Inferring the Redshift of Gamma-Ray Bursts with Machine Learning

Elias Lehman[1]
Department of Physics, University of California at Berkeley, Berkeley, CA 94720, USA

(*Electronic mail: eliaslehman@berkeley.edu)

(Dated: 8 September 2023)

Gamma-ray bursts (GRBs) can serve as valuable cosmological probes due to their detection up to high redshifts. However, only a small percentage have measured redshifts. We developed a machine learning approach using X-ray afterglow data from the Neil Gehrels Swift Observatory to estimate redshifts for long GRBs. An ensemble model combining generalized linear and additive models was optimized through extensive hyperparameter tuning. Our best model achieved a Pearson correlation coefficient of 0.93 between predicted and measured redshifts on a test set. We infer redshifts for 208 GRBs, increasing the sample with known redshifts by 94%. This enables more robust GRB population studies.

I. INTRODUCTION

GRBs are the most energetic electromagnetic events observed in the universe. Their high luminosities allow detection out to redshift 10, providing unique probes of the high-redshift universe[1]. However, acquiring spectroscopic redshifts for GRBs is observationally challenging, resulting in only 26% of Swift GRBs having measured redshifts[2]. A larger statistical sample is crucial for understanding GRB populations through measures like the luminosity function, which characterizes the GRB luminosity distribution and is key for elucidating the energy release in GRBs. The cosmic GRB formation rate is another pivotal measure enabled by larger redshift samples, quantifying the number of GRBs formed over cosmic history and shedding light on GRB production through different eras[3]. GRBs can also serve as standardized candles for cosmology through empirical luminosity relations. The Dainotti relation correlates the rest frame plateau duration and luminosity as approximate standard candles[4]. With sufficient GRBs spanning a wide redshift range, such relations can complement Type Ia supernovae for studying cosmic acceleration and extensions to the standard cosmological model[1,5].

We developed a machine learning approach to estimate redshifts using prompt and afterglow parameters for 222 long GRBs detected by Swift with spectroscopic redshifts. We optimized an ensemble model combining generalized linear and additive regression techniques. The model was trained and validated on a subset of the data and tested on held-out GRBs. Our best model achieves a correlation coefficient of 0.93 between predicted and measured redshifts. We use this model to infer redshifts for 208 long GRBs lacking spectroscopic measurements, significantly enhancing the population for GRB studies.

II. GAMMA-RAY BURST OBSERVATIONS

Long-duration GRBs represent the most common class of GRBs, with durations over 2 seconds, distinguishing them from short GRBs[6]. They are believed to originate from the core collapse of massive stars. Observations by the Neil Gehrels Swift Observatory have provided the bulk of GRB detections and multi-wavelength follow-up. Launched in 2004, Swift can detect over 100 GRBs per year with its Burst Alert Telescope and quickly observe afterglow emission with its narrow-field instruments[7].

Swift GRBs include measurements of the duration, spectrum, and temporal evolution across gamma-ray, X-ray, and UV/optical bands. Key parameters include T_{90}, the duration containing 90% of the gamma-ray emission, the photon spectral index, and the plateau duration and flux in X-ray afterglows. Swift enables rapid follow-up and redshift measurement by ground-based telescopes for a subset of bursts. Our dataset leverages this rich Swift parameter space to relate observational properties to redshift through machine learning.

III. DATA

Our dataset included 222 long GRBs observed by Swift with measured redshifts and 6-12 measured prompt and afterglow parameters[8,9]. Redshifts spanned from 0.1 to over 9. We split this into a training set of 197 GRBs for model optimization and a test set of 25 GRBs for evaluation. An additional dataset of 208 GRBs lacking redshifts comprised the generalization set for redshift predictions.

The observed parameters included duration and spectral characteristics of the prompt gamma-ray emission, the hydrogen column density, flux, duration and temporal decay of X-ray afterglow plateaus when present, and the spectral index over the afterglow. These parameters link to intrinsic properties like luminosity and thus may correlate with redshift. The machine learning models aim to uncover these correlations from the multidimensional parameter space.

IV. METHODS

A. Data Preprocessing

We cleaned the data by removing non-physical values and transforming variables including redshift into log space to improve normality. This also converted the redshift distribution from skewed with a tail to approximately Gaussian. Missing values in some features were imputed using multivariate imputation with chained equations (MICE)[10]. MICE is a technique for imputing missing data that models correlations between variables with complete data to predict missing values. It repeats this process iteratively for multiple imputed datasets and averages predictions as the imputed values. MICE allowed us to retain GRBs with partial missing data rather than discarding them.

