PROCEEDINGS OF SCIENCE

Machine learning applications for the study of AGN physical properties using photometric observations

Sarah Mechbal,[a,*] **Markus Ackermann**[a] **and Marek Kowalski**[a]

[a] *DESY, Zeuthen, Germany*

E-mail: sarah.mechbal@desy.de

Enlarging the sample and sky coverage of AGN observations with reliably estimated physical parameters is particularly important for multimessenger astronomy, where signals from individual sources are often weaker, such that searching for correlations between a population class (e.g, AGN) and a messenger (e.g., neutrinos or cosmic rays) is common. However, knowledge of physical parameters of AGN, such as the mass of the central black hole M_{BH} and the Eddington ratio λ_{Edd}, are limited by the feasibility of large spectroscopic follow-up surveys. We show an application of machine learning (ML) techniques to reconstruct AGN physical parameters using multi-wavelength photometric observations only, in the soft X-ray, mid-infrared, and optical bands, as a way to increase the number of characterized AGN. We present a catalogue of 21 364 newly reconstructed AGN, with redshifts ranging from $0 < z < 2.5$. The redshift z, the bolometric luminosity L_{Bol}, M_{BH}, λ_{Edd}, and AGN class (obscured or unobscured) are estimated with their associated uncertainty, using a Support Vector Regression and Random Forest algorithms.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

*Speaker

PoS(ICRC2023)1569

1. Introduction

Active Galactic Nuclei (AGN), the most luminous sources in the Universe, consist of a central supermassive black hole (SMBH) around which an accretion disk is formed. They are favored as a source of cosmic ray acceleration and neutrino production [1–3]. Collecting a large and unbiased sample of AGN physical parameters, such as their redshift z, SMBH mass M_{BH} and Eddington ratio λ_{Edd}, is an important task for multimessenger studies. Traditionally, spectroscopic techniques are needed to derive these variables, but the discrepancy between photometric observations and spectroscopic follow-up of AGN remain large. We report on a new study using multi-wavelength photometric observations and machine learning (ML) regression to reconstruct fundamental parameters for 21 364 AGN, and ML classification to distinguish Type 1 (unobscured) and Type 2 AGN (obscured).

2. Data

Our goal is to compile the largest possible catalogue of non-blazar AGN sources observed in the X-ray, optical and infrared (IR) bands, but have not been followed-up spectroscopically. Crucially, a *training* sample is needed, for the ML model to learn the correlations between photometric and spectroscopic parameters: we use the recent SPIDERS-AGN spectroscopic survey [4–6], which has released a sample of 7616 Type 1 AGN with information on z, L_{Bol}, M_{BH}, λ_{Edd}. The multi-wavelengths surveys used to construct the inputs parameters for the catalogue of 21 364 sources to be reconstruct by the ML model are listed in Table 1. All sources were observed by WISE and ROSAT or XMM, and cross-matched in a previous study [7]. We require all AGN to have been observed in SDSS, which limits the study to a quarter of the sky. A subset of 9944 sources were already observed spectroscopically in previous SDSS runs, and have added information on their AGN classification and spectroscopic z. The left panel of Fig. 1 shows the spatial distribution of the training (SPIDERS-AGN) and the compiled catalogue sources.

Observation type	Instrument	Spectral band	Input provided	N_{AGN}[1]	Reference
Soft X-ray	*ROSAT*	0.1-2.4 keV	X-ray flux and err.	19 896	[8]
	XMM-Newton	0.2-12 keV		1468	[9]
Mid-Infrared photometry	WISE	3.4 - 22 µm	W1,W2,W3,W4 mag. and err.	21 364	[10]
Optical photometry	SDSS I-IV	3543-9134 Å	*ugriz* mag. and err.	21 364	[11]
	Gaia	330–1050 nm	Flux and err.		[**?**]
Optical spectroscopy	SDSS I-III	380–920 nm	Classification	9944	[5]
	see Ref.		redshift		[12]

Table 1: Catalogues and their references used to build the multiwavelength inputs to the machine learning algorithm.

