Astronomy & Astrophysics manuscript no. agna_paper_aa_2_columns ©ESO 2023
April 3, 2023

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

Estimation of obscuration, redshift, luminosities, black hole mass, and Eddington ratio

Sarah Mechbal[1], Markus Ackermann[1], and Marek Kowalski[1]

DESY, Platanenallee 6, D-15738, Zeuthen, Germany
e-mail: sarah.mechbal@desy.de

Received November 15, 2022; accepted November 16, 2022

ABSTRACT

Context. We investigate the physical nature of Active Galactic Nuclei (AGN) using machine learning (ML) tools.
Aims. We show that the redshift z, the bolometric luminosity $L_{\rm Bol}$, the central mass of the supermassive black hole (SMBH) $M_{\rm BH}$, the Eddington ratio $\lambda_{\rm Edd}$ as well as the AGN class (obscured or unobscured) can be reconstructed through multi-wavelength photometric observations only.
Methods. A Support Vector Regression (SVR) ML-model is trained on 7616 of spectroscopically observed AGN from the SPIDERS-AGN survey, previously cross-matched with soft X-ray observations (from ROSAT or XMM), WISE mid-infrared photometry, and optical photometry from SDSS *ugriz* filters. We build a catalogue of 21 364 AGN to be reconstructed with the trained SVR: for 9944 sources, we found archival redshift measurements. All AGN are classified as either Type 1/2 using a Random Forest (RF) algorithm on a subset of known sources. All known photometric measurement uncertainties are incorporated using a simulation-based approach.
Results. We present the reconstructed catalogue of 21 364 AGN with redshifts ranging from $0 < z < 2.5$. z estimations are made for 11 420 new sources, with an outlier rate within 10%. Type 1/2 AGN can be identified with respective efficiencies of 88% and 93%: the estimated classification of all sources is given in the dataset. $L_{\rm Bol}$, $M_{\rm BH}$, and $\lambda_{\rm Edd}$ values are given for 16 907 new sources with their estimated error. These results have been made publicly available.
Conclusions. The release of this catalogue will advance AGN studies by presenting key parameters of the accretion history of 6 dex in luminosity over a wide range of z. Similar applications of ML techniques using photometric data only will be essential in the future, with large datasets from eROSITA, JSWT and the VRO poised to be released in the next decade.

Key words. AGN – machine learning – regression – classification – obscuration – SMBH – Eddington ratio – catalogue

1. Introduction

Active Galactic Nuclei (AGN) are known to be the most luminous sources in the Universe. These systems consist of a central supermassive black hole (SMBH), around which an accretion disk is formed. Although much is yet to be learned about the feedback mechanisms linking the SMBHs growth and the evolution of their host galaxies, we already know that their mass $M_{\rm BH}$ scales with a number of galaxy properties, like the stellar velocity dispersion σ, bulge mass and luminosity (see review by Kormendy & Ho (2013)). Furthermore, the extreme energetics of these objects also makes them a favoured source of cosmic ray acceleration (Murase & Stecker 2022; Abbasi et al. 2022), as underlined by the recent discovery of neutrinos originating from the AGN NGC 1068 with IceCube (ICECUBE COLLABORATION et al. 2022).

Collecting physical parameters — such as their redshifts z, black hole mass $M_{\rm BH}$, Eddington luminosity and ratio $L_{\rm Edd}$ and $\lambda_{\rm Edd}$ — from a complete and unbiased sample of AGN, becomes a necessary challenge in order to study their accretion history. However, spectroscopic techniques are almost always needed to measure these variables, and although the number of spectroscopically observed AGN has undoubtedly grown in the last decade (Comparat et al. 2020), the discrepancy between photometrically identified AGN and those followed up with spectroscopic surveys remains large. Fortunately, AGN have been well covered by multi-wavelength surveys: X-ray telescopes have observed the extragalactic sky, revealing a pattern of stars, black holes in binary systems and AGN, while IR telescopes have allowed to distinguish the latter from the large stellar population.

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 weak. This limitation can be overcome by searching for correlations between a messenger, e.g., neutrinos or cosmic rays, and a population of AGN instead. However, the power of correlation searches increases with the sky coverage of the counterpart observations and, additionally, requires a model for the expected production of the messenger in question for each object included in the correlation study. Such models usually depend on physical parameters of the AGN.