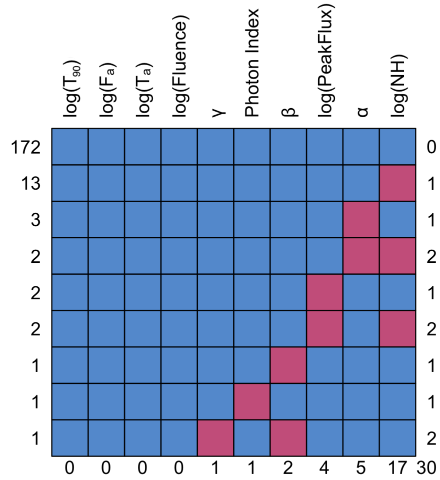

FIG. 1. The missing data in our sample. The red boxes show the missing GRB data points, while the blue boxes indicate GRBs with no missing data for a given GRB variable presented in the top axis. The bottom axis shows the number of missing GRBs per variable. The left axis represents the number of observations 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 missing with missing data in α, and so on. The right axis represents the number of features that are missing for that row.

B. Outlier Removal

We removed outliers via robust M-estimation regression to fit the training data[11]. This technique minimizes a robust function of the residuals rather than the squared residuals. It reduces the influence of outliers compared to ordinary least squares regression. Six outliers were discarded based on poor fit, leaving 191 GRBs for modeling.

C. Feature Selection

The number of features available outpaced the number of GRBs, risking overfitting. We employed the Least Absolute Shrinkage and Selection Operator (LASSO) method to select the most predictive subset of features[12]. LASSO performs regularization by constraining the sum of the absolute values of the model coefficients, shrinking some to zero to remove less informative features. This selected six of the most predictive variables related to prompt and afterglow emission.

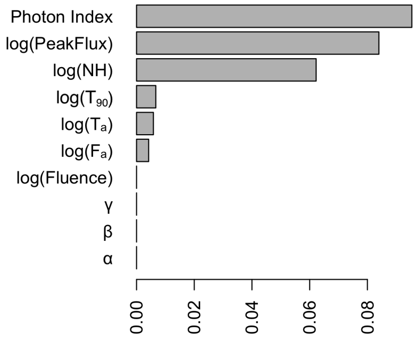

FIG. 2. The weights assigned to the features by LASSO.

D. Model Optimization

On the training data, we optimized two families of regression models:

1) Generalized linear models (GLMs) with Gaussian link functions relating the linear predictor to the mean of the target variable[13]. GLMs estimate coefficients by maximum likelihood. We tested thousands of model formulas combining linear and squared terms of the features.

2) Generalized additive models (GAMs) incorporating nonparametric smoothing functions to model nonlinear relationships between features and redshift[14]. We focused on models with only original features, avoiding overfitting from higher-order terms.

We performed an extensive hyperparameter search, evaluating models based on cross-validated correlation and root mean squared error between predicted and measured redshifts on held-out data. This identified the best-performing model formula for each method (See FIG. 3).

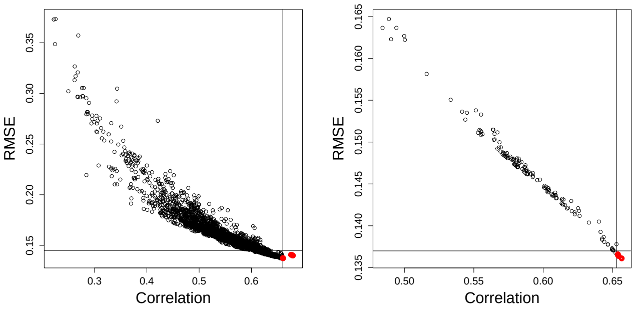

FIG. 3. The plot of the cross-validation results of RMSE and the r in the $\log(z+1)$ of GLM (left) and GAM (right) formulae. Each dot on either plot represents a formula performance within the 10fCV. The red dots represent the formulae that were above the chosen RMSE and Correlation.

E. Ensemble Modeling

The optimized GLM and GAM were combined using SuperLearner ensemble modeling. SuperLearner weights component models by performance on cross-validation. It combines models to leverage their complementary strengths. After testing many algorithms, linear techniques like GLM and GAM proved optimal for the limited GRB dataset.

F. Redshift Inference

We applied bias correction to address sample imbalance using optimal transport. Bias correction adjusts systematic errors in the predicted values compared to the true values. Our redshift sample had imbalances, with more low-z GRBs than high-z. We used optimal transport to reorder the predicted and observed redshifts and fit a linear correction function. Applying this to the predictions removed the bias. The optimized ensemble was then used to predict redshifts for 208 GRBs lacking spectroscopic measurements. To estimate uncertainties, we propagated measured errors on GRB parameters through Monte Carlo simulations.

