



Predicting the Redshift of Gamma-Ray Loud AGNs Using Supervised Machine Learning. II

Aditya Narendra¹ , Spencer James Gibson², Maria Giovanna Dainotti^{3,4} , Malgorzata Bogdan^{5,6}, Agnieszka Pollo^{1,7},
Ioannis Liodakis⁸, Artem Poliszczuk⁷, and Enrico Rinaldi^{9,10} 

¹ Astronomical Observatory of Jagiellonian University, Krakow, Poland

² Carnegie Mellon University, PA, USA

³ National Astronomical Observatory of Japan, Mitaka, Japan; maria.dainotti@nao.ac.jp

⁴ Space Science Institute, 4750 Walnut St, Suite 205, Boulder, CO 80301, USA

⁵ Department of Mathematics, University of Wroclaw, Poland

⁶ Department of Statistics, Lund University, Sweden

⁷ National Centre for Nuclear Research, Warsaw, Poland

⁸ Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Finland

⁹ University of Michigan, Ann Arbor, MI, US

¹⁰ Riken, iTHEMS Wako, Saitama, Japan

Received 2021 November 16; revised 2022 January 25; accepted 2022 January 26; published 2022 April 6

Abstract

Measuring the redshift of active galactic nuclei (AGNs) requires the use of time-consuming and expensive spectroscopic analysis. However, obtaining redshift measurements of AGNs is crucial as it can enable AGN population studies, provide insight into the star formation rate, the luminosity function, and the density rate evolution. Hence, there is a requirement for alternative redshift measurement techniques. In this project, we aim to use the Fermi Gamma-ray Space Telescope's 4LAC Data Release 2 catalog to train a machine-learning (ML) model capable of predicting the redshift reliably. In addition, this project aims at improving and extending with the new 4LAC Catalog the predictive capabilities of the ML methodology published in Dainotti et al. Furthermore, we implement feature engineering to expand the parameter space and a bias correction technique to our final results. This study uses additional ML techniques inside the ensemble method, the SuperLearner, previously used in Dainotti et al. Additionally, we also test a novel ML model called Sorted L-One Penalized Estimation. Using these methods, we provide a catalog of estimated redshift values for those AGNs that do not have a spectroscopic redshift measurement. These estimates can serve as a redshift reference for the community to verify as updated Fermi catalogs are released with more redshift measurements.

Unified Astronomy Thesaurus concepts: [Active galactic nuclei \(16\)](#); [High energy astrophysics \(739\)](#); [Blazars \(164\)](#)

Supporting material: machine-readable table

1. Introduction

The Fermi Gamma-ray Space Telescope observes the sky in the energy range of 50 MeV–1 TeV. With its γ -ray observational capabilities, it has observed 3511 γ -ray loud active galactic nuclei (AGNs) to this day. These AGNs are gathered in the latest Fourth Fermi-LAT Data Release 2 (4LAC) catalog (Abdollahi et al. 2020; Ajello et al. 2020), along with their measured γ -ray properties. The majority (>98%) of AGNs in the 4LAC catalog are blazars, which according to the Unified Model of AGN, are identified with galaxies that have their relativistic emission jets pointed directly along our line of sight (Blandford et al. 2019). Two major subcategories of blazars are present in the 4LAC catalog: BL Lacertae objects (BLLs) and flat-spectrum radio quasars (FSRQs). BLLs are classified as sources with weak or no emission lines in their spectra, whereas FSRQs are classified as sources that have broad emission lines (Ghisellini et al. 2011; Madejski & Sikora 2016).

Measuring the redshift of blazars proves to be problematic as it often requires multiwavelength observations and coordination

across multiple facilities. Further complications arise because the most dominant category of γ -ray loud blazars, BLLs, often do not show detectable emission lines. Not surprisingly, only 50% of γ -ray AGNs have redshift estimates, highlighting the difficulty in measuring it, as well as the need for a reliable photometric redshift estimation method. In light of these hurdles, a predictive approach for redshift estimation using the more readily available photometric γ -ray features is an attractive proposition.

Photometric redshifts can play a significant role in helping us understand the origin of extragalactic background light, crucial for blazars emitting GeV–TeV γ -rays, which are typically BLLs (Wakely & Horan 2008).¹¹ This can, in turn, enhance our knowledge about the cosmic evolution of blazars (Chiang et al. 1995; Ackermann et al. 2012; Singal et al. 2012, 2013a, 2013b, 2014; Ackermann et al. 2015; Singal 2015), the structure of the magnetic field in the intergalactic medium, (Venters & Pavlidou 2013; Fermi-LAT Collaboration et al. 2018; Marcotulli et al. 2020) as well as constraining cosmological parameters (Petrosian 1976; Singal et al. 2013b; Domínguez et al. 2019). In recent years, the number of studies focusing on photometric redshift estimation of high-redshift AGN, using a machine learning (ML) approach, has increased significantly (Cavuoti et al. 2014; Jones & Singal 2017; Yang



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

¹¹ <http://tevcad.uchicago.edu/>

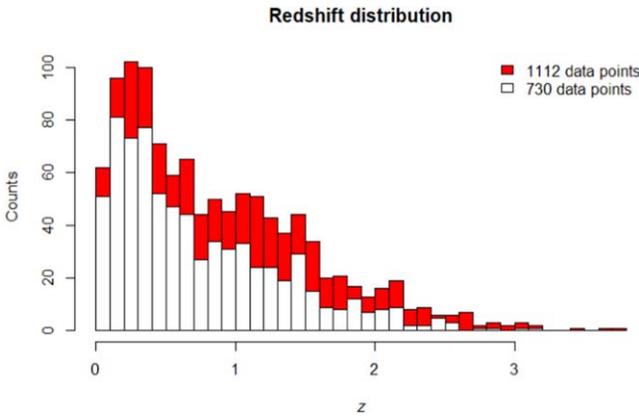


Figure 1. The histogram showing the overlapped distributions of z in Dainotti et al. (2021) training data set (white) and the current training data set (red).

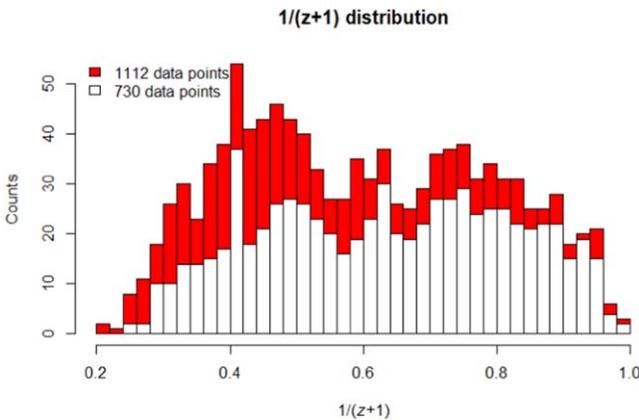


Figure 2. The overlapped histogram between Dainotti et al. (2021) data set and the current one, in the $1/(z+1)$ scale.

et al. 2017; Fotopoulou & Paltani 2018; Pasquet-Itam & Pasquet 2018; Zhang et al. 2019; Curran 2020; Logan & Fotopoulou 2020; Nakoneczny et al. 2020). This is primarily due to the availability of large data sets from all-sky surveys like the Sloan Digital Sky Survey (SDSS) (Aihara et al. 2011) and Wide-field Infrared Survey Explorer (WISE) (Ilbert et al. 2008; Hildebrandt et al. 2010; Brescia et al. 2013; Wright et al. 2010; D’Isanto & Polsterer 2018; Brescia et al. 2019). However, when it comes to γ -ray loud AGNs, ML methods have been primarily used for AGN classification (Hassan et al. 2013; Doert & Errando 2014; Chiaro et al. 2016; Einecke 2016; Lefaucheur & Pita 2017; Kang et al. 2019; Liodakis & Blinov 2019).

Dainotti et al. (2021) remains the only work in the field, as of this writing, that investigates the capabilities of Fermi in estimating photometric redshifts. However, in that study, a significant hindrance was the small data set, which also contained a large number of missing values. Since AGNs need to have both observed spectroscopic redshift and complete observational data to be included in the ML models, this led to only 50% of the data set being eligible. As a result, the small training sample contained 730 AGNs. Hence, in this research, one of our primary focuses was on increasing the training sample size, which was achieved by including AGNs from the 4LAC catalog that are present below the Galactic plane. A detailed description of the differences between the data set of

the current work and Dainotti et al. (2021) is provided in Section 6.

We also increase the number of predictors using feature-engineering techniques (see Section 3.6). We use additional ML models in our ensemble (see Section 3.3), chosen specifically for this updated data set, and an independent method called Sorted L-One Penalized Estimation (SLOPE) (see Section 3.5). Furthermore, we investigate the effects of bias correction methods (see Section 3.7) on our results. Finally, we present a list of photometric redshift predictions for the AGNs in the 4LAC that do not have a spectroscopic redshift measurement (see the Appendix).

With these updates, we were able to maintain low numbers of catastrophic outlier percentage (5%), small values of the root mean square error (RMSE), and the normalized median absolute deviation (σ_{NMAD}) as in Dainotti et al. (2021). We also managed to increase the Pearson correlation value, r , between our predicted and observed redshifts. The results on the validation set are also improved by 10%. We here stress that this has been achieved on training and validation data samples 52% larger than Dainotti et al. (2021) (see Figure 1).

2. The Sample

Since 2008, the instruments on board Fermi (LAT, GBM) have been continuously detecting γ -ray sources (e.g., γ -ray bursts, AGNs, among others). Out of the 3511 AGNs observed in the latest 4LAC catalog, only 1767 have measured redshifts, 50.3% of the total data set. In the 4LAC catalog, two categories of γ -ray loud AGNs dominate: BLLs and FSRQs. The catalog contains a total of 1190 BLLs and 733 FSRQs. A significant percentage (43%) of the AGNs are unclassified and are denoted as a Blazar Candidate of Unknown Type (BCU). We also have nine narrow-line Seyferts (NLSY1), 44 radio galaxies, and 19 AGNs of other categories. Keeping this analysis consistent with the work previously done in Dainotti et al. (2021), see Figures 1 and 2 in z and $1/(1+z)$ scale, respectively, and to improve the readability and interpretability of the method, we choose to omit all the non-BLL and non-FSRQ type AGNs from our data.

Furthermore, Fermi measures 13 properties of the 4LAC AGNs. These observations are used as predictors for the redshift in our ML models. The corresponding Gaia g -band magnitudes (Jordi et al. 2010) of the AGNs are also included. Out of these 13 features or predictors, we consider 11, excluding *Fractional_Variability* and *Highest_Energy*. Both of them were excluded due to a lack of observational data in 43% of the AGNs. The definitions of the 11 predictors are provided below.

1. *Flux*: The integral photon flux, in photons $\text{cm}^{-2} \text{s}^{-1}$, from 1 to 100 GeV.
2. *Energy_Flux*: The energy flux, in $\text{erg cm}^{-2} \text{s}^{-1}$, in the 100 MeV–100 GeV range obtained by the spectral fitting in this range.
3. *Significance*: The source detection significance in Gaussian sigma units, on the range from 50 MeV to 1 TeV.
4. *Variability_Index*: Summation of the $\log(\text{likelihood})$ differences between the fitted flux in every interval of time and the mean flux over the 50 MeV to 1 TeV range.
5. ν : The synchrotron peak frequency, measured in Hz, in the observer frame.