Machine learning (ML) techniques have been applied in recent years to characterize AGN sources, mainly for classification tasks and redshift determination (Sadeh et al. 2016; Fotopoulou & Paltani 2018; Stevens et al. 2021). In this paper, we report on a novel attempt to employ ML regression tasks to reconstruct fundamental parameters of AGN. We train a ML algorithm to estimate z, L_{x}, L_{Bol}, the soft X-ray (SXR) and bolometric luminosities, as well as M_{BH}, L_{Edd} and λ_{Edd} on 21 364 AGN, all observed in the IR, optical and X-ray bands photometrically but not spectroscopically. To train the model, we use the recent SPIDERS-AGN spectroscopic survey (Clerc et al. 2016; Dwelly et al. 2017), which has compiled and released a sample of ~ 7600 Type 1 AGN (Coffey et al. 2019). We also train a ML classifier to identify Type 2 (or obscured), from Type 1 (unobscured) AGN.

The structure of the paper is as follows: we detail the catalogues used to expand and build both the training and to-be-reconstructed AGN data samples in Sect. 2, and describe the procedures to select AGN from stellar, galactic and blazar populations. We detail in Sect. 3 how the errors on the input parameters are incorporated in order to generate pseudo-sets for the classification, training and reconstruction of AGN, and underline the advantages of the simulation-based method, before classifying the unlabeled sources using ML tools them in Sect. 4). Sect. 5 dives into the details of the main ML regression task built to parametrize the core of AGN: comparison of several models and final results on the spectroscopic parameters predictions are discussed. We present in Sect. 6 the largest to date catalogue of AGN physical properties, stemming from the ML reconstruction of 21 364 sources, including 11 420 new z measurements, and L_{Bol}, M_{BH}, λ_{Edd} values for all. The limits of the Type 2 AGN reconstruction are discussed. We turn to future studies the release of this dataset makes possible in Sect. 7, and discuss the role ML tools will play with the advent of future missions. The catalogue columns are described in Appendix A.

2. Data

The goal of our study is to compile as wide as possible a catalogue — both in number of entries as well as astronomical features — of non-blazar AGN sources in order to find photometric parameters (X-ray, optical and IR magnitudes, etc) that are highly correlated with features relating to the accretion activity, traditionally only accessible with spectroscopic follow-up surveys.

In this paper, we shall call feature any parameter or column from the catalogue: for instance, the W1 or u-band magnitude. Concurrently, we will refer to a catalogue entry as an AGN point source, or row of the catalogue. For matters pertaining to machine learning methods a few terms also need be clearly defined:

- by *training* and *testing* sample, we mean the dataset of 7616 sources built from the SPIDERS catalogue (Coffey et al. 2019), used to train the ML model tasked to learn correlations between photometric and spectroscopic parameters. The performance of the ML model is assessed by comparing the true and predicted values of target parameters (also referred in the ML literature as the *validation* set),
- we call the *reconstructed* or *full* dataset the AGN catalogue we have built for which we make estimations on the target parameters using the previously trained ML model.

Our aim is then to find a compromise between the addition of new features and the completeness of the final catalogue: since any new parameter added to the training sample must also be available in the reconstructed dataset — and null features are not admitted in machine learning regression tasks — entries with null columns would then have to be sacrificed (for a more detailed treatment of null entries see Appendix C). We detail in the following section the multiwavelength observations used to construct our input parameters. Table 1 summarizes all features and catalogues of origin used. The analysis sequence to create the full reconstructed catalogue of AGN is described in Fig. 1.

2.1. The SPIDERS-AGN catalogue

SPIDERS (SPectroscopic IDentification of ERosita Sources) is a completed SDSS-IV (Blanton 2017) 5128.9 deg^2 survey over the SDSS footprint. The AGN sources were originally pre-selected based on the 1RXS and XMMSL1 (Saxton et al. 2008) catalogues, which were then later updated once the 2RXS (Boller et al. 2016) and XMMSL2 [2] were released. The details of the mission targeting and summary are documented in Dwelly et al. (2017) and Comparat et al. (2020), respectively. The spectroscopic data was made available in the 16th SDSS data release (DR16) (Ahumada 2020) as a catalogue of Type 1 AGN containing X-ray fluxes, optical spectral and photometric measurements, black holes estimates and other derived quantities[3]. We refer the reader to Coffey et al. (2019) for a detailed description of the dataset and to Wolf et al. (2020) for a principal component analysis (PCA) of Type 1 AGN properties.

