doi: 10.3847/0004-637X/819/1/79	012
764	Hastie, T., & Tibshirani, R. 1990, Generalized	013
765	Additive Models (Routledge),	814
766	doi: https://doi.org/10.1201/9780203753781	815
767	Huber, P. J. 1964, The Annals of Mathematical	816
768	Statistics, 35, 73,	817
769	doi: 10.1214/aoms/1177703732	818
770	Jones, E., & Singal, J. 2020, Publications of the	819
771	Astronomical Society of the Pacific, 132, 024501	820
772	Kann, D. A., Klose, S., & Zeh, A. 2006, The	821
773	Astrophysical Journal, 641, 993,	822
774	doi: 10.1086/500652	823
775	Karson, M. 1968, Journal of the American	824
776	Statistical Association, 63, 1047,	825
777	doi: 10.1080/01621459.1968.11009335	826
778	Kuhn, & Max. 2008, Journal of Statistical	827
779	Software, 28, 1–26, doi: 10.18637/jss.v028.i05	828
780	Levine, D., Dainotti, M., Zvonarek, K. J., et al.	829
781	2022, The Astrophysical Journal, 925, 15,	830
782	doi: 10.3847/1538-4357/ac4221	831
783	Li, L., Wu, XF., Lei, WH., et al. 2018, The	832
784	Astrophysical Journal Supplement, 236, 26,	833
785	doi: 10.3847/1538-4365/aabaf3	834
786	Liang, EW., Zhang, BB., & Zhang, B. 2007,	835
787	The Astrophysical Journal, 670, 565,	035
788	doi: 10.1086/521870	030
789	Lien, A., Sakamoto, T., Barthelmy, S. D., et al.	001
790	2016, The Astrophysical Journal, 829, 7,	838
791	doi: 10.3847/0004-637x/829/1/7	839
792	Little, R. J., & Rubin, D. B. 2019, Statistical	840
793	analysis with missing data, Vol. 793 (John	841
794	Wiley & Sons)	842
795	Narendra, A., Gibson, S. J., Dainotti, M. G., et al.	843
796	2022, The Astrophysical Journal Supplement	844
797	Series, 259, 55, doi: $10.3847/1538-4365/ac545a$	845
798	Nelder, J., & Wedderburn, R. 1972, Journal of the	846
799	Royal Statistical Society. Series A, 135, 370,	847
800	doi: https://doi.org/10.2307/2344614	848
801	Norris, J. P., & Bonnell, J. T. 2006, The	849
802	Astrophysical Journal, 643, 266,	850
803	doi: 10.1086/502796	851

	Nousek, J. A., Kouveliotou, C., Grupe, D., et al.
	2006, The Astrophysical Journal, 642, 389,
;	doi: 10.1086/500724
,	Oates, S. R., Page, M. J., De Pasquale, M., et al.
;	2012, Monthly Notices of the Royal

2012, Monthly Notices of the Royal Astronomical Society, 426, L86, doi: 10.1111/j.1745-3933.2012.01331.x

O'Brien, P. T., Willingale, R., Osborne, J., et al. 2006, The Astrophysical Journal, 647, 1213, doi: 10.1086/505457

Panaitescu, A., & Vestrand, W. T. 2008, Monthly
Notices of the Royal Astronomical Society, 387, 497, doi: 10.1111/j.1365-2966.2008.13231.x

 —. 2011, Monthly Notices of the Royal Astronomical Society, 414, 3537, doi: 10.1111/j.1365-2966.2011.18653.x

 Petrosian, V., Kitanidis, E., & Kocevski, D. 2015, The Astrophysical Journal, 806, 44, doi: 10.1088/0004-637x/806/1/44

Piro, L., Amati, L., Antonelli, L. A., et al. 1998,
Astronomy and Astrophysics, 331, L41.
https://arxiv.org/abs/astro-ph/9710355