3. Pseudo-set method

In both the classification and regression ML tasks, we make use of the measurement uncertainties that are associated with all the photometric observations presented in Table 1. We adopt the method of [13] to create 200 "pseudo-sets" for each single AGN source, based on the smearing of each input value μ_{value} by its measurement error σ_{err}, assuming σ_{err} to be gaussian. The right panel

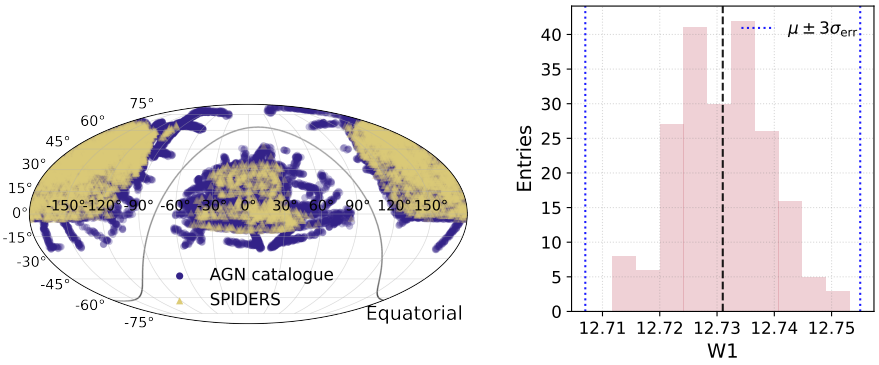

Figure 1: *Left*: Spatial distribution of sources in equatorial Mollweide projection for the for the selected AGN sample (in blue) and the SPIDERS AGN sample (yellow). The requirement for all sources to have been observed by SDSS constrain their distribution to the Northern Sky footprint. The galactic plane is shown as a gray line. *Right*: Distribution of W1 input smeared by the measurement uncertainty for a single source. Each point is drawn from a normal distribution centered at the given catalogue input feature $\mu_{\rm value}$ (black dashed line) and extending to $\pm 3\sigma_{\rm err}$ (blue dotted lines), from the given photometric measurement error.

of Fig. 1 shows such an example for the IR W1 band. Drawing randomly from each independent "smeared" input, we can create N pseudo-sets for each AGN, reconstructing them N times, and thus having a handle on the performance and reconstruction of the ML classification and regression tasks.

4. AGN type classification

Type 2, or obscured AGN, are systems where the emission from the accretion disk gets absorbed by the surrounding dusty torus, suppressing the emission in certain wavelengths [14]. Traditionally, Type 2 AGN are identified spectroscopically looking narrow emission lines. In our study, we use the 9944 sources with a classification to train a random forest (RF) model N times. The ratio of Type 1/ Type 2 AGN in the training sample is highly imbalanced (16:1), since all sources were soft X-ray selected, a band that is highly suppressed in obscured AGN. To help the model identify the minority class, we undersample the majority class to reach a 1:1 training ratio. The left panel of Fig. 2 shows the confusion matrix for the validation set: the RF selects Type 2 AGN with a 93% efficiency. The effect of the pseudo-set method in classification of the unlabeled data is shown in the right panel of Fig. 2: the reconstructed obscuration for each is then the average of all N reconstructions, $\mu_{\rm obscuration}$, with its associated standard deviation value $\sigma_{\rm obscuration}$. We establish a custom decision threshold t for an AGN source to be considered as eiher Type 1 or Type 2. To enhance the purity of the classification, we set t=0.8, such that an AGN source is considered of Type 2 if $\mu_{\rm obscuration} > 0.8$ and of Type 1 if $\mu_{\rm obscuration} < 0.2$. Doing so, we find that 7852 are marked as Type 1, 5228 as Type 2, and 2448 as "ambiguous". This corresponds to a n_1/n_2 ratio of $\sim$ 1.5:1, which is markedly smaller than the 16:1 ratio from the labeled dataset, due to the bias of spectroscopic follow-ups towards brighter optical sources in the training data.