The survey probed the brightest X-ray sources in the sky, at the higher end of the luminosity distribution with $41 < \log_{10}(\mathrm{L_X}/\mathrm{ergs}^{-1}) < 46$ for a mean redshift $\bar{z} = 0.47$. The bolometric luminosity L_{Bol} was also derived from the monochromatic luminosity $L_{3000\text{Å}}$ and $L_{5100\text{Å}}$ using bolometric corrections. Fitting the Hβ and MgII emission lines, the SPIDERS-AGN study has derived M_{BH}, $L_{\mathrm{Edd}} = 1.26 \times 10^{38}(\frac{M_{\mathrm{BH}}}{M_{\odot}})$ erg s^{-1} and $\lambda_{\mathrm{Edd}} = L_{\mathrm{Bol}}/L_{\mathrm{Edd}}$, with $M_{\odot}$ being the solar mass. For 2337 sources, both the Hβ and MgII lines were observed, and two estimates of M_{BH} and derived quantities were provided: in such cases, we select the values with the smallest associated error $\sigma_{\mathrm{M_{BH}}}$, which has a typical value of ~ 0.02 dex. Table 2 presents a list of the key properties found in the SPIDERS-AGN catalogue, with their respective range and median estimate error. Out of 7670 AGN, 7616 have complete spectroscopic information, which we use as the basis of our training sample, that we expand with several other astronomical catalogues.

2.2. X-ray data

X-ray band observations are some of the most effective to identify AGN: emission is believed to come from above the accretion disk, from where photon scatter onto the hot corona gas and emit X-rays via inverse Compton. Although binary systems such as accreting neutron stars and stellar-mass black holes are also X-ray emitters, AGN are in general an order of magnitude more luminous ($L_{\mathrm{X}} > 10^{42}$ erg s^{-1}) (Hickox & Alexander 2018).

The *ROSAT* telescope (Trümper 1982) performed the first all-sky survey (RASS) between 1990 and 1991 in the 0.1-2.4 keV band: two catalogues, one for faint and another for bright sources were back then released (Voges et al. 2000). The data was reprocessed decades later, leading to a second data release, the 2RXS catalogue, comprising ~135 000 sources (Boller et al. 2016). XMMSL2 is the second catalogue of X-ray sources found in slew data taken by the *XMM-Newton* European Photon Imaging Camera pn (EPIC-pn) in 3 bands: 0.2–12 keV (B8), 0.2–2 keV (B7), and 2–12 keV (B6). The B8 band is the most complete and the one of interest. The starting point of our reconstructed catalogue building is the work of Salvato et al. (2018): 106 573 X-ray sources from 2RXS and 17 665 sources from

[2] https://www.cosmos.esa.int/web/xmm-newton/xmmsl2-ug

[3] https://data.sdss.org/datamodel/files/SPIDERS_ANALYSIS/spiders_quasar_bhmass.html