V. RESULTS

Our ensemble model achieved a Pearson correlation coefficient of 0.93 between predicted and measured redshifts for the 25 GRB test set after bias correction. The root mean squared error was 0.46, with a median absolute deviation of 0.68. After removing 4 catastrophic outliers exceeding 2σ, the sample correlation increased to 0.96. The model accurately estimated redshifts from 0.1 to over 5.

The plateau parameters, including flux and duration, proved most informative for predicting redshift. Their inclusion in the models substantially improved accuracy. The distribution of predicted redshifts for the 208 GRB generalization set appeared consistent with the measured redshift distribution, passing statistical tests.

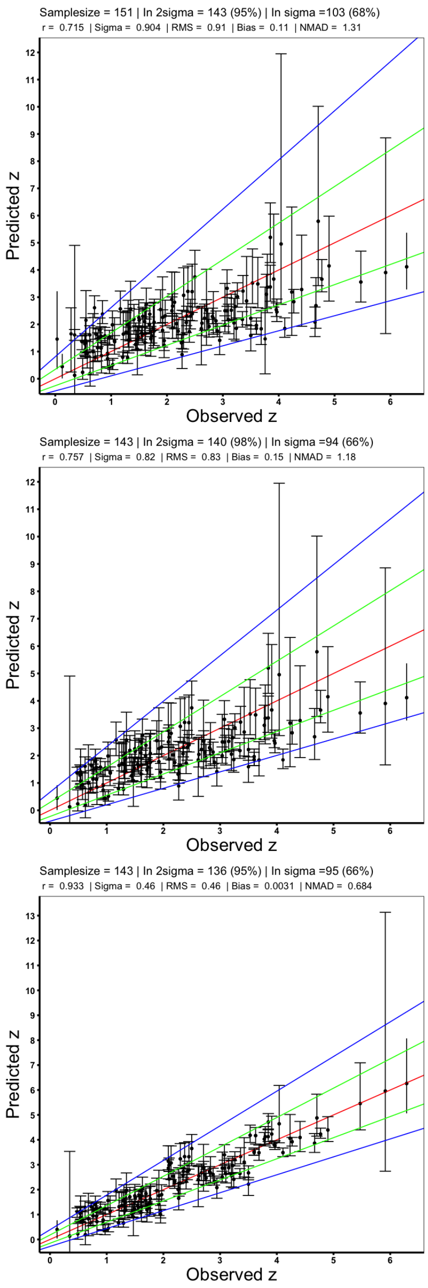

FIG. 4. The scatter plot between z_{obs} and z_{pred}. Upper panel: Predictions before removal of catastrophic outliers and bias correction. Middle panel: Predictions after the removal of catastrophic outliers. Lower panel: Predictions after the removal catastrophic outliers and application of bias correction.

VI. DISCUSSION

This machine learning approach yielded the most accurate GRB redshift predictions to date. Our study demonstrated the power of modeling diverse prompt and afterglow parameters instead of relying solely on gamma-ray information. While the plateau emission was known to correlate with luminosity, its usefulness for redshift inference was unknown. The plateau's prominence in our feature importance ranks validates its value.

By inferring redshifts for 208 GRBs, we significantly expand the sample available for Population studies and cosmology. We increased the GRBs with measured or predicted redshifts and plateaus by 94%. This enables near-term improvements in characterizing GRB populations. With continued GRB observations, our method can achieve cosmological constraints competitive with Type Ia supernovae in the next decade[1].

VII. CONCLUSION

We developed an accurate machine learning approach to infer redshifts of long GRBs from Swift data. Applying this model to GRBs lacking spectroscopic redshifts significantly expands the statistical sample for population studies and cosmology. The addition of afterglow parameters was key to improving predictions. As the GRB dataset grows, machine learning techniques show strong promise to enhance our understanding of the high-redshift universe.

ACKNOWLEDGMENTS

This work was supported by NASA Swift GI Cycle XIX grant 22-SWIFT22-0032, The Department of Energy through the Summer Undergraduate Laboratory Internship program, and the faculty and associates of the Stanford Linear Accelerator Collider National Laboratory.

[1] M. G. Dainotti, B. D. Simone, T. Schiavone, G. Montani, E. Rinaldi, G. Lambiase, M. Bogdan, and S. Ugale, “On the evolution of the hubble constant with the SNe ia pantheon sample and baryon acoustic oscillations: A feasibility study for GRB-cosmology in 2030,” Galaxies **10**, 24 (2022).