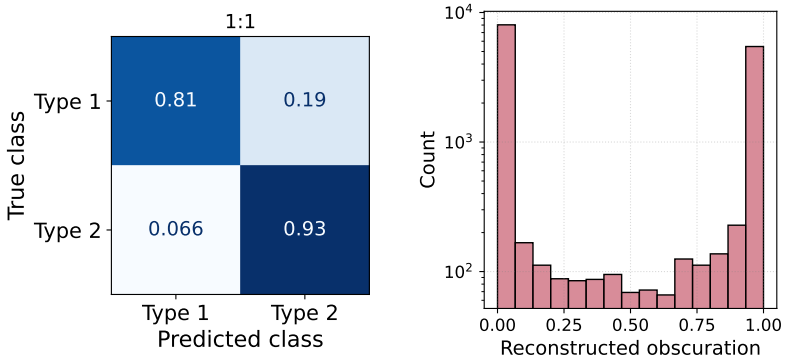

Figure 2: *Left*: Confusion matrix for training ratios 1:1. The undersampled (1:1) classifier is much more apt to identify Type 2 AGN, while still performing well in the identification of Type 1 AGN (91% and 87% efficiency respectively). *Right*: Histogram of the averaged reconstructed obscuration values for all unlabeled data. While the majority of sources have an obscuration value equal to 0 or 1, a non-negligible number of them lie in the region between the two hard values: a hard classifier, softened

5. Regression for AGN properties estimation

We now use a Support Vector Regressor (SVR) model to predict the parameters z, $L_{\rm Bol}$, $M_{\rm BH}$, and $\lambda_{\rm Edd}$, using as inputs the features presented in Table 1. Since redshift measurements are available for almost half of the 21 364 AGN sources (see Sec. 2), but not for the other, we train and test two separate models, which we call $ML_{\rm w/z}$, where z is added as an input, and $ML_{\rm wo/z}$, where z is one of the outputs. Just as it was done for the classification task (see in Sect. 4), we use the N=200 pseudo-sets to propagate both the uncertainties in the photometric measurements in the training and reconstructed datasets, as well as fluctuations of the regressor's reconstruction.

Many ML applications are readily available to use for such a supervised learning task notably through the `scikit-learn` python library [15]. We use a single output, multi-step chain regression, in order for the ML model to learn the correlations between target parameters. In the first pass of the chain regressor, the initial 18 inputs are used to predict the first output, the redshift z. In the next pass, the model takes 18+1 inputs, the extra-one being the predicted z, and outputs the next parameter, $L_{\rm Bol}$, and so on.

5.1 Prediction performance

Fig. 3 summarizes the performance of the SVR for the $ML_{\rm w/z}$ (left column) and $ML_{\rm wo/z}$ (top panel and right column) cases. From these response matrices, we first note that the quality of the reconstruction is higher when the redshift of the source is known and used as an input parameter. In general the performance worsens slightly as we move into the tails of the bin edges, and the data sample to train on become scarce: the reconstructed parameters tends to be overestimated for low values of the true parameter, while they are underestimated for high values. We quote in the following section values taken from the mean pull distributions for all training sources, taking all $\mu_{\rm true} - \mu_{\rm pred}$ values, and fitting a gaussian to derive $\mu_{\rm pull} \pm \sigma_{\rm pull}$.
For $ML_{\rm wo/z}$, the redshift prectiction (top) yields an accuracy $\sigma_{\rm NMAD}$=1.48× median(| Δz- median(Δz) |

$/(1+z_{\rm true})$ of 7.1%, with an outlier rate of 3.48%, performing just as well as the estimation of photometric redshifts using AGN SDE templates [16]. The worsening reconstruction quality for higher z can be explained by the fewer number of sources at such redshifts found in the SPIDERS-AGN training sample.
The bolometric luminosity $L_{\rm Bol}$, being the convolution of multiple wavelength observations presents the first moderate challenge for the model to predict: it however gives reliable reconstructed values, with $\sigma_{\rm pull}$ = 0.12 (0.31) for $ML_{\rm w/z}$ ($ML_{\rm wo/z}$).
When it comes to estimations of $M_{\rm BH}$, the knowledge of z of the source is the most determining factor for the performance. The width of the pull $\sigma_{\rm pull}$ goes from 0.54 to 0.66 in units of $\log(M_\odot)$. However, we see no such discrepancy in the prediction of the Eddington ratio $\lambda_{\rm Edd}$, even though both parameters $L_{\rm Bol}$ and $M_{\rm BH}$ ($\lambda_{\rm Edd} \propto \frac{M_{\rm BH}}{L_{\rm Bol}}$): this is due to the high level of correlation between predicted parameters in the case of $ML_{\rm wo/z}$, leading to a high covariance error, and a decreased total propagated errors.