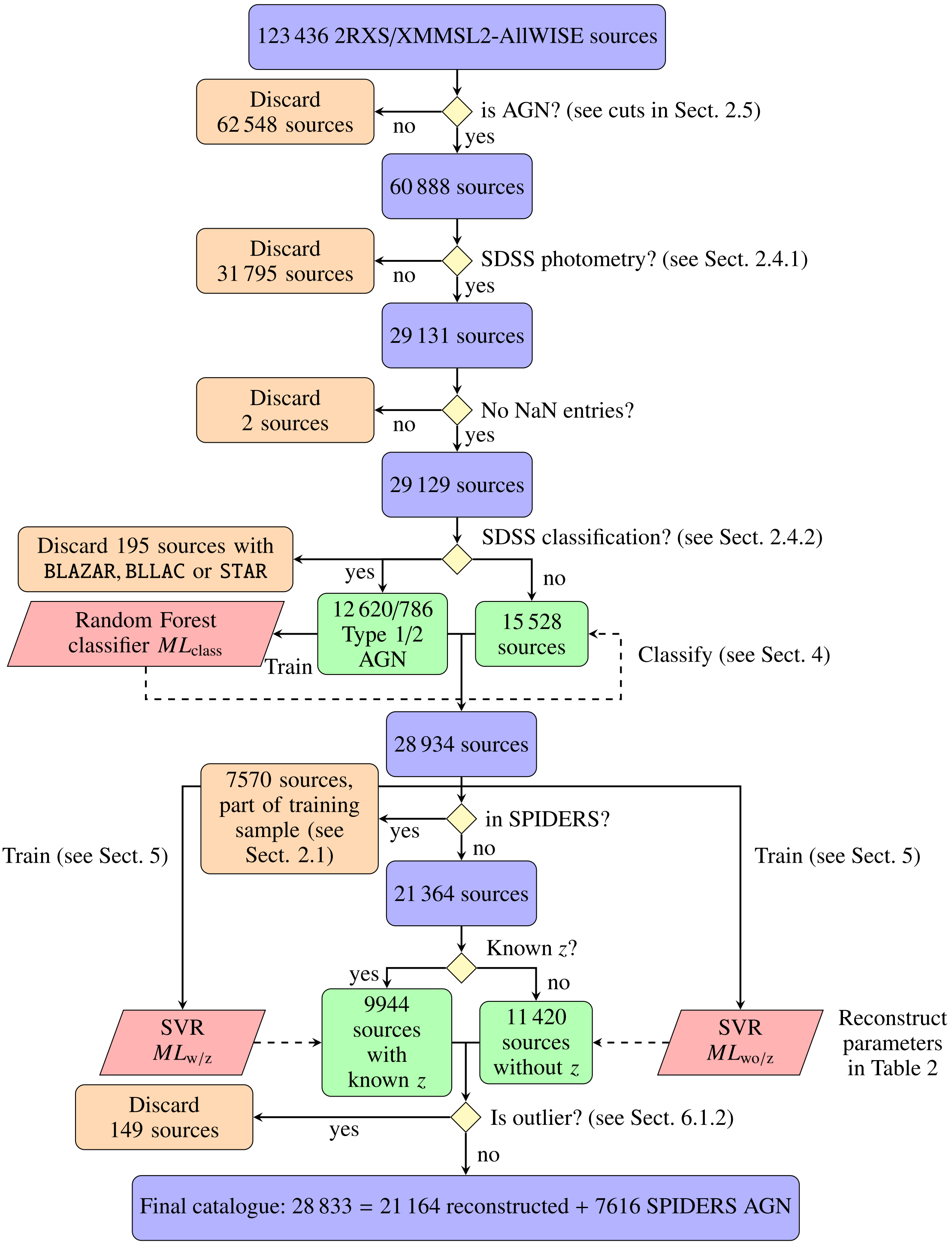

123 436 2RXS/XMMSL2-AllWISE sources
is AGN? (see cuts in Sect. 2.5)
no
yes
Discard
62 548 sources
60 888 sources
SDSS photometry? (see Sect. 2.4.1)
no
yes
Discard
31 795 sources
29 131 sources
No NaN entries?
no
yes
Discard
2 sources
29 129 sources
SDSS classification? (see Sect. 2.4.2)
yes
no
Discard 195 sources with BLAZAR, BLLAC or STAR
12 620/786 Type 1/2 AGN
15 528 sources
Random Forest classifier ML_{class}
Train
Classify (see Sect. 4)
28 934 sources
7570 sources, part of training sample (see Sect. 2.1)
in SPIDERS?
yes
no
Train (see Sect. 5)
Train (see Sect. 5)
21 364 sources
Known z?
yes
no
9944 sources with known z
11 420 sources without z
SVR $ML_{\text{w/z}}$
SVR $ML_{\text{wo/z}}$
Reconstruct parameters in Table 2
Is outlier? (see Sect. 6.1.2)
yes
no
Discard
149 sources
Final catalogue: 28 833 = 21 164 reconstructed + 7616 SPIDERS AGN

Fig. 1. Flowchart of the analysis.The starting point 2RXS/XMMSL2-AllWISE catalogue released in Salvato et al. (2018), leading to the catalogue of 21 364 reconstructed AGN sources presented in this work.

XMMSL2 (with $| b |> 15°$) were cross-matched with their AllWISE (Wright et al. 2010) and *Gaia* (Gaia Collaboration et al. 2018) counterparts using a newly developed Bayesian algorithm to overcome the large positional uncertainties of the X-ray observations. Two catalogues were released: 2RXS-AllWISE and XMMSL2-AllWISE[4].