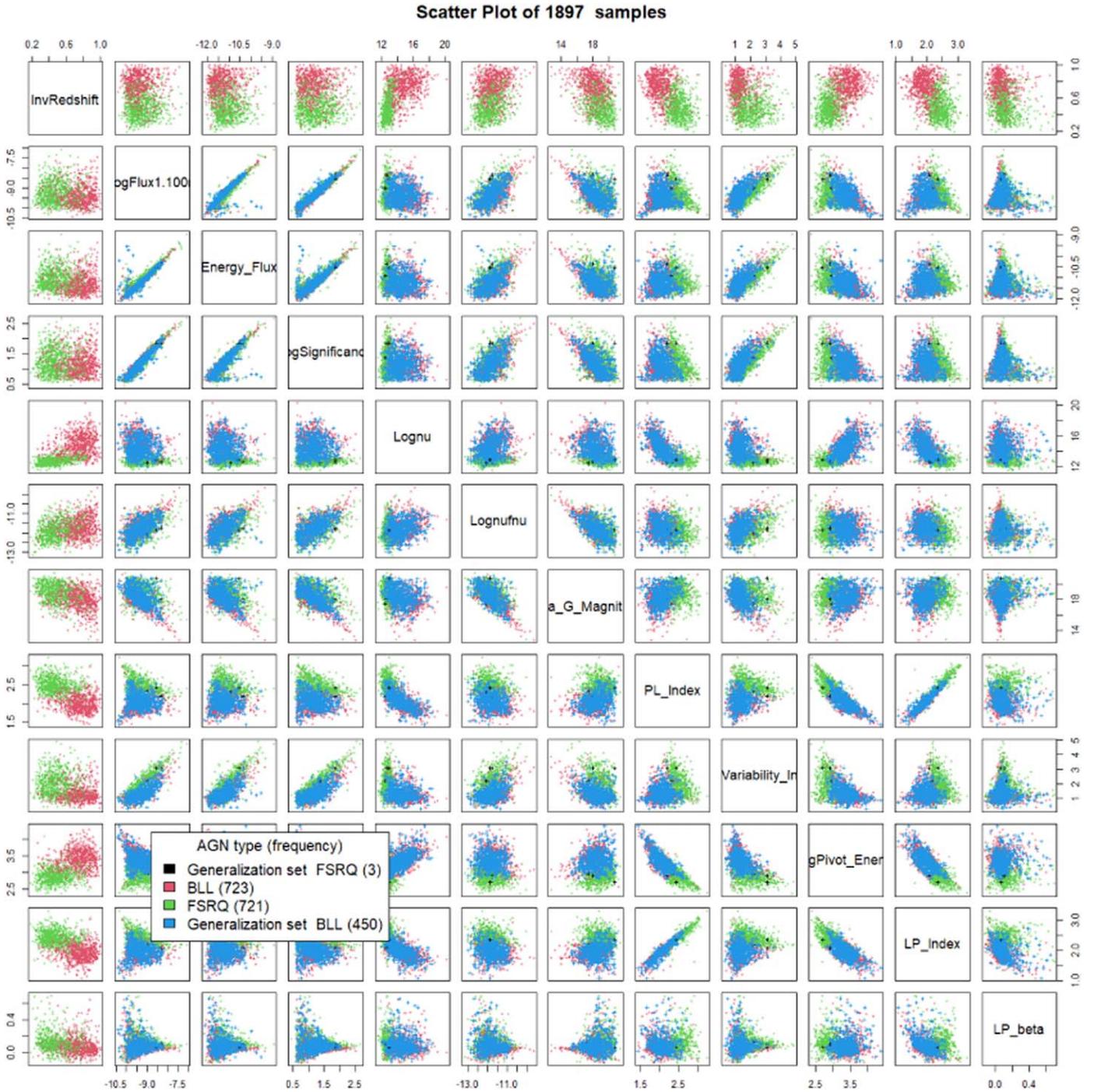


Figure 3. The full scatter matrix plot of all the variables defined above. Here *InvRedshift* denotes $1/(\tau + 1)$ scaled data. The generalization set refers to samples which do not have a redshift observation.

6. $\nu f\nu$: Spectral energy distribution at the synchrotron peak frequency in the $\nu f\nu$ spectrum.
7. *PL_Index*: The photon index when fitting the spectrum with a power law, in the energy range from 50 MeV to 1 TeV.
8. *Pivot_Energy*: The energy, in MeV, at which the error in the differential photon flux is minimal, is derived from the likelihood analysis in the range from 100 MeV–1 TeV.
9. *LP_Index*: Photon index at pivot energy (α) when fitting the spectrum (100 MeV–1 TeV) with Log Parabola.

10. *LP_beta*: the spectral parameter (β) when fitting with Log Parabola spectrum from 50 MeV to 1 TeV.
11. *Gaia_G_Magnitude*: Gaia magnitude at the *g*-band provided by the 4LAC, taken from the Gaia Survey.

2.1. Training, Generalization, and Validation Sets

Before we can train the ML models, we need to transform our predictors and create the training and generalization sets. For the predictors' *Flux*, *Energy_Flux*, *Significance*, *Variability_Index*, ν , $\nu f\nu$, and *Pivot_Energy*, a transformation to their

logarithmic scale is necessary as they span over several orders of magnitude. Also, note that we do not train our ML models to predict the redshift directly, instead, it is trained to predict $1/(z+1)$, where z is the redshift. With such a transformation, we predict the scale factor a , which is equal to $1/(z+1)$; furthermore, we found this transformation to be better for the overall performance of our ML models. The $1/(z+1)$ is indeed a much more physical quantity to estimate, as it is directly used for cosmological and redshift evolution studies in both AGNs and GRBs. (Dainotti et al. 2013a, 2013b, 2017, 2021; Singal 2015; Dainotti & Del Vecchio 2017; Dainotti & Amati 2018).

Furthermore, we also make cuts in the sample to remove data points with an $LP_{\beta} < 0.7$, $LP_{Index} > 1$, and $LogFlux > 10.5$, as they are outliers of the distribution. These cuts leave us with 1897 AGNs, out of which 1444 have observed redshift and the remaining 453 do not. The scatter matrix plot in Figure 3 shows the distribution of these 1897 samples.

The training set or the training sample is the data on which our ML models are trained to learn the underlying relations between the response variable and the predictors. Thus, it consists of AGNs with observed redshift (the response variable) and predictors with complete data. Due to missing values in three predictors, namely $Log\nu$, $Logf\nu$, and $Gaia_G_Magnitude$, we had to remove an additional 332 AGNs. This leads to a training sample of 1112 data points, which is 52% larger than Dainotti et al. (2021). The overlapped histogram of the redshifts of the entire data and the training data is presented in Figure 4.

On the other hand, the generalization set contains AGNs without observed redshift but with complete observational data. These are the AGNs for which we want to estimate the photometric redshift. Similar to the training set, if we remove the data points with missing values, we are left with 320 AGNs in the generalization sample. These 320 AGNs are made up of 319 BLLs and 1 FSRQ.

The validation set consists of AGNs, which contain measured redshift and complete predictor data similar to the training set. Hence, we take a cut of 10% from the training set to create the validation set. This set of 111 AGNs will not be used in the model's training and will serve as an independent test to verify the cross validated results.

3. Methodology

For this project, we expand upon the methodology published in Dainotti et al. (2021), which includes a more accurate version of feature selection where we select the best predictors for estimating $1/(z+1)$, additional ML models, feature engineering, and bias corrections. All the techniques and methods are described in the following subsections.

The metrics we use for measuring the goodness of our results are the following:

1. Bias, mean of the difference between the observed and predicted values.
2. σ_{NMAD} , normalized median absolute deviation between the predicted and observed measurements.
3. Pearson correlation coefficient, r , between the predicted and observed measurements.
4. Root mean square error (RMSE).
5. Standard deviation σ .

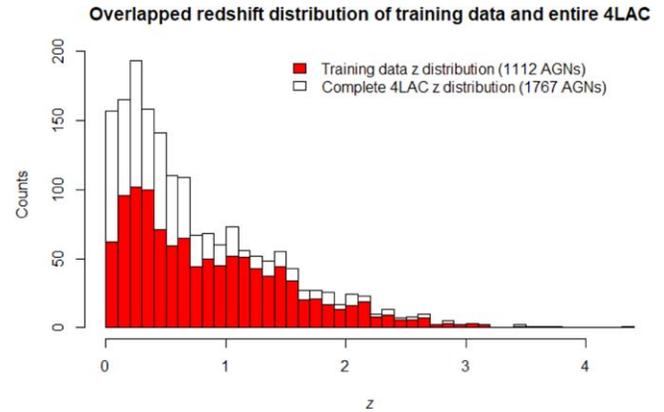


Figure 4. Overlapped histogram showing the training set and the total 4LAC redshift distributions.

We quote the measured values of these parameters for Δz_{norm} and Δz , where

$$\Delta z = z_{\text{observed}} - z_{\text{predicted}}$$

$$\Delta z_{\text{norm}} = \frac{\Delta z}{1 + z_{\text{observed}}}$$

Along with these, we also quote the catastrophic outlier percentage, which is defined as the percentage of predictions that lies beyond the 2σ error. The quoted metrics used here are the same as in Dainotti et al. (2021) and other works in the field, thus allowing for easy comparison.

3.1. Cross Validation

Cross validation (CV) is a frequently used method (Browne 2000; Refaeilzadeh et al. 2009), which estimates the expected prediction error of ML models, which in our case is the RMSE. Commonly, a 10-fold cross validation (10fCV) is used. In 10fCV, 10 equal partitions or folds are created at random. At each step, nine folds are used as training data to fit a model and the 10th fold is used as the test set, on which the prediction error is evaluated. This process is iterated such that each fold is used exactly once as a test set. The averaged errors of each test set provide a metric to judge the performance of the ML model on novel data samples. Since the folds are generated randomly, there is a risk of overfitting to a specific data partition. Therefore, we perform 10fCV 100 times, and for each iteration, the folds are generated randomly. The results are an average of these 100 iterations. Such a step is necessary to mitigate the uncertainty of 10fCV and derandomize our results. However, it should also be noted that this step is computationally expensive, as we need to train multiple ML algorithms $10 \times 100 = 1000$ times.

3.2. Feature Selection: LASSO inside 10fCV

One of the most crucial steps in any ML project is feature selection. It is the process where we choose a subset of all the predictors present in the data set that are better at predicting the response variable. Feature selection is crucial because it makes the model easier to interpret and reduces the size of the high-dimensional data, which enables algorithms to work faster. As a consequence, this also reduces the risk of overfitting. Here we use the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani 1996).