[2] A. Lien, T. Sakamoto, S. D. Barthelmy, W. H. Baumgartner, J. K. Cannizzo, K. Chen, N. R. Collins, J. R. Cummings, N. Gehrels, H. A. Krimm, C. B. Markwardt, D. M. Palmer, M. Stamatikos, E. Troja, and T. N. Ukwatta, “THE THIRDiSWIFT/iBURST ALERT TELESCOPE GAMMA-RAY BURST CATALOG,” The Astrophysical Journal **829**, 7 (2016).

[3] V. Petrosian, E. Kitanidis, and D. Kocevski, “COSMOLOGICAL EVOLUTION OF LONG GAMMA-RAY BURSTS AND THE STAR FORMATION RATE,” The Astrophysical Journal **806**, 44 (2015).

[4] M. G. Dainotti, S. Livermore, D. A. Kann, L. Li, S. Oates, S. Yi, B. Zhang, B. Gendre, B. Cenko, and N. Fraija, “The optical luminosity–time correlation for more than 100 gamma-ray burst afterglows,” The Astrophysical Journal Letters **905**, L26 (2020).

[5] S. Cao, N. Khadka, and B. Ratra, “Standardizing dainotti-correlated gamma-ray bursts, and using them with standardized amati-correlated gamma-ray bursts to constrain cosmological model parameters,” Monthly Notices of the Royal Astronomical Society **510**, 2928–2947 (2021).

[6] J. P. Norris and J. T. Bonnell, “Short Gamma-Ray Bursts with Extended Emission,” The Astrophysical Journal **643**, 266–275 (2006), arXiv:astro-ph/0601190 [astro-ph].

[7] N. Gehrels, G. Chincarini, P. Giommi, K. O. Mason, J. A. Nousek, A. A. Wells, N. E. White, S. D. Barthelmy, D. N. Burrows, L. R. Cominsky, K. C. Hurley, F. E. Marshall, P. Meszaros, P. W. A. Roming, L. Angelini, L. M. Barbier, T. Belloni, S. Campana, P. A. Caraveo, M. M. Chester, O. Citterio, T. L. Cline, M. S. Cropper, J. R. Cummings, A. J. Dean, E. D. Feigelson, E. E. Fenimore, D. A. Frail, A. S. Fruchter, G. P. Garmire, K. Gendreau, G. Ghisellini, J. Greiner, J. E. Hill, S. D. Hunsberger, H. A. Krimm, S. R. Kulkarni, P. Kumar, F. Lebrun, N. M. Lloyd-Ronning, C. B. Markwardt, B. J. Mattson, R. F. Mushotzky, J. P. Norris, J. Osborne, B. Paczynski, D. M. Palmer, H.-S. Park, A. M. Parsons, J. Paul, M. J. Rees, C. S. Reynolds, J. E. Rhoads, T. P. Sasseen, B. E. Schaefer, A. T. Short, A. P. Smale, I. A. Smith, L. Stella, G. Tagliaferri, T. Takahashi, M. Tashiro, L. K. Townsley, J. Tueller, M. J. L. Turner, M. Vietri, W. Voges, M. J. Ward, R. Willingale, F. M. Zerbi, and W. W. Zhang, “Theiswift/igamma-ray burst mission,” The Astrophysical Journal **611**, 1005–1020 (2004).

[8] M. Dainotti, D. Levine, N. Fraija, and P. Chandra, “Accounting for selection bias and redshift evolution in GRB radio afterglow data,” Galaxies **9**, 95 (2021).

[9] G. P. Srinivasaragavan, M. G. Dainotti, N. Fraija, X. Hernandez, S. Nagataki, A. Lenart, L. Bowden, and R. Wagner, “On the investigation of the closure relations for gamma-ray bursts observed by swift in the post-plateau phase and the GRB fundamental plane,” The Astrophysical Journal **903**, 18 (2020).

[10] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, Vol. 793 (John Wiley & Sons, 2019).

[11] P. J. Huber, “Robust Estimation of a Location Parameter,” The Annals of Mathematical Statistics **35**, 73 – 101 (1964).

[12] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society **58**, 267–228 (1996).

[13] J. Nelder and R. Wedderburn, “Generalized linear models,” Journal of the Royal Statistical Society. Series A **135**, 370–384 (1972).

[14] T. Hastie and R. Tibshirani, *Generalized Additive Models* (Routledge, 1990).