In order to combine the two X-ray datasets, several steps must be taken, to match the different response functions and detection range of the instruments. We first convert the ROSAT fluxes from the original 0.1-2.4 keV into the classical soft X-ray band 0.5-2 keV (Dwelly et al. 2017):

$$\frac{F[E'_{\text{max}} : E'_{\text{min}}]}{F[E_{\text{max}} : E_{\text{min}}]} = \frac{E'^{2-\Gamma}_{\text{max}} - E'^{2-\Gamma}_{\text{min}}}{E^{2-\Gamma}_{\text{max}} - E^{2-\Gamma}_{\text{min}}}, \tag{1}$$

where $\Gamma = 1.7$ for 2RXS sources.

In Dwelly et al. (2017), $\Gamma = 2.4$ was chosen for XMMSL2 fluxes: however, the 2RXS/XMMSL2 datasets were kept separate. About a thousand sources from the SPIDERS AGN catalogue have been observed with both instruments: we use these as

[4] https://www.mpe.mpg.de/XraySurveys/2RXS_XMMSL2

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	Boller et al. (2016)
	XMM-Newton	0.2-12 keV		1468	Saxton et al. (2008)
Mid-Infrared photometry	WISE	3.4 - 22 µm	W1,W2,W3,W4 mag. and err.	21 364	Cutri et al. (2021)
Optical photometry	SDSS I-IV *Gaia*	3543-9134 Å 330–1050 nm	*ugriz* mag. and err. Flux and err.	21 364	Blanton (2017) Arenou et al. (2017)
Optical spectroscopy	SDSS I-III see Ref.	380–920 nm	Classification redshift	9944	Dwelly et al. (2017) Véron-Cetty & Véron (2010)

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

Target parameter	Notes	Training range	Error on the estimate
z	Redshift	[0.008,2.5]	-
log L_{X}	X-ray luminosity	[40.8,45.9]	~0.05 dex
log L_{bol}	Bolometric luminosity	[42.9,47.6]	~0.02 dex
log $M_{\mathrm{BH}}/M_{\odot}$	Black hole mass	[6.2,10.4]	~0.02 dex
log λ_{Edd}	Eddington ratio	[-3.2,0.54]	~0.01 dex

Table 2. Target variables, their domain range and error estimate from the SPIDERS catalogue (Coffey et al. 2019). The variables are presented in the order of their sequential prediction by the ML algorithm, as executed by the chain regressor.

a control group to match the XMM to the converted RXS fluxes, by varying the Γ power-law index of Eq. 1 in order to match the peaks of the two X-ray flux distributions, as is shown in Fig. 2. Choosing Γ_{XMM}=1.25, 94% of sources present in both datasets have a flux ratio $\frac{\log F_{XMM_{0.5-2.0keV}}}{\log F_{RXS_{0.5-2.0keV}}}$ within 5% of one another. In the converted SRX band, the distribution of X-ray fluxes is contained between $10^{-14} < F_{0.5-2\mathrm{keV}} < 10^{-9}$ erg cm^{-2}s^{-1}. From the X-ray catalogues, we only keep the the X-ray fluxes and corresponding errors in the converted 0.5-2 keV band as as input to the ML-model. Whereas the hardness ratio and/or column density would have given great information about the class of AGN, both being a known proxy for the obscuration level of accretion disks, this parameter was neither complete, nor very accurate in the case of ROSAT observations, such that it had to be dropped.

2.3. Infrared observation

AGN are bright in the mid-infrared (MIR, 3-30 µm). The dusty torus is responsible for this thermal emission, as it absorbs shorter-wavelength photons from the accretion disk and re-emits them in the MIR. Although star-forming galaxies are also bright in this band, their SED is cooler and can be distinguished from those of AGN (Padovani et al. 2017). The Wide-field Infrared Survey Explorer (WISE) (Wright et al. 2010) is a satellite launched in 2009. The missions was then later extended under a new appellation, NEOWISE (Mainzer et al. 2011). The combination of WISE and NEOWISE data was made available to the public with the release of the AllWISE catalogue (Cutri et al. 2021). The WISE survey scanned the sky at 3.4, 4.6, 12 and 22 µm (the bands designated as W1, W2, W3, and W4, respectively), at a depth at which the majority of the resolved 2RXS and XMMSL2 populations are to be detected (Salvato et al. 2018). In addition to the 4 MIR magnitudes and their associated errors, we explicitly record the relative magnitudes W1-W2, W2-W3, W3-W4. These values are readily available from Salvato et al. (2018), as previously mentioned.