5.2 Reconstruction of new sources

Once the robustness of the prediction of the target parameters has been evaluated, we reconstruct the ~22 000 sources without spectroscopic information. Just as it was done for the training sample, all AGN were reconstructed N=200 times with the method outlined in Sec. 3. The mean $\mu_{\rm reco}$ and standard deviation $\sigma_{\rm reco}$ from a gaussian fit to the posterior probability distribution are recorded for each source. Encoded in $\sigma_{\rm reco}$ are pointers to the regressor's ability to reconstruct AGN that are further away from the input range, revealing differences between population type. Fig. 4 presents the distributions of the reconstruction uncertainty $\sigma_{\rm reco}$ on the $M_{\rm BH}$ parameter for Type 1 and 2 AGN, with and without known z. We have reconstructed the 5362 sources in our catalogue identified as Type 2 AGN, using the classifier and criteria presented in Sect. 4, although the ML-model was trained with Type 1 AGN only. The regressor is able to reconstruct the $M_{\rm BH}$ Type 2 AGN with known z (purple and blue distributions), but is unable to do so for Type 2 AGN with unknown z, as exemplified by the flat green curve, which is characteristic of reconstructed noise. For these sources, only the redshift z is reconstructed. We take a look at the AGN classified as Type 1, as that population follows the training dataset more closely. The left panel of Fig. 5 presents the AGN number source density over a wide range of redshifts for several bins in bolometric luminosity. Reconstructed Type 1 AGN are shown in full circles, and SPIDERS AGN are represented in open circles for the same luminosity bins. The same trends are observed in the spectroscopically observed and reconstructed samples: the number density of lower-luminosity AGN peaks later in cosmic time than that of more luminous ones. This effect is known as AGN downsizing (see review [17]). The right panel of Fig. 5 shows the black hole masses of these sources, using the same binning in $L_{\rm Bol}$. Not only does the scaling trend of increasing $M_{\rm BH}$ with $L_{\rm Bol}$ remain, but the peaks of the distribution is also coincident between the spectroscopically observed and ML-reconstructed Type 1 AGN samples, a proof that the reconstructed sources match the training ones when binned in multidimensional space.