The LASSO algorithm works as a shrinkage method by constraining the ℓ^1 -norm of the solution vector to be less than or equal to a positive number known as the tuning parameter (λ). A vector's ℓ^1 -norm is defined as the sum of the magnitudes of the components of that vector. Such a constraint on the solution vector penalizes the predictor coefficients, resulting in some of the coefficients being set to 0, effectively eliminating those predictors. The λ parameter controls the shrinkage that is applied to the coefficients of the predictors. Our analysis, uses the R implementation of the GLMNET function and the LASSO feature selection mode by setting the parameter *alpha* to 1 (Tibshirani et al. 2012; Hastie et al. 2017). The function varies λ while fitting a linear model to the data and assigns coefficients to the predictors to reduce the error. As mentioned previously, the actual features that are chosen depending on the value of λ . Thus, we select those predictors with a nonzero coefficient for the highest λ value such that the prediction errors are within 1 standard deviation (Birnbbaum 1962; Hastie & Tibshirani 1987, 1990; Friedman et al. 2010a, 2010).

In this project, we perform LASSO feature selection during the CV step. This means that instead of choosing a feature set for the entire training data, we pick optimal features from the data of the nine folds that will be used to train the model at that stage. Then the model predicts on the test set (10th fold) using these features only, such that each of the 10 folds acts as a test set exactly once. Note that the LASSO feature selection technique uses the response variable to choose the optimal subset of predictors. Hence, this method assures us that we are not using information from the test set of the 10FCV, and allows us to substantiate the claim that the CV results accurately estimate the expected prediction error for this method.

3.3. The ML Algorithms Used in Our Analysis

The choice of the best ML algorithms depends significantly on the problem being addressed and the data set being used. That holds in our study as well. Previously in Dainotti et al. (2021), we tested multiple ML algorithms in an ensemble and selected the ones with the highest coefficients. We perform the same process once again, but with an expanded data set and different predictors. Unsurprisingly, different ML algorithms were chosen to be best suited for this analysis compared to Dainotti et al. (2021). In Figure 5, we present the coefficients assigned to the various ML models we tested. Those with a higher coefficient are better at predicting the redshift. Therefore, we define a threshold value of 0.5 and choose all the models with a coefficient greater than this. Consequently, we end up with six algorithms in our final ensemble. These are Enhanced Adaptive Regression Through Hinge (EARTH), KSVM, Cforest, Ranger, Random Forest, and Linear Model. We describe these methods below.

3.3.1. EARTH

EARTH is an implementation of the multivariate adaptive regression splines method (MARS), based on the work by Friedman & Roosen (1995). MARS is a nonparametric regression technique that can better model the interactions of predictors and nonlinearities than a simple linear model. MARS starts with a constant term then fits a weighted sum of functions where each function is either another constant term, a hinge function, or a product of hinge functions. Each term added to

the model results in the maximal reduction of the sum of square residual error.

3.3.2. KSVM

In this study, we use the support vector regression method in the ksvm algorithm from the SuperLearner package. Support vector regression (SVR) is the continuous predictive counterpart to support vector machine classification (Cortes & Vapnik 1995). SVR uses a kernel function that maps an input vector onto a space of higher dimension. Then it fits a hyperplane to the data, such that the predictions are within a certain error margin. In this study, the default parameters of the ksvm model in the SuperLearner package are used. Specifically, the kernel used is the radial basis kernel, i.e., the Gaussian kernel.

3.3.3. Cforest, Ranger, and Random Forest

The Cforest, Ranger, and Random Forest algorithms belong to the tree class of supervised ML models. A single decision tree algorithm may perform very well on the training data, but it is prone to overfitting. Besides the true relationships, such models learn the noise present in the training set and are referred to as models with high variance. To obtain an optimal bias–variance trade-off, one can often use ensembles made of multiple decision trees. In such models, several stages of randomness are introduced to reduce the variance of the final prediction.

Random Forest (Breiman 2001) works by generating multiple decision trees. While building the tree, the data is partitioned based on the values of a randomly chosen subset of features in the training data (Ho 1995). This partitioning repeats until the pre-specified depth of the tree is reached. At the nodes, the values of the response variables are averaged over and that average is assigned as the prediction for an AGN that falls into that node. The algorithm finally outputs the prediction as an average over all the independent trees.

The Ranger algorithm is nearly equivalent to Random Forest with the difference being a slightly quicker implementation and the ability to use extremely randomized trees (ERT). ERTs are built in the same top-down procedure as decision trees in Random Forest with the difference being that the cuts are randomly generated, assuming a uniform distribution for each of the features in the training set. The cut that reduces prediction error the most is chosen (Geurts et al. 2006). ERTs have the advantage of reducing variance further than Random Forest due to the increased randomness in the node-splitting procedure.

The Cforest algorithm (Hothorn et al. 2006) is similar to Random Forest and Ranger except that it uses conditional inference trees (Hothorn & Zeileis 2015), which splits based on significance tests instead of information gain. These tree-based regression methods can fit very complex, sophisticated models.

3.3.4. Linear Model

The linear model is the simplest regression model. Given a set of features X_1, \dots, X_n , the linear model fits a function of the form $f(X_1, \dots, X_n) = \beta_0 + \sum_{i=1}^n \beta_i * X_i$, where the coefficients β_i are real numbers. This project uses the default ordinary least squares linear model from the SuperLearner package. Here, the optimal β parameter vector is picked such that it minimizes the average squared error.

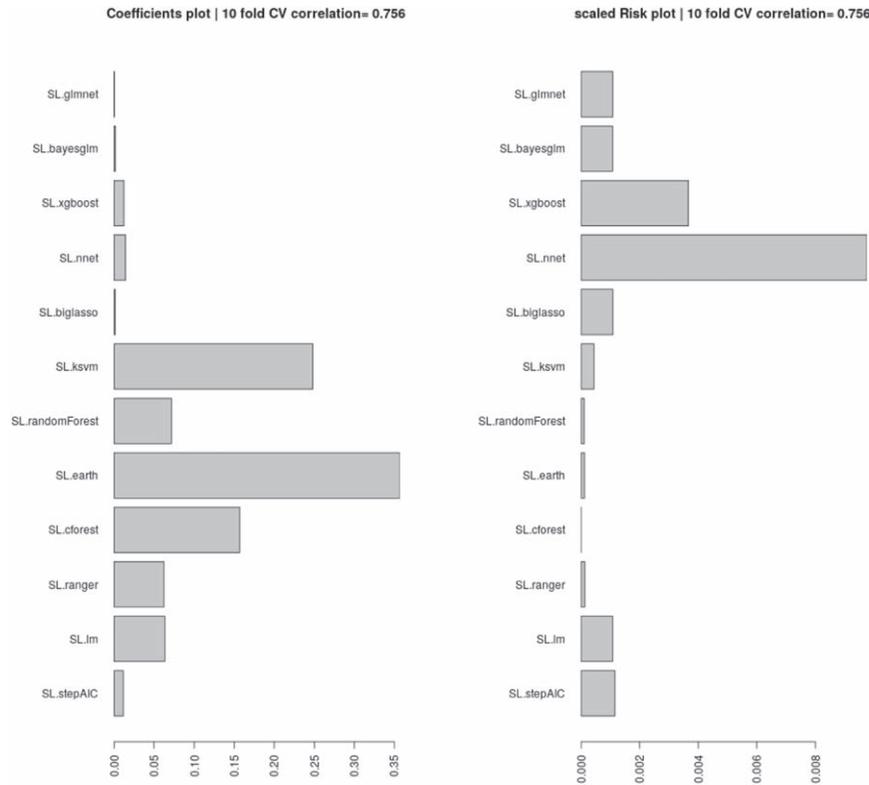


Figure 5. Left panel: the coefficients assigned by SuperLearner to the algorithms tested. We select the algorithms that have a coefficient above 0.5 to be incorporated into our ensemble. Right panel: the RMSE error (risk) of each of the algorithms, scaled to show the minimum risk algorithm at 0, which is cforest.

3.4. SuperLearner

SuperLearner (Van der Laan et al. 2007) is an ensemble learning algorithm that utilizes 10fCV. Ensemble learning is an ML technique that combines multiple individual algorithms into a single one. There are multiple methods to do this (Opitz & Maclin 1999). The advantage of such a method is that the shortcomings of one constituent algorithm can be mitigated by leveraging other algorithms. This overall has a positive impact, thereby boosting performance. In the case of SuperLearner, this ensemble is created by assigning coefficients to its constituent models. The coefficients are determined based on the RMSE of each algorithm in 10fCV. Algorithms with lower RMSE are assigned higher coefficients. The sum of all the coefficients always equals 1, and they are greater than or equal to 0. This ensemble created by SuperLearner intrinsically has the lowest RMSE among all its constituent models (Polley et al. 2010).

3.5. SLOPE

Sorted L-One Penalized Estimation, or SLOPE (Bogdan et al. 2015), is a supervised learning method for both prediction and feature selection. Given a vector $b \in \mathbb{R}^p$, and a sequence of nonnegative scalars $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p \geq 0$ where $\lambda \neq 0$, the Sorted L1-Norm is defined as $\lambda_1 b_1 + \lambda_2 b_2 + \dots + \lambda_p b_p$. SLOPE is a natural extension of the LASSO method. It aims to minimize a cost function with the constraint that the Sorted L1 Norm is less than or equal to a positive constant (Opitz & Maclin 1999). It does this by assigning larger weights for larger parameter values. We test this independent method that is not included in the SuperLearner ensemble.

3.6. Feature Engineering

It is common in ML to create new predictors from the preexisting features in the data sample (Ng 2013). These techniques, called feature engineering, can help boost the performance of the final model. The importance of feature engineering is clear when considering that the response variable may depend on combinations of the initial features that certain ML algorithms cannot learn. These include, but are not limited to, features such as ratios for neural networks, and count variables for tree algorithms (Heaton 2016). Continuing the work from Dainotti et al. (2021) where we considered models trained with the features present in 4LAC, here we extended the set of predictors. Our engineered features are simple cross multiplications of the existing features, as described in Section 2 to distinguish the feature sets, we denote the original features obtained from the 4LAC catalog as Order 1 (O1). The cross multiplied features are called Order 2 (O2). Additionally, the feature selection with LASSO and the algorithms used is done in the same way for O1 and O2 cases. The results are presented for both.

3.7. Bias Correction

We define bias as the difference between the predicted and the observed values of a response variable. Training models on an unbalanced sample often leads to predictions that have a high bias, although it should be noted that an unbalanced sample is not the sole reason for predictions to be biased. This outcome can occur for other reasons, such as the ML methods used, and the transformations applied to the predictors and the response variable. In the case of an unbalanced sample, the model is better trained on the redshift range that contains more

observations. As a result, the model predicts more frequently in those particular redshift ranges and may provide erroneous results.

We apply the technique of Optimal Transport for our bias correction. In this technique, the predicted and observed $1/(z+1)$ values are sorted in ascending order, and then a linear fit is obtained between them. The bias is corrected using the slope and intercept of this linear fit. The analysis is performed for results which are obtained in the case of O1 predictors with SuperLearner and is applied for BLLs and FSRQs independently. We correct the respective predictions using the following formula:

$$C_{\text{preds}} = B + A * U_{\text{phot}}$$

where C_{preds} is the corrected $1/(z+1)$ predictions, U_{phot} is the SuperLearner $1/(z+1)$ predictions, and A and B are the slope and intercept of the linear fits, respectively. C_{pred} values are then converted to the z -scale to obtain the bias-corrected redshift estimates.