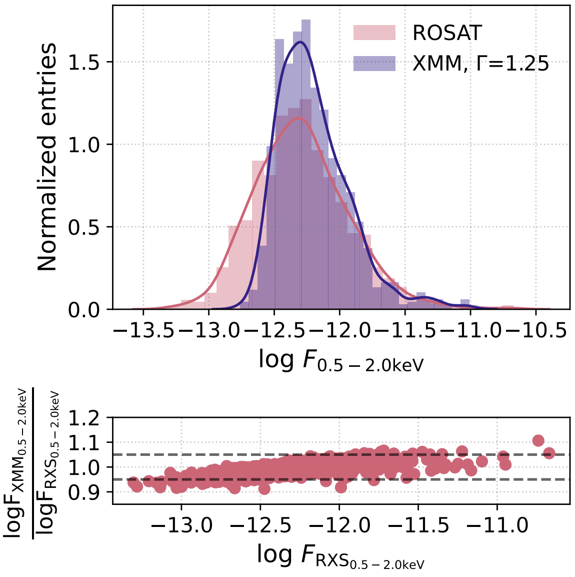


Fig. 2. *Top*: 2RXS and XMMSL2 fluxes for the ~ 1000 SPIDERS-AGN sources observed with both instruments. The X-ray fluxes were converted to the soft X-ray band 0.5-2.0 keV using Γ=1.7 for 2RXS and Γ=1.25 for XMMSL2, value chosen to match the peaks of the two distributions. *Bottom*: Ratio of the converted flux logs as a function of the 2RXS fluxes for the same sources shown in the above panel.The dashed lines represent the ± 5% level on the ratio, within which 94% of the converted fluxes are.

2.4. Optical data

2.4.1. Photometry

SDSS The Sloan Digital Sky Survey (SDSS) has in the course of its runs observed over 700 000 quasars in the optical band, most of them in broadband photometry in the instrument specific *ugriz* (3543-9134 Å) filters (Lyke et al. 2020). To pair the AGN sources from our training and unknown sets with their SDSS observations, the `astroquery` software tool (Ginsburg et al. 2019) is used: we cross-match the best AllWISE counterpart to the X-ray sources in our training and reconstructed samples with an optical counterpart from the DR17 photometric catalogue, setting a maximum radius of 5 arcsec. For the matched sources, we add as features the SDSS PSF magnitudes `psfMag` and their associated error `psfMagErr` for the 5 *ugriz* bands. These values are most appropriate when studying the photometry of distant quasars. While logically, all 7616 SPIDERS sources have SDSS photometry counterparts, 47 739 (5985) 2RXS (XMMSL2) sources have been observed photometrically, which we add as a requirement (see Fig. 1).

Gaia Salvato et al. (2018) also cross-matched the 2RXS/XMMSL2 sources to the first release catalogue of the *Gaia* mission (Arenou et al. 2017). The astrometric instrument performs broadband photometry in Gaia's white-light G-band (330–1050 nm): we keep the mean flux and mean flux error for all sources, expressed in photo-electrons s^{-1}.

2.4.2. Spectroscopy and redshift

Prior to the start of the SPIDERS mission, X-ray+AllWISE AGN targets were cross-matched with the already observed SDSS I-II-III runs (Dwelly et al. 2017): ∼12 000 ROSAT and ∼1500 XMM–Newton sources were found to already have been observed with spectroscopy. After performing a visual inspection of the optical spectra, two Value Added Catalogues (VAC) were released [5]. These included redshift measurements, object main and sub-classification. Matching sources from the reconstructed catalogue by their X-ray name, we find 21 287 (2540) 2RXS (XMMSL2) sources previously observed spectroscopically: of these, 11 242 (1250) contain redshift information (but no black hole mass/Eddington ratio). As we shall see in Sect.5.3, this latter variable greatly improves the ML predictions. Furthermore, we will make use of the sub-sample of sources for which we have an AGN classification to correlate multiwavelength observations with the AGN obscuration level (see Sect. 4). Following the classification scheme of Comparat et al. (2020), we mark AGN as either Type 1[6] or 2[7].

Additional spectroscopic classification and redshift measurements are found in the VERONCAT catalogue (Véron-Cetty & Véron 2010), a collection of some ∼ 150 000 quasars from multiple surveys. The X-ray positions of sources are then cross-matched with the optical or radio positions given by VERONCAT, using a maximum matching radius of 60 arcsec, a value taken from a past study (Abbasi et al. 2022). This allows us to get additional AGN classification and spectral class features. We collect some ∼ 9000 redshifts from VERONCAT, on top of the ones already found from previous SDSS surveys. To verify the accuracy of the cross-matching, we compare the redshift entries from the 6163 SPIDERS sources that are already present in VERONCAT, $\Delta_z = | z_{\rm SPIDERS} - z_{\rm VERONCAT} |$. We found that 98% of the sources have $\Delta_z < 0.01$, comforting us in the adequateness of the cross-matching radius used.