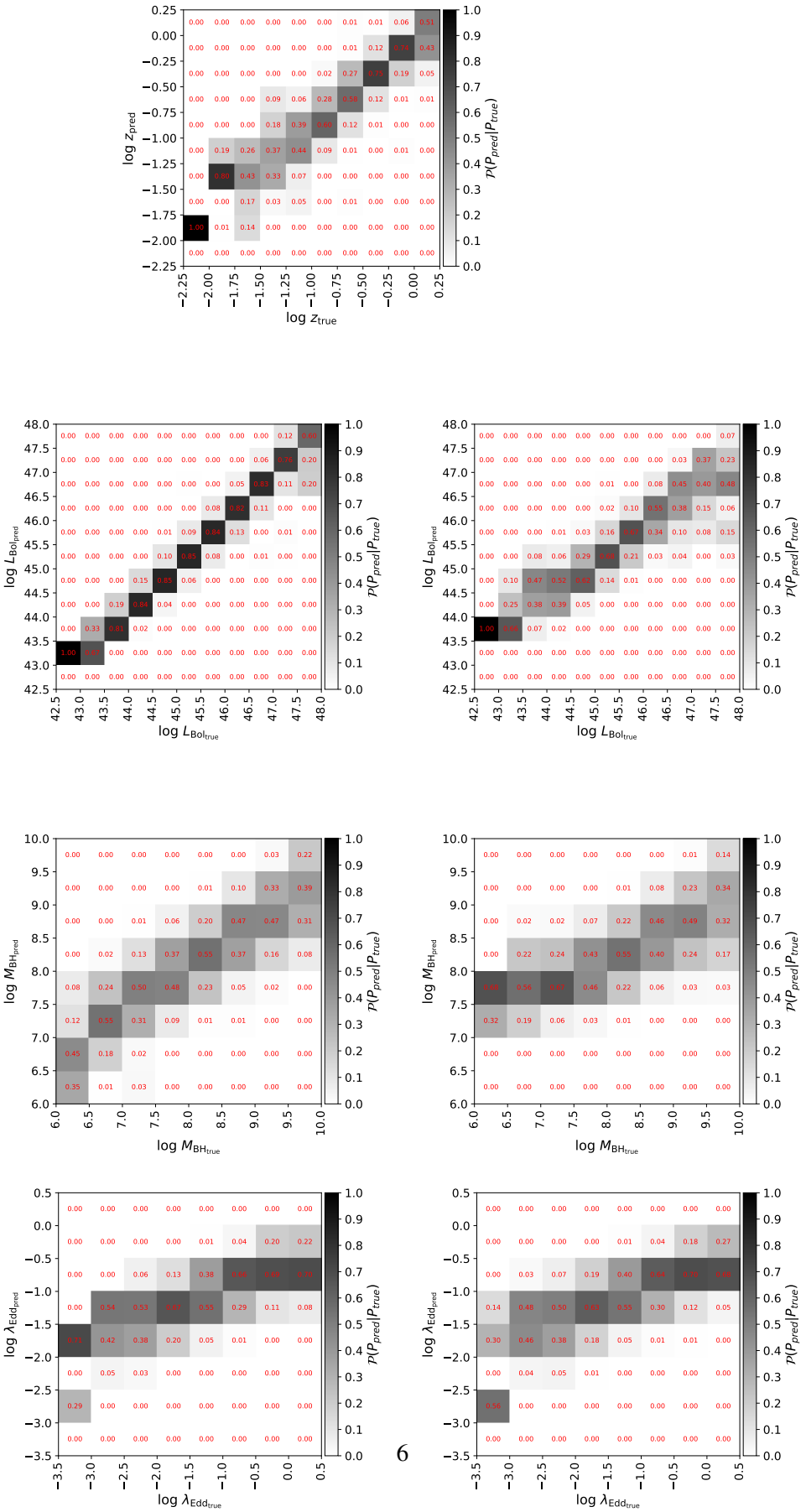

Figure 3: Normalized performance matrices for ML-estimator with known redshift as an input ($ML_{w/z}$, left) and without ($ML_{wo/z}$, right). The true and reconstructed parameters are plotted on the x and y axis, respectively. The error on the reconstruction is used as a weight to the histogram.

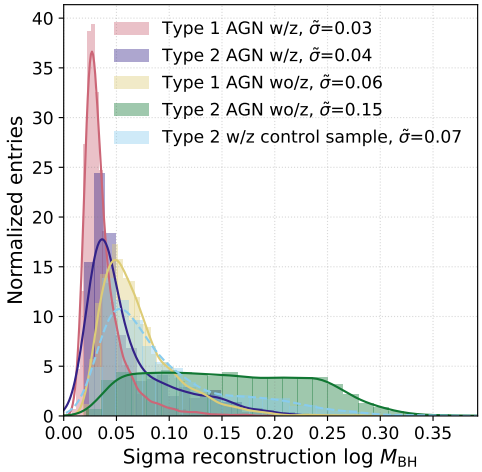

Figure 4: Uncertainty $\sigma_{\rm reco}$ on the log of the reconstructed black hole mass $M_{\rm BH}$ for Type 1 and 2 AGN, with and without z information (estimated with $ML_{\rm w/z}$ and $ML_{\rm wo/z}$ respectively). The $\tilde{\sigma}$ values correspond to the median of their distributions. The mass of the black hole for sources without z identified as Type 2 AGN (in green) cannot be reconstructed, as proven by the flat, structureless uncertainty PDFs.