The effect of these corrections is presented in Section 5.4. Following this procedure, we also correct the predictions on the generalization set and we present these results in Table 3, along with the predictions without the bias correction.

4. Comparison

As mentioned, this study aims at extending the method developed by Dainotti et al. (2021). Here we provide a comparison between the two studies.

Previously, in Dainotti et al. (2021), some of us worked with 2863 sources from the 4LAC, comprising of 658 FSRQs, 1067 BLLs, 1074 BCUs, and 64 sources categorized as radio galaxies, narrow-line Seyferts, or other nonblazar AGNs. Out of those 2863 sources, only BLLs and FSRQs with no missing data were used, resulting in 657 AGNs for the training set and 73 AGNs for the validation set. In the current study, the 4LAC catalog has been significantly expanded from 2863 total AGNs to 3511, an increment of 52%. This increment has been achieved by including the lower-Galactic-latitude γ -ray loud AGNs. Out of the total number of AGNs under consideration, 50.3% have a measured redshift. Thus, the first difference between the current and Dainotti et al. (2021) is the enlarged sample. Indeed, in Section 2.1, we describe how we obtained a training set of 1112 AGNs, a 52% increase compared to Dainotti et al. (2021).

In addition, part of our extension of the work in Dainotti et al. (2021) is the consideration of new ML algorithms. They had obtained the highest weights for the algorithms Extreme Gradient Boosting (XGBoost), Random Forest, Big LASSO, and Bayes Generalized Linear Model (Bayes GLM) using SuperLearner. For this study, we obtain the highest weights for KSVM, Random Forest, Linear Model, Ranger, Cforest, and EARTH (see Section 3.3) within the same SuperLearner framework.

Previously, Dainotti et al. (2021) used the property *Highest_Energy* as one of the predictors and did not take into account for *Log ν* , as this was already considered a cross product term. In this study, we are including *Log ν* and excluding *Highest_Energy* (see Section 2 for additional details), along with introducing the use of O2 predictors (see Section 3.6).

In Dainotti et al. (2021), LASSO feature selection was performed on the entire training set. Here, we perform LASSO feature selection on a fold-by-fold basis inside the 10fCV leading to more accurate results (see Section 3.2). This procedure was applied in all the trails that use SuperLearner, regardless of the predictor set.

5. Results

In this study, we tested multiple models with different parameter spaces to obtain the best possible model for predicting the redshift of γ -ray loud AGNs. For each of the methods, we perform 10fCV 100 times. The method which leads to the most promising results is the SuperLearner ensemble using O1 predictors. We then contrast this with two other experiments: SuperLearner ensemble with O2 predictors, and the SLOPE method with O1 predictors.

The experiment with SuperLearner using O2 predictors is intended to test how feature engineering affects the predictive capabilities of the model. For this we have 78 O2 predictors compared to 12 O1 predictors.

With SLOPE, we want to test how this single algorithm performs with respect to the SuperLearner ensemble. We are using the O1 predictor set with SLOPE and the same feature selection method using LASSO.

5.1. SuperLearner Ensemble with O1 Predictors

Using the O1 predictor set with the SuperLearner ensemble presents us the best results among the experiments we tested.

The results obtained are shown in the Figure 6. The top left and right panels show the cross validated results, in the $1/(z+1)$ and linear scales, respectively. In the $1/(z+1)$ scale, we obtain a Pearson correlation of 78.1% between the predicted and observed values, an RMSE of 0.119, σ of 0.119, and σ_{NMAD} of 0.113. In the linear scale, the correlation between predicted and observed redshift is 74%, RMSE is 0.467, bias is 0.094, and σ_{NMAD} is 0.321. We obtain a low catastrophic outlier percentage of 5%: 1052 AGNs are predicted within the 2σ range.

The lines in blue depict the 2σ curves for each plot, where the σ is calculated in the $1/(z+1)$ scale as follows:

$$\frac{1}{z_p + 1} = \frac{1}{z_s + 1} \pm 2\sigma,$$

where z_s is the spectroscopic redshift and z_p is the photometric redshift. The choice of our scaling causes the 2σ curve to not be a straight line on the linear scale. Thus, we use the formula mentioned below to draw them in the linear scale

$$z_p = z_s \left[\frac{1 \pm 2\sigma(z_p + 1)}{1 \mp 2\sigma} \right] \pm \frac{2\sigma}{1 \mp 2\sigma}.$$

In Figure 7, the top left and right panels show the histogram distribution of the Δz and Δz_{norm} . The bias is denoted by the red line and the blue lines show the σ values. The bottom left panel shows the relative influence of our predictors with *LP β* with the highest relative influence, followed by the *Log ν* and *Log ν* . The categorical predictor *LabelNo*, which distinguishes between BLL and FSRQ, is the seventh most influential predictor. These relative influence values presented here are an average over 100 iterations. Such a step is necessary

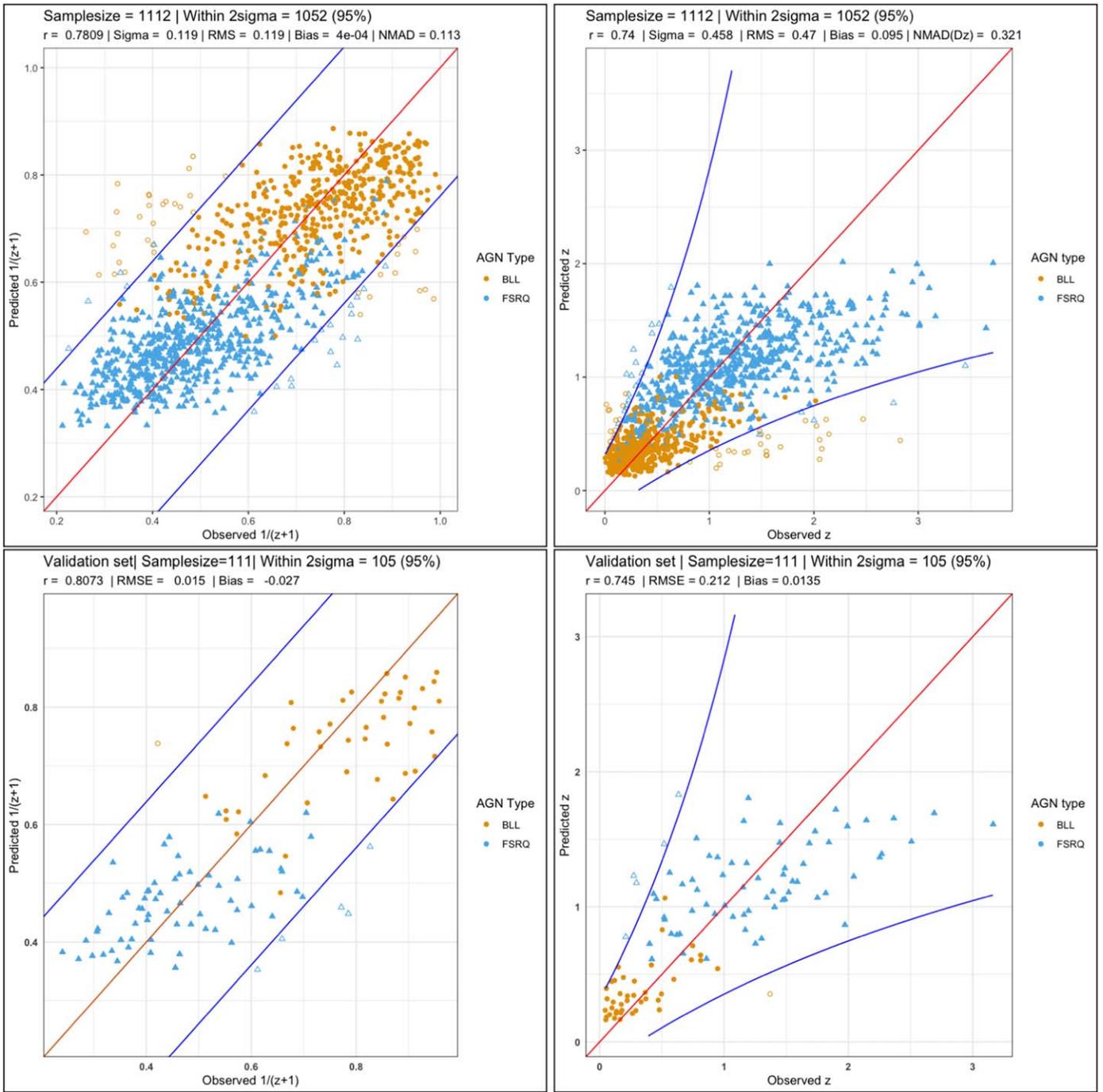


Figure 6. 10fCV results. The orange filled circles represent the BLLs and the blue filled triangles represent the FSRQs. The empty circles and triangles represent the BLLs and FSRQs that are catastrophic outliers, respectively. Top left panel: cross validation correlation plot in the $1/(z+1)$ scale. Top right panel: cross validation correlation plot in the linear scale. Bottom left panel: validation set correlation plot in the $1/(z+1)$ scale. Bottom right panel: validation set correlation plot in the linear scale.

as it serves to derandomize and stabilize the relative influence results, and thus they can be compared with Dainotti et al. (2021). Finally, the bottom right panel shows the distribution of the correlation value between observed and predicted redshift across the 100 iterations of SuperLearner. As shown, these correlations lie completely within 0.730 to 0.745, with the majority falling between 0.735 and 0.740.

In Figure 8 we are presenting the distributions of two parameters, namely RMSE and σ_{NMAD} . The top left and right panels show the distribution of σ_{NMAD} in the $1/(z+1)$ and the

linear scales, respectively. Similarly, the bottom left and right panels show the distribution of RMSE in $1/(z+1)$ and the linear scales, respectively.

5.2. Comparison with the Other Methods

In this section, we present a comparative analysis between the O1 predictors with SuperLearner and the two experiments we mentioned previously.