2.5. AGN selection

The multiwavelength data collected in the previous sections can now be used to select AGN from a larger sample comprising of blazars, galaxies, and stars using X-ray and IR colors observations .

Following the source characterization methods already established in Salvato et al. (2018) and references therein, we proceed to a first selection of AGN in the X-ray/MIR plane (see top panel of Fig.3). An empirical relationship has been found, which separates AGN from stars and galaxies.

$$W1 \geq 1.625 \times \log F_{0.5-2\rm keV} - 8.8 \qquad (2)$$

We can confirm the validity of such a selection by overlaying the confirmed AGN in the SPIDERS sample, which all clearly lie above the cut-off line.

We then use the AllWISE W1-W2, W2-W3, and W3-W4 relative magnitudes to isolate AGN from blazars, starbust and normal galaxies, as was developed in Assef et al. (2013): in the W1-W2 vs W2-W3 digram (middle panel of Fig. 3), stars and elliptical galaxies have colors near zero, located in the lower left quadrant, spiral galaxies are red in W2–W3 but not in W1-W2, while Ultra-Luminous InfraRed Galaxies (ULIRGS) are red in both colors, lying in the upper right quadrant of the diagram (Wright et al. 2010). We select the sources for which $1.5 < W2 - W3 < 4.5$ and $0.2 < W1 - W2 < 1.75$ (black square in middle panel of Fig. 3). Once again, we use the SPIDERS AGN sample to justify the selection criteria being made in the W1-W2 vs W2-W3 color-color space. Similarly, a visual cut is made on the W1-W2 vs W3-W4 plane following the SPIDERS-AGN locus (black line of bottom panel in Fig. 3) (Abbasi et al. 2022).

After these selections, 60 952 AGN are identified, from an original sample of 123 436 X-ray-AllWISE sources. The spatial distribution of the final catalogue is shown in Fig. 4: the sources follow the SDSS footprint, as the requirement to have been observed photometrically by SDSS marks the most stringent cut on the data.

3. Measurement uncertainties and pseudo-sets

Each photometric observation used as an input, presented in Table 1, comes with a measurement uncertainty of non-constant variance (also called "heteroscedastic" error). Properly taking them into account is an active area of study in astrostatistics (Feigelson et al. 2021). Considering the stochastic nature of both input features and ML models, we adopt the approach outlined in Shy et al. (2022): all measurement errors $\sigma_{\rm err}$ are assumed to be gaussian, such that any photometric input for a single source is represented as a normal distribution, centered around the given value $\mu_{\rm value}$, extending to $\pm 3\sigma_{\rm err}$. Fig. 5 shows such an example of the W1 for a single training source with the measurement given by $\mu_{\rm value}$ (black dashed line), and $\mu_{\rm value} \pm 3\sigma_{\rm err}$ (blue dotted lines).

Drawing randomly from each independent "smeared" input distributions, we can thus create N pseudo-sets for each AGN source, where the photometric inputs differ within $\pm\sigma_{\rm err}$ between each realization. We create N=200 pseudo-sets of both the training (for the regression) and the reconstructed datasets. Sect. 4 and 5 will develop how this simulation-based treatment of measurement errors helps characterize both the performance and reconstruction of unlabeled/unknown data in the context of ML classification and regression tasks.

4. Machine learning classifier for Type 2 AGN identification

Once AGN have been identified, a further step is needed: since our training sample exclusively contains Type 1 AGN (broad emission line, unobscured), we must distinguish Type 1 from Type 2 (narrow emission line, obscured) in our reconstructed sample, to study any potential biases in the spectroscopic predictions.

Obscured AGN, also called Type 2, are systems where the emission from the accretion disk gets absorbed and scattered by dust or gas surrounding it, masking some of the characteristic

[5] https://data.sdss.org/datamodel/files/SPIDERS_ANALYSIS

[6] CLASS_BEST=="BALQSO", "QSO_BAL, "QSO", "BLAGN"

[7] CLASS_BEST=="NLAGN", "GALAXY"