Postnikov, S., Dainotti, M. G., Hernandez, X., &
 Capozziello, S. 2014, ApJ, 783, 126,
 doi: 10.1088/0004-637X/783/2/126

 Racz, I. I., Ribli, D., Bagoly, Z., et al. 2017, in
 Proceedings of the 7th International Fermi Symposium, 79

Rácz, I. I., Ribli, D., Bagoly, Z., et al. 2017, in
Proceedings of 7th International Fermi
Symposium — PoS(IFS2017) (Sissa Medialab),
doi: 10.22323/1.312.0079

Reichart, D. E. 2001, The Astrophysical Journal, 554, 643, doi: 10.1086/321428

⁸ Rodney, S. A., & Riess, A. G. 2015, AJ

Roming, P. W. A., Kennedy, T. E., Mason, K. O.,
 et al. 2005, Space Science Reviews, 120, 95,
 doi: 10.1007/s11214-005-5095-4

Rowlinson, A., Gompertz, B. P., Dainotti, M.,
et al. 2014, Monthly Notices of the Royal
Astronomical Society, 443, 1779,
doi: 10.1093/mnras/stu1277

Sakamoto, T., Hill, J. E., Yamazaki, R., et al. 2007, The Astrophysical Journal, 669, 1115, doi: 10.1086/521640

Scolnic, D. M., Jones, D. O., Rest, A., et al. 2018,
 The Astrophysical Journal, 859, 101,
 doi: 10.3847/1538-4357/aab9bb

- 852 Srinivasaragavan, G. P., Dainotti, M. G., Fraija,
- N., et al. 2020, The Astrophysical Journal, 903,
 18, doi: 10.3847/1538-4357/abb702
- Stephens, M. A. 1974, Journal of the American
 Statistical Association, 69, 730,
- doi: 10.1080/01621459.1974.10480196
- Tanvir, N. R., Fox, D. B., & Levan, A. J. 2008,
 Nature
- Tibshirani, R. 1996, Journal of the Royal
 Statistical Society, 58, 267
- ⁸⁶² Ukwatta, T. N., Woźniak, P. R., & Gehrels, N.
- 2016, Monthly Notices of the Royal
 Astronomical Society, 458, 3821,
- doi: 10.1093/mnras/stw559
- ⁸⁶⁶ Ukwatta, T. N., Woźniak, P. R., & Gehrels, N.
- ⁸⁶⁷ 2016, Monthly Notices of the Royal
- Astronomical Society, 458, 3821,
- doi: 10.1093/mnras/stw559
- Van Buuren, S., & Groothuis-Oudshoorn, K. 2011,
- Journal of statistical software, 45, 1
- van Paradijs, J., Groot, P. J., Galama, T., et al.
- ⁸⁷³ 1997, Nature, 386, 686, doi: 10.1038/386686a0

- Venables, W. N., & Ripley, B. D. 2002, in Modern
 applied statistics with S (Springer), 271–300
- Vestrand, W. T., Wozniak, P. R., Wren, J. A.,
 et al. 2005, Nature, 435, 178,
 doi: 10.1038/nature03515
- Vestrand, W. T., Wren, J. A., Panaitescu, A.,
 et al. 2014, Science, 343, 38,
 doi: 10.1126/science.1242316
- Yonetoku, D., Murakami, T., Nakamura, T., et al.
 2004, The Astrophysical Journal, 609, 935,
 doi: 10.1086/421285
- Zaninoni, E., Bernardini, M. G., Margutti, R.,
 Oates, S., & Chincarini, G. 2013, Astronomy
 and Astrophysics, 557, A12,
 doi: 10.1051/0004-6361/201321221
- Zeh, A., Klose, S., & Kann, D. A. 2006, The
 Astrophysical Journal, 637, 889,
 doi: 10.1086/498442
- ⁸⁹² Zhang, B., Fan, Y. Z., Dyks, J., et al. 2006, The
 ⁸⁹³ Astrophysical Journal, 642, 354,
- doi: 10.1086/500723