When using the O2 predictor set with the SuperLearner ensemble, we obtain comparable results to the O1 case. In the

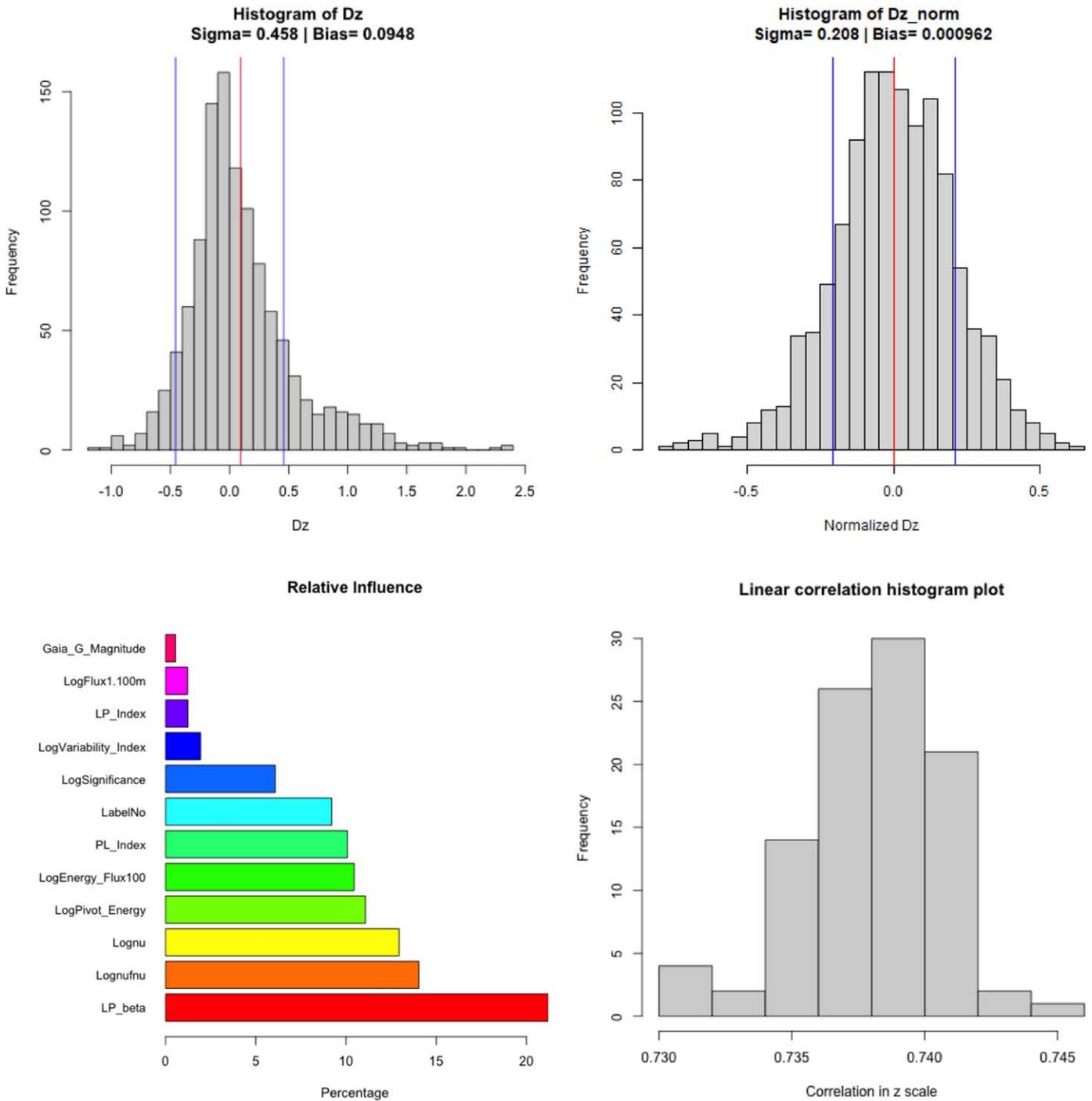


Figure 7. Top left panel: distribution of Δz with average bias (red) and sigma lines (blue). Top right panel: distribution of the Δz_{norm} with the average bias (red) and sigma values (blue). Bottom left panel: relative influence of the O1 predictors. Bottom right panel: distribution of the correlations in the linear scale from the 100 iterations of 10fCV.

$1/(z+1)$ scale, the correlation is 0.782, which is within 1%. Similarly, we obtain a RMSE of 0.119, bias of 4.5×10^{-4} , and σ_{NMAD} of 0.114.

In the linear scale, the correlation is 0.74, which is equal to the value obtained with the O1 predictors. The RMSE is 0.463, which is a 1% improvement. The bias obtained with O2 is 0.088, which is similarly a 1% improvement. The σ_{NMAD} obtained is 0.328.

As mentioned previously, we also used the SLOPE method. Compared to the SuperLearner ensemble with O1 predictors, it

performs nearly as well despite being a singular algorithm. Specifically, we obtain a Pearson correlation of 76% in the $1/(z+1)$ scale, a RMSE of 0.124, bias of 3×10^{-5} , and σ_{NMAD} of 0.12.

In the linear scale, the correlation is 72%, the RMSE is 0.479, bias is 0.1 and σ_{NMAD} is 0.348. Compared to the O1 case, in the linear scale, the R is down by 2.7%, RMSE is up by 2.5%, and the σ_{NMAD} is up by 8.4%. In both the linear scale and the $1/(z+1)$ scale, the O1 predictors with SuperLearner outperforms SLOPE, although by only a small margin.

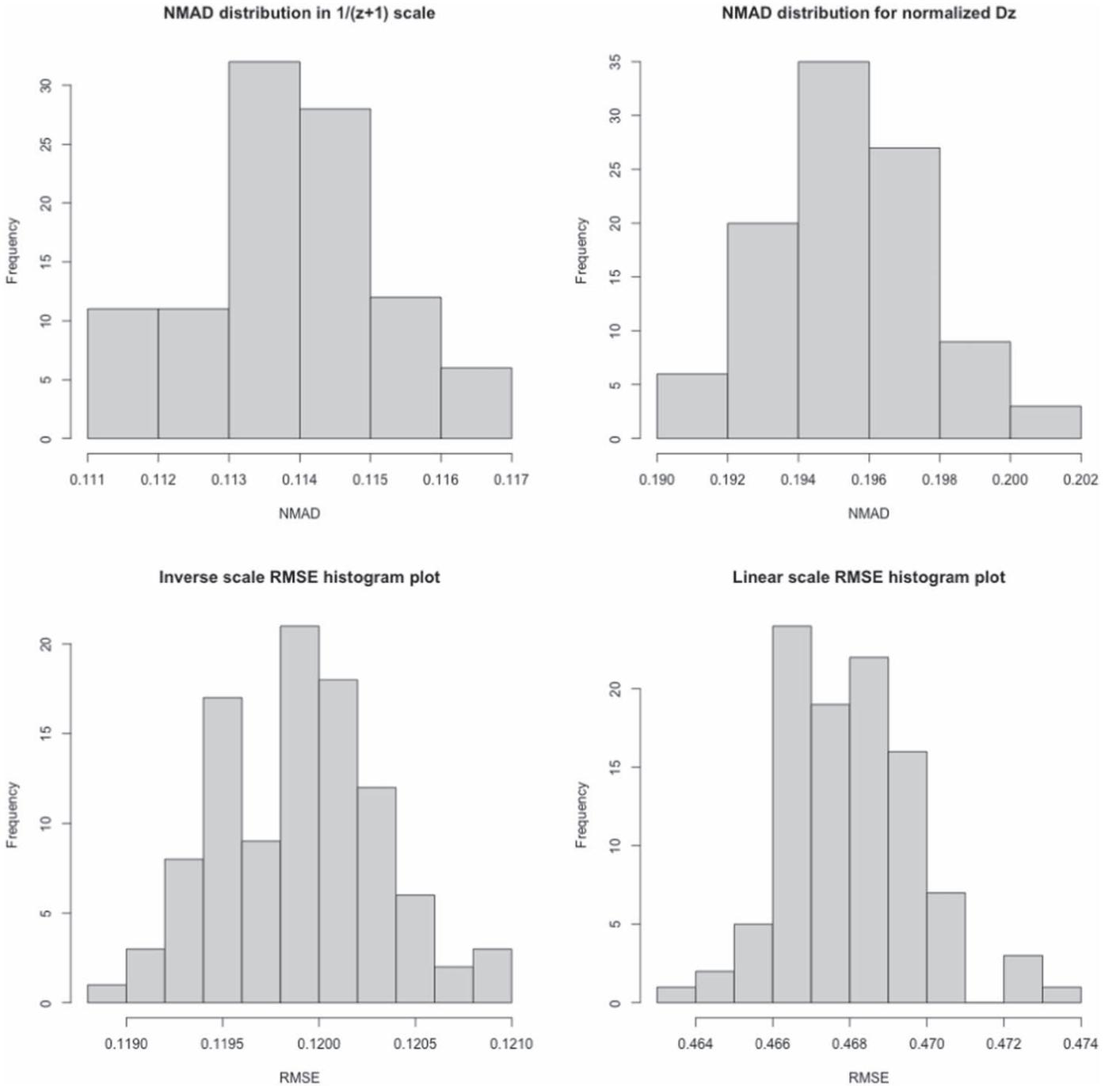


Figure 8. These distribution plots are derived from the 100 iterations of 10fCV. They represent how dispersed these metrics are for the data and algorithms at hand. Top left panel: distribution of σ_{NMAD} in the $1/(z+1)$ scale. The σ_{NMAD} is written as NMAD on the plots. Top right panel: distribution of the σ_{NMAD} in the linear scale. Bottom left panel: distribution of the RMSE in the $1/(z+1)$ scale. Bottom right panel: distribution of the RMSE in the linear scale.

All the results are presented in Tables 1 and 2, for the $1/(z+1)$ and the linear scales, respectively.

5.3. Prediction on the Generalization Set

As described in Section 2.1, the generalization set consists of AGNs that do not have a measured redshift. When predicting on this set, it is crucial to ensure that the data points fall within the parameter space of the trained model. Otherwise, the model will extrapolate those predictions and our confidence in them will be drastically reduced. We ensure this does not happen by

only keeping those AGNs in the generalization set whose maximum and minimum values for each predictor fall within the range of the same predictor in the training set. This is demonstrated in Figure 9, where the blue points (generalization-set BLLs) lie within the range of the red points (training-set BLLs). This leaves us with 305 AGNs, all of which are BLLs.

In Figure 10, we show the distribution of the predicted values of 305 BLLs, overlapped with the redshift distribution of the BLLs in the training sample. These predictions are made by using the O1 predictor set with the SuperLearner ensemble.