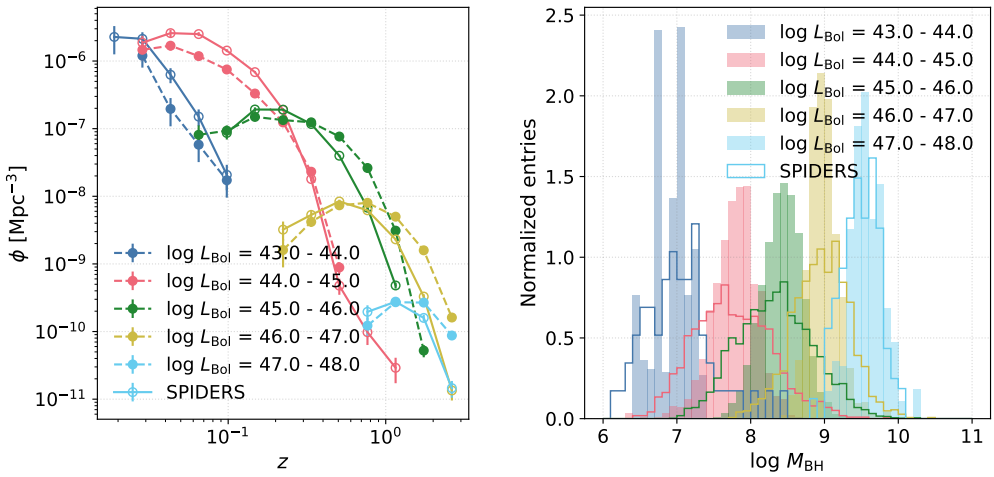

Figure 5: (*Left*) AGN downsizing: comoving number density vs. redshift for Type 1 AGN from this work's catalogue (full circles) and the SPIDERS AGN catalogue (open circles) for different bins of $L_{\rm Bol}$ in units of log(erg s^{-1}). (*Right*) Distribution of $M_{\rm BH}$ for the same bins of bolometric luminosities, for the reconstructed AGN (colored bars) and SPIDERS AGN (colored steps). A flat ΛCDM cosmology with H_0 = 70 km s^{-1} Mpc^{-1}, $\Omega_{\rm M}$ = 0.3, and Ω_Λ = 0.7 is assumed to calculate the comoving number density.

6. Summary and catalogue release

The result of the work presented in this paper has been compiled into a single catalogue available in https://www.zeuthen.desy.de/nuastro/ML_reconstructed_AGN_catalogue/. This includes the 21 364 reconstructed sources, with results from the obscuration classifier and estimation of z, L_X, L_{Bol}, M_{BH}, L_{Edd} and λ_{Edd} with associated reconstruction uncertainties. For 4457 sources, of Type 2 AGN without previous z information, entries for L_X, L_{Bol}, M_{BH}, L_{Edd} and λ_{Edd} are left blank. The release of this new dataset is of particular use for multimessenger astronomy studies, where one needs to know these physical parameters for a large sample of sources while maximizing the sky coverage.

References

[1] **IceCube** Collaboration, R. Abbasi *et al.* *Science* **378** no. 6619, (Nov., 2022) .

[2] K. Murase *Science* **378** no. 6619, (2022) 474–475.

[3] **IceCube Collaboration** Collaboration, R. Abbasi *et al.* *Phys. Rev. D* **106** (Jul, 2022) 022005.

[4] N. Clerc *et al.* *Monthly Notices of the Royal Astronomical Society* (08, 2016) .

[5] T. Dwelly *et al.* *Monthly Notices of the Royal Astronomical Society* **469** no. 1, (July, 2017) 1065–1095.

[6] D. Coffey *et al.* *Astronomy & Astrophysics* **625** (May, 2019) A123.

[7] M. Salvato *et al.* *Monthly Notices of the Royal Astronomical Society* **473** no. 4, (Feb., 2018) 4937–4955.

[8] T. Boller *et al.* *Astronomy & Astrophysics* **588** (Apr., 2016) A103.

[9] R. D. Saxton *et al.* *Astronomy & Astrophysics* **480** no. 2, (Mar., 2008) 611–622.

[10] R. M. Cutri *et al.* *VizieR Online Data Catalog* (Feb., 2021) II/328.

[11] M. R. Blanton *et al.* *The Astronomical Journal* **154** no. 1, (Jun, 2017) 28.

[12] M.-P. Véron-Cetty *et al.* *Astronomy and Astrophysics* **518** (July, 2010) A10.

[13] S. Shy *et al.* *The Astronomical Journal* **164** no. 1, (July, 2022) 6.

[14] R. C. Hickox *et al.* *Annual Review of Astronomy and Astrophysics* **56** no. 1, (2018) 625–671.

[15] F. Pedregosa *et al.* *Journal of Machine Learning Research* **12** (01, 2012) .

[16] B. Luo *et al.* *The Astrophysical Journal Supplement Series* **187** no. 2, (Apr., 2010) 560–580.

[17] W. N. Brandt *et al.* *The Astronomy and Astrophysics Review* **23** no. 1, (Dec., 2015) 1.