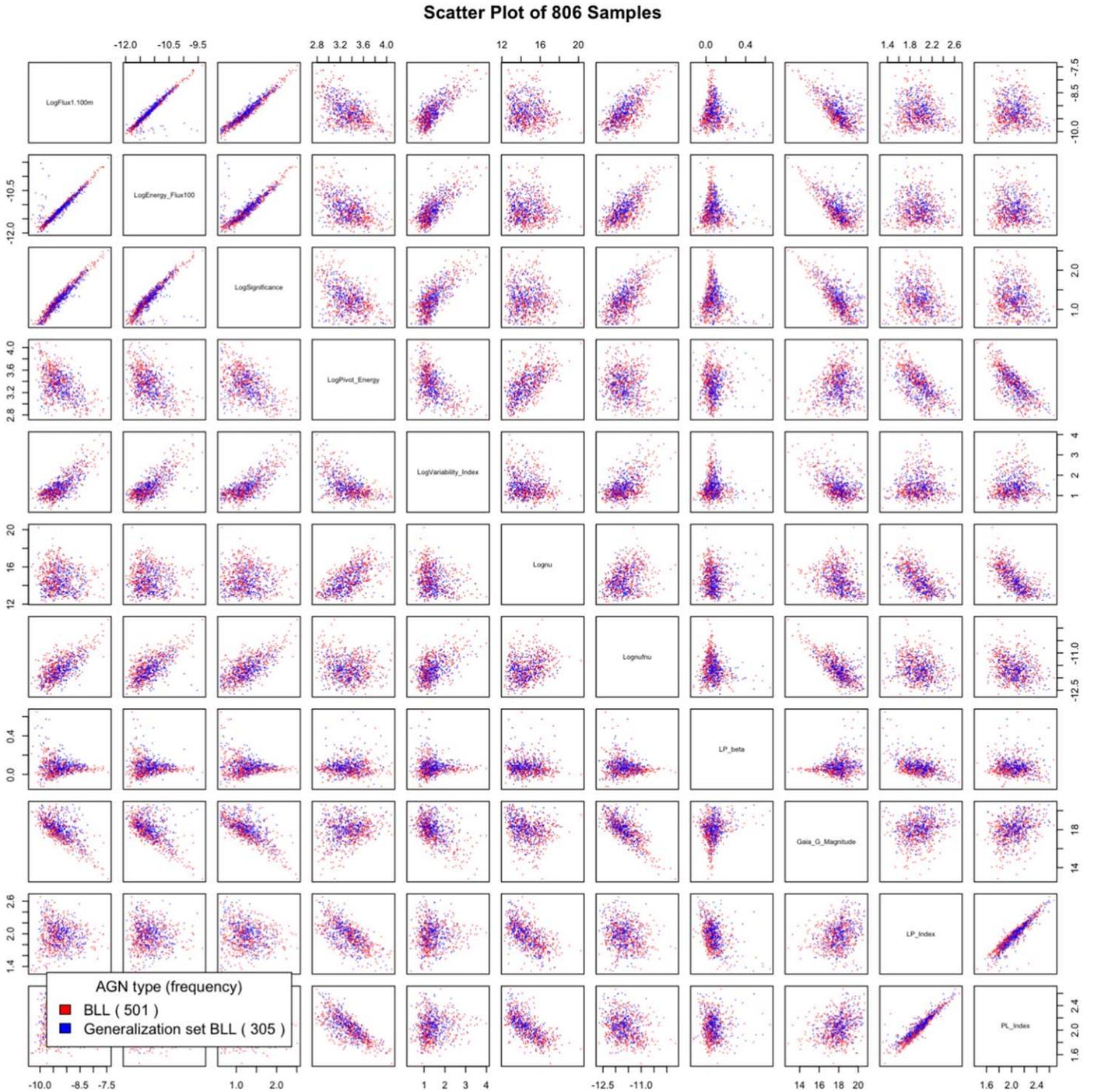


Figure 9. The scatter matrix plot for the BLLs in the training (red) and generalization (blue) samples.

5.3.1. Comparative Results

Here we consider the intersection of the generalization sets of both the current study and Dainotti et al. (2021), which consists of 240 AGNs. By plotting the estimated redshifts of these AGNs from the two studies, we see that there is a Pearson correlation of 87%. This correlation plot is presented in Figure 11. We also observe a low RMSE and σ of 0.078 and 0.079, respectively. The red line denotes the equality line, while the blue lines denote the 2σ error curves. Based on the position of the points in the plot, we note here that the redshift estimate of only eight AGNs (3%) lie beyond the 2σ error limit.

It should be noted that both these sets of estimates have not been bias corrected.

5.4. Bias Correction for the Generalization Set

Following the bias correction described in Section 3.7, the left and right panels of Figure 12 present the optimal transport linear fit of BLLs and FSRQs, respectively and the modified 10fCV results are presented in Figure 13.

We note that there is a 8.3% reduction in the correlation in the linear scale, 15.6% increase in the RMSE, and a 3.4% decrease in the σ_{NMAD} . We also see an increase in catastrophic

Table 1

 Metrics to Compare the Different Experiments Performed in the $1/(z+1)$ Scale

Metric	SL with O1	SL with O2	SLOPE with O1
R	0.781	0.782	0.762
RMSE	0.119	0.119	0.124
Bias	4×10^{-4}	4.5×10^{-4}	3×10^{-5}
σ_{NMAD}	0.113	0.114	0.12
σ	0.119	0.119	0.124

Table 2

Metrics to Compare the Different Experiments Performed in the Linear Scale

Metric	SL with O1	SL with O2	SLOPE with O1
R	0.74	0.74	0.72
RMSE (Δz)	0.467	0.463	0.479
Bias (Δz)	0.095	0.089	0.1
σ_{NMAD} (Δz)	0.321	0.328	0.348
σ (Δz)	0.458	0.455	0.47
Bias (Δz_{norm})	9.6×10^{-4}	-5.1×10^{-4}	6.9×10^{-4}
σ_{NMAD} (Δz_{norm})	0.195	0.195	0.205
σ (Δz_{norm})	0.208	0.209	0.217

outliers from 5% to 6%. However, in actual numbers, there are only two additional AGNs that are outside the 2σ cones, as compared to the O1 scenario. These results are quite expected since, with bias correction, we are reducing the accuracy in the bulk of the redshift distribution in order to get better results for the tails of the distribution. The application of this bias correction to the predictions on the generalization set lead to the results shown in Figure 14. We present the full set of bias-corrected predictions in Table 3.

6. Conclusions and Discussion

In Dainotti et al. (2021), some of us had obtained a Pearson correlation of 0.71 and an RMSE of 0.434 in the linear scale. In this study, using the O1 predictors with the SuperLearner, we achieve a Pearson correlation of 0.74 and an RMSE of 0.467. Specifically, an improvement of 4.2% in the correlation and an increase of 7.6% in RMSE. At the same time, the percentage of catastrophic outliers remains the same at 5% between this study and Dainotti et al. (2021).

We found that using O2 predictors leads to performance equivalent to O1 predictors. Thus, we can deduce that O1 predictors contain all the information needed to predict the redshift. This is evident in Figure 15 as the top 10 O2 predictors with the highest relative influence are either O1 predictors or cross products of other high influence O1 predictors (see bottom left panel of Figure 7). Such an outcome could possibly be exclusive to the 4LAC properties, and indeed cross product terms in other data sets might lead to better performance.

Additionally we tested SLOPE independently of SuperLearner. On average we observe that the results from SLOPE are comparable to those from the ensemble via the metrics

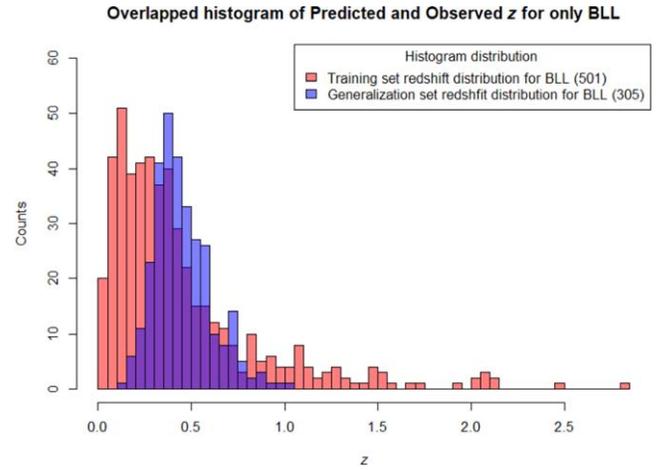


Figure 10. The distribution of the predicted redshifts for the generalization set from the model trained on O1 predictors, overlapped with the redshift distribution of the BLLs in the training set.

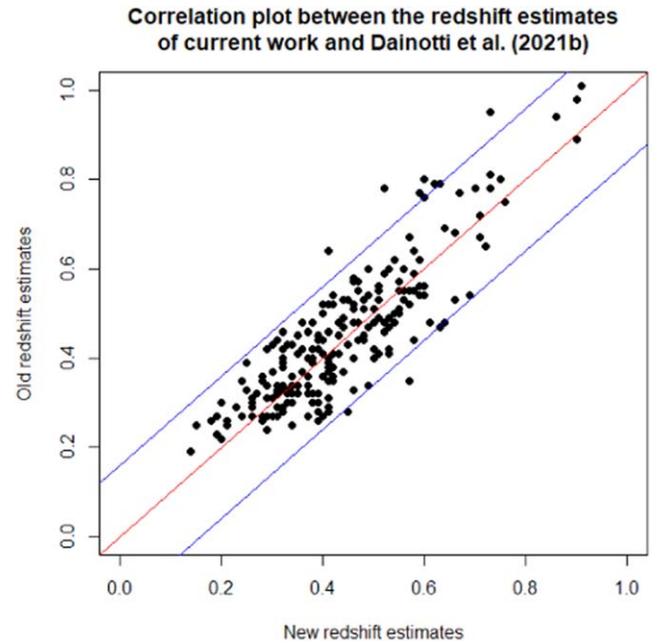


Figure 11. Comparison between the redshift estimates of 240 generalization-set AGNs, common between this study and Dainotti et al. (2021). We obtain a correlation of 0.87, RMSE of 0.078, and σ of 0.079.

presented in Table 1. However, the advantage of SLOPE is that it works much faster than the SuperLearner ensemble, along with the added benefit of high interpretability of the algorithm. Hence, when the requirement is to obtain estimates quickly, SLOPE can be used as a reliable alternative.

Here we summarize about the predictors used in our study.

1. LP_{β} , $Logv_{fv}$, $Logv$, $LogPivot_Energy$, $LogEnergy_Flux$, PL_Index , $LabelNo$, and $LogSignificance$ have highest influence in this study.
2. LP_{β} , $LogPivot_Energy$, $LogSignificance$, $LogEnergy_Flux$, $Gaia_G_Magnitude$, $Logv$, and $Log_Highest_Energy$ had the highest influence in Dainotti et al. (2021).
3. In this study and Dainotti et al. (2021), LP_{β} has the highest influence at 25%.

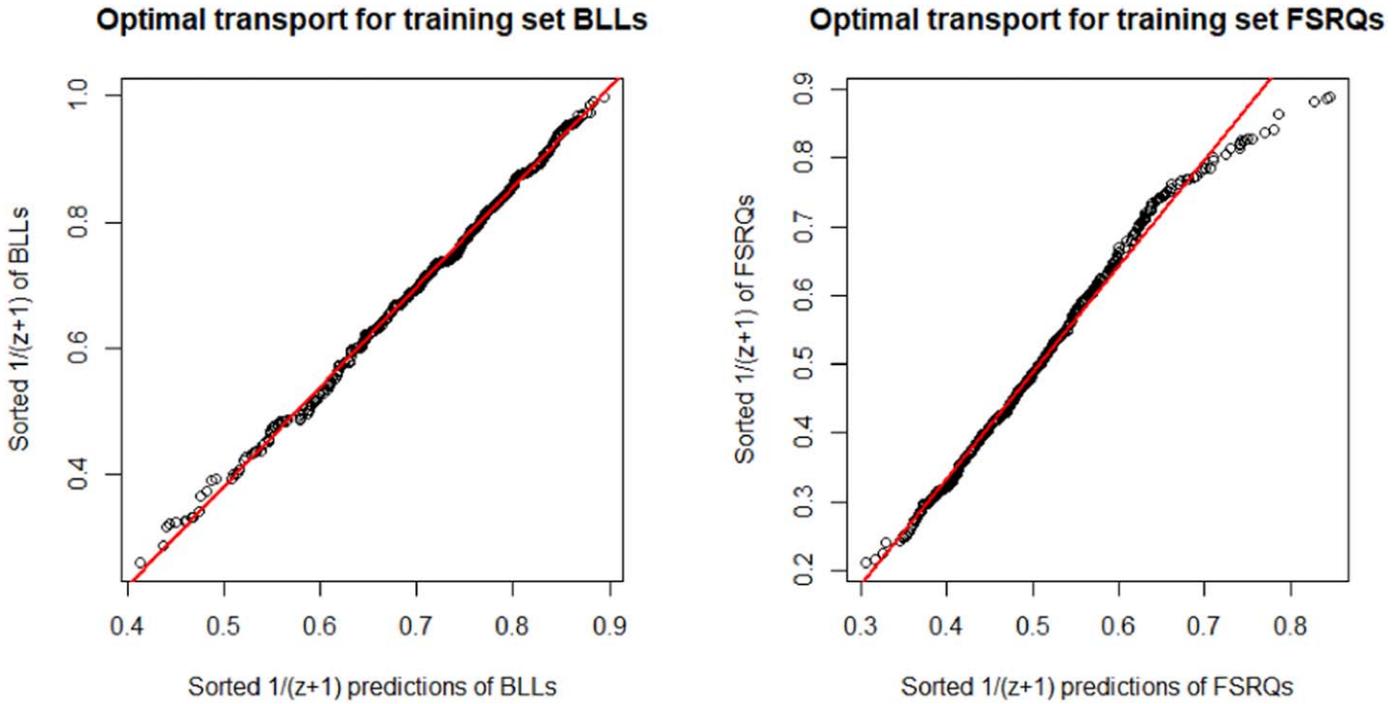


Figure 12. Optimal transport linear fits of BLLs, left panel, and FSRQs, right panel.

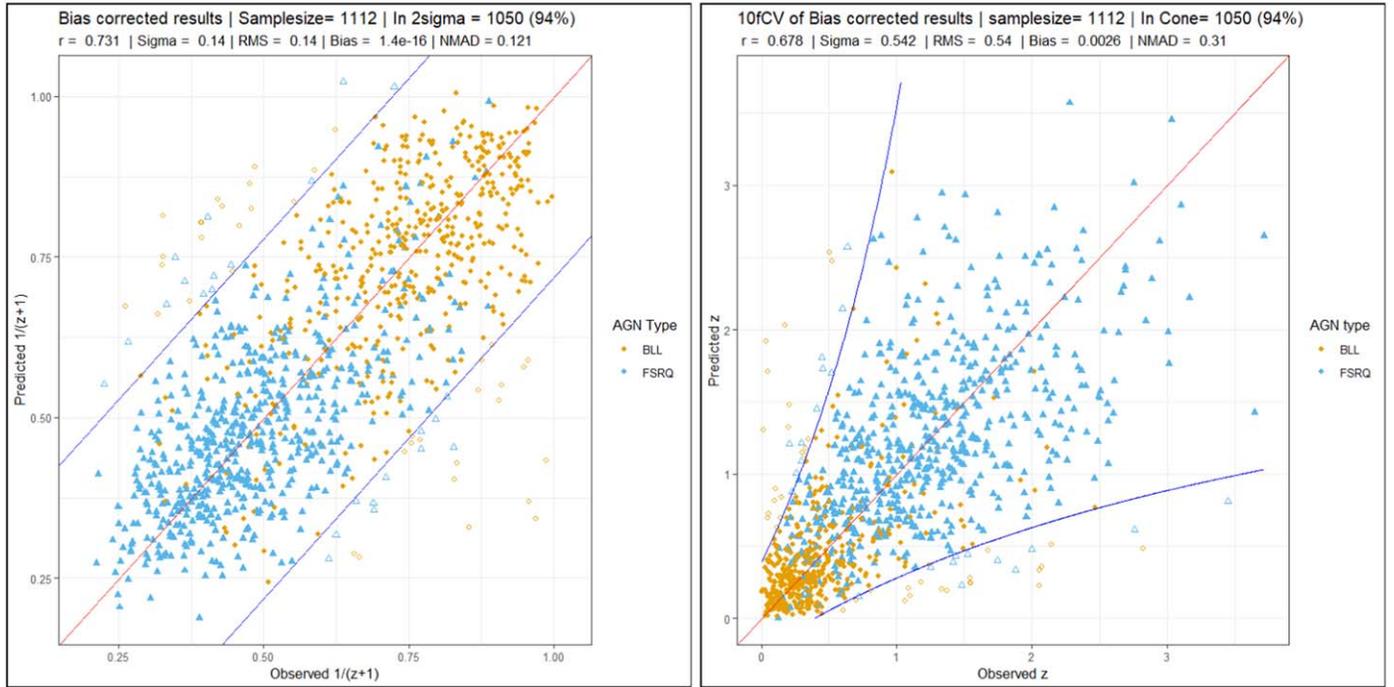


Figure 13. The bias-corrected predictions of the 10fCV. The symbols here retain the same meaning as those in Figure 6. Left panel: correlation plot between the bias-corrected prediction and observed redshift in the $1/(z+1)$ scale. Right panel: correlation plot between the bias-corrected prediction and observed redshift in the linear scale.

4. With O2 predictors, LP_β has a relative influence of 3% at the 10th position. Three O2 predictors more influential than LP_β are cross products of it. These are LP_β^2 , $LogSignificance*LP_\beta$ and $LogVariability_Index*LP_\beta$, with $LogVariability_Index*LP_\beta$ being the second most influential at 10%.

5. Hence, we can conclude that for redshifts estimation of 4LAC AGNs, LP_β is crucial.
 6. $Gaia_G_Magnitude$ has a relative influence of 1% at 12th position. In Dainotti et al. (2021) however it was the 5th most influential at 12% relative influence. This shift in the relative influence of $Gaia_G_Magnitude$ can be attributed

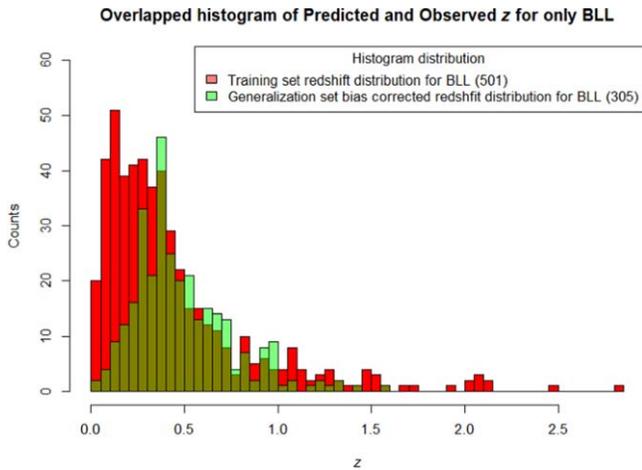


Figure 14. Histogram distributions of bias-corrected predictions of the generalization-set AGNs, overlapped with the training-set AGNs.

to the enlarged data set. However, this is not conclusive and merits further investigation.

7. PL_Index has the highest influence when paired with O2 predictors at 15%. With O1 predictors, it was the sixth most influential at 11% influence.

Finally, to make the 4LAC more complete, we present two lists containing the predictions on the generalization set. One contains the predictions from the SuperLearner ensemble with O1 predictors (shown in Figure 10) and the second contains the bias-corrected results (shown in Figure 14). Additionally, we provide a comparison between the redshift estimates of 240 AGNs that are common between the generalization sets of this study and Dainotti et al. (2021) (see Figure 11). These two results have a Pearson correlation of 0.87, RMSE and σ of 0.078 and 0.079, respectively.

This is the most complete investigation into the predictive capabilities of the 4LAC dealing with redshift estimation. Compared to the Dainotti et al. (2021), we have been able to maintain the predictive capabilities of our methodology with an even larger data sample. These estimates will likely serve as a reference point for the community to be used as tools for

cosmological investigations. The final aim is to use the estimation of the redshift to investigate AGN with interesting properties.

This work presents results from the European Space Agency (ESA) space mission, Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is <https://www.cosmos.esa.int/gaia>. The Gaia archive website is <https://archives.esac.esa.int/gaia>. M.G.D. thanks Trevor Hastie for the interesting discussion on overfitting problems. We also thank Raymond Wayne for the initial computation and discussions about balanced sampling techniques which will be implemented in subsequent papers. This research was 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. We would also like to acknowledge Biagio De Simone for their invaluable discussions regarding bias correction and SLOPE implementation. This research was also supported by the MNS2021 grant N17/MNS/000057 by the Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University.

Appendix

Here we present Table 3 containing the predictions of 50 generalization-set AGNs. It contains the names of AGNs, as assigned in the 4LAC catalog, followed by the predicted redshift from the SuperLearner ensemble with O1, and finally the bias-corrected redshifts. The predictions for the complete generalization set can be found in the data file provided.

The relative influence of the top 10 O2 predictors are shown in Figure 15.

Due to BLLs being the major category of AGNs present in the generalization set of the 4LAC catalog, we here present in Figure 16 the distribution of predicted redshift made from a Superlearner model trained exclusively on the BLLs of the training set (501 BLL). We perform the Kolmogorov–Smirnov (KS) test to compare the distribution of the predicted redshifts of generalization samples of BLLs, obtained from the model trained only on BLLs and the O1 SuperLearner model. We

conclude based on the KS test ($p = 0.66$) that we cannot reject the null hypothesis that the two distributions are drawn from the same population. Thus, we have stable results, which are resistant to the construction of the training sample, either with only BLLs or with the total 4LAC sample.

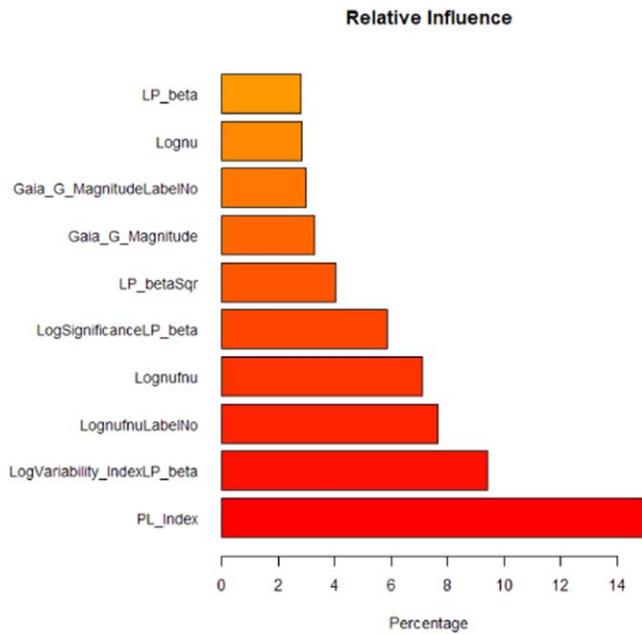


Figure 15. This plot presents the relative influence of the top 10 O2 predictors.

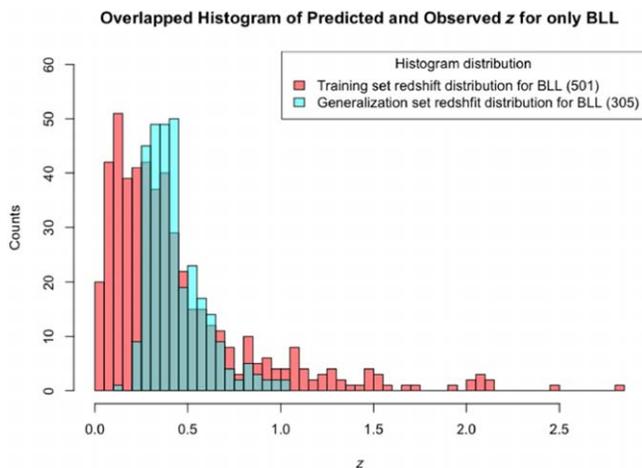


Figure 16. Distribution of the redshifts of the training data (red) and the predicted redshifts of the generalization set, from a Superlearner trained only on BLLs (turquoise).

Table 3
List of Predictions for the Generalization Set

Name	Predicted z_{pred}	Bias-corrected z_{pred}
4FGL J0009.3+5030	0.57	0.62
4FGL J0001.2-0747	0.39	0.39
4FGL J0013.1-3955	0.7	0.90
4FGL J0019.3-8152	0.24	0.17
4FGL J0019.6+2022	0.79	1.09
4FGL J0021.5-2552	0.48	0.51
4FGL J0022.1-1854	0.57	0.65
4FGL J0022.5+0608	0.76	1.03
4FGL J0023.9+1603	0.63	0.78
4FGL J0029.0-7044	0.59	0.72
4FGL J0031.3+0726	0.32	0.27
4FGL J0045.3+2128	0.34	0.28
4FGL J0056.4-2118	0.51	0.54
4FGL J0058.0-3233	0.54	0.62
4FGL J0058.3+1723	0.37	0.36
4FGL J0107.4+0334	0.4	0.48
4FGL J0112.8-7506	0.39	0.38
4FGL J0115.6+0356	0.48	0.52
4FGL J0116.0-2745	0.31	0.26
4FGL J0116.2-6153	0.39	0.37
4FGL J0120.4-2701	0.34	0.28
4FGL J0125.3-2548	0.71	0.95
4FGL J0127.2+0324	0.34	0.29
4FGL J0132.8-4413	0.37	0.35
4FGL J0136.5+3906	0.34	0.28
4FGL J0138.0+2247	0.26	0.19
4FGL J0139.0+2601	0.33	0.29
4FGL J0142.7-0543	0.36	0.32
4FGL J0143.7-5846	0.21	0.11
4FGL J0144.6+2705	0.73	0.95
4FGL J0153.0+7517	0.31	0.26
4FGL J0155.0+4433	0.54	0.62
4FGL J0158.5-3932	0.4	0.39
4FGL J0159.7-2740	0.39	0.37
4FGL J0202.7+4204	0.51	0.57
4FGL J0203.6+7233	0.47	0.51
4FGL J0208.3-6838	0.44	0.46
4FGL J0209.3+4449	0.29	0.23
4FGL J0219.5+0724	0.42	0.44
4FGL J0226.5-4441	0.32	0.28
4FGL J0232.9+2608	0.35	0.30
4FGL J0241.0-0505	0.58	0.69
4FGL J0243.4+7119	0.66	0.84
4FGL J0245.1-0257	0.45	0.48
4FGL J0258.1+2030	0.6	0.73
4FGL J0304.9-0606	0.48	0.50
4FGL J0309.7-0745	0.32	0.26
4FGL J0310.6-5017	0.28	0.21
4FGL J0312.5-2221	0.37	0.34
4FGL J0314.3-5103	0.55	0.64
4FGL J0316.2-6437	0.41	0.39

Note. The first column has the names of the AGNs as they appear in the 4LAC catalog. The second column contains the redshift predictions from the SuperLearner ensemble using O1 predictors. The third column contains the same predictions with bias correction. Here z_{pred} denotes the predicted redshift. (This table is available in its entirety in machine-readable form.)

ORCID iDs

Aditya Narendra  <https://orcid.org/0000-0003-1265-2981>Maria Giovanna Dainotti  <https://orcid.org/0000-0003-4442-8546>Enrico Rinaldi  <https://orcid.org/0000-0003-4134-809X>

References

- Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, *ApJS*, **247**, 33
- Ackermann, M., Ajello, M., Allafort, A., et al. 2012, *ApJ*, **755**, 164
- Ackermann, M., Ajello, M., Albert, A., et al. 2015, *ApJL*, **813**, L41
- Aihara, H., Prieto, C. A., An, D., et al. 2011, *ApJS*, **193**, 29
- Ajello, M., Angioni, R., Axelsson, M., et al. 2020, *ApJ*, **892**, 105
- Birnbaum, A. 1962, *J. Am. Stat. Assoc.*, **57**, 269
- Blandford, R., Meier, D., & Readhead, A. 2019, *ARA&A*, **57**, 467
- Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., & Candès, E. J. 2015, *AnApS*, **9**, 1103
- Breiman, L. 2001, *Mach. Learn.*, **45**, 5
- Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, *ApJ*, **772**, 140
- Brescia, M., Salvato, M., Cavuoti, S., et al. 2019, *MNRAS*, **489**, 663
- Browne, M. W. 2000, *J. Math. Psychol.*, **44**, 108
- Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2014, *MNRAS*, **437**, 968
- Chiang, J., Fichtel, C., Von Montigny, C., Nolan, P., & Petrosian, V. 1995, *ApJ*, **452**, 156
- Chiaro, G., Salvetti, D., La Mura, G., et al. 2016, *MNRAS*, **462**, 3180
- Cortes, C., & Vapnik, V. 1995, *Mach. Learn.*, **20**, 273
- Curran, S. 2020, *MNRAS*, **493**, L70
- Dainotti, M. G., & Amati, L. 2018, *PASP*, **130**, 051001
- Dainotti, M. G., Bogdan, M., Narendra, A., et al. 2021, *ApJ*, **920**, 118
- Dainotti, M. G., Cardone, V. F., Piedipalumbo, E., & Capozziello, S. 2013a, *MNRAS*, **436**, 82
- Dainotti, M. G., De Simone, B., Schiavone, T., et al. 2021, *ApJ*, **912**, 150
- Dainotti, M. G., & Del Vecchio, R. 2017, *NewAR*, **77**, 23
- Dainotti, M. G., Hernandez, X., Postnikov, S., et al. 2017, *ApJ*, **848**, 88
- Dainotti, M. G., Petrosian, V., Singal, J., & Ostrowski, M. 2013b, *ApJ*, **774**, 157
- D'Isanto, A., & Polsterer, K. L. 2018, *A&A*, **609**, A111
- Doert, M., & Errando, M. 2014, *ApJ*, **782**, 41
- Domínguez, A., Wojtak, R., Finke, J., et al. 2019, *ApJ*, **885**, 137
- Einecke, S. 2016, *Galax*, **4**, 14
- Fermi-LAT Collaboration, Abdollahi, S., Ackermann, M., et al. 2018, *Sci*, **362**, 1031
- Fotopoulou, S., & Paltani, S. 2018, *A&A*, **619**, A14
- Friedman, J., Hastie, T., & Tibshirani, R. 2010, *J. Stat. Softw.*, **33**, 1
- Friedman, J. H., & Roosen, C. B. 1995, *Stat. Methods Med. Res.*, **4**, 197
- Geurts, P., Ernst, D., & Wehenkel, L. 2006, *Mach. Learn.*, **63**, 42
- Ghisellini, G., Tavecchio, F., Foschini, L., & Ghirlanda, G. 2011, *MNRAS*, **414**, 2674
- Hassan, T., Mirabal, N., Contreras, J., & Oya, I. 2013, *MNRAS*, **428**, 220
- Hastie, T., & Tibshirani, R. 1987, *J. Am. Stat. Assoc.*, **82**, 371
- Hastie, T., Tibshirani, R., & Tibshirani, R. J. 2017, arXiv:1707.08692
- Hastie, T. J., & Tibshirani, R. J. 1990, *Generalized Additive Models*, Vol. 43 (Boca Raton, FL: CRC Press)
- Heaton, J. 2016, in *SoutheastCon 2016* (Piscataway, NJ: IEEE)
- Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, *A&A*, **523**, A31
- Ho, T. K. 1995, in *Proc. Third Int. Conf. on Document Analysis and Recognition, ICDAR '95* (Los Alamitos, CA: IEEE Computer Society), 278
- Hothorn, T., Hornik, K., & Zeileis, A. 2006, *J. Comput. Graph. Stat.*, **15**, 651
- Hothorn, T., & Zeileis, A. 2015, *J. Mach. Learn. Res.*, **16**, 3905
- Ilbert, O., Capak, P., Salvato, M., et al. 2008, *ApJ*, **690**, 1236
- Jones, E., & Singal, J. 2017, *A&A*, **600**, A113
- Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, *A&A*, **523**, A48
- Kang, S.-J., Fan, J.-H., Mao, W., et al. 2019, *ApJ*, **872**, 189
- Lefaucheur, J., & Pita, S. 2017, *A&A*, **602**, A86
- Lioudakis, I., & Blinov, D. 2019, *MNRAS*, **486**, 3415
- Logan, C., & Fotopoulou, S. 2020, *A&A*, **633**, A154
- Madejski, G. G., & Sikora, M. 2016, *ARA&A*, **54**, 725
- Marcotulli, L., Ajello, M., & Di Mauro, M. 2020, *AAS Meeting*, **235**, 405.06
- Nakoneczny, S. J., Bilicki, M., Pollo, A., et al. 2020, *A&A*, **649**, A81
- Ng, A. 2013, *Machine Learning and AI via Brain Simulations*, <https://forum.stanford.edu/events/2011/2011slides/plenary/2011plenaryNg.pdf>
- Opitz, D., & Maclin, R. 1999, *J. Artif. Intell. Res.*, **11**, 169
- Pasquet-Itam, J., & Pasquet, J. 2018, *A&A*, **611**, A97
- Petrosian, V. 1976, *ApJL*, **209**, L1
- Polley, E. C., Rose, S., & Van der Laan, M. J. 2010, *Targeted Learning: Causal Inference for Observational and Experimental Data* (New York: Springer), 43
- Refaeilzadeh, P., Tang, L., & Liu, H. 2009, *Encyclopedia of Database Systems* (Boston, MA: Springer US), 532
- Singal, J. 2015, *MNRAS*, **454**, 115
- Singal, J., Ko, A., & Petrosian, V. 2013a, in *IAU Symp. 304, Multiwavelength AGN Surveys and Studies* (Cambridge: Cambridge Univ. Press), 149
- Singal, J., Ko, A., & Petrosian, V. 2014, *ApJ*, **786**, 109
- Singal, J., Petrosian, V., & Ajello, M. 2012, *ApJ*, **753**, 45
- Singal, J., Petrosian, V., & Ko, A. 2013b, *AAS/High Energy Astrophysics Division*, **13**, 300.07
- Tibshirani, R. 1996, *J. R. Stat. Soc. B*, **58**, 267
- Tibshirani, R., Bien, J., Friedman, J., et al. 2012, *J. R. Stat. Soc. B*, **74**, 245
- Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. 2007, *Stat. Appl. Genet. Mol. Biol.*, **6**
- Venters, T. M., & Pavlidou, V. 2013, *MNRAS*, **432**, 3485
- Wakely, S. P., & Horan, D. 2008, *Proc. ICRC (Mérida)*, **30**, 1341
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Yang, Q., Wu, X.-B., Fan, X., et al. 2017, *AJ*, **154**, 269
- Zhang, K., Schlegel, D. J., Andrews, B. H., et al. 2019, *ApJ*, **883**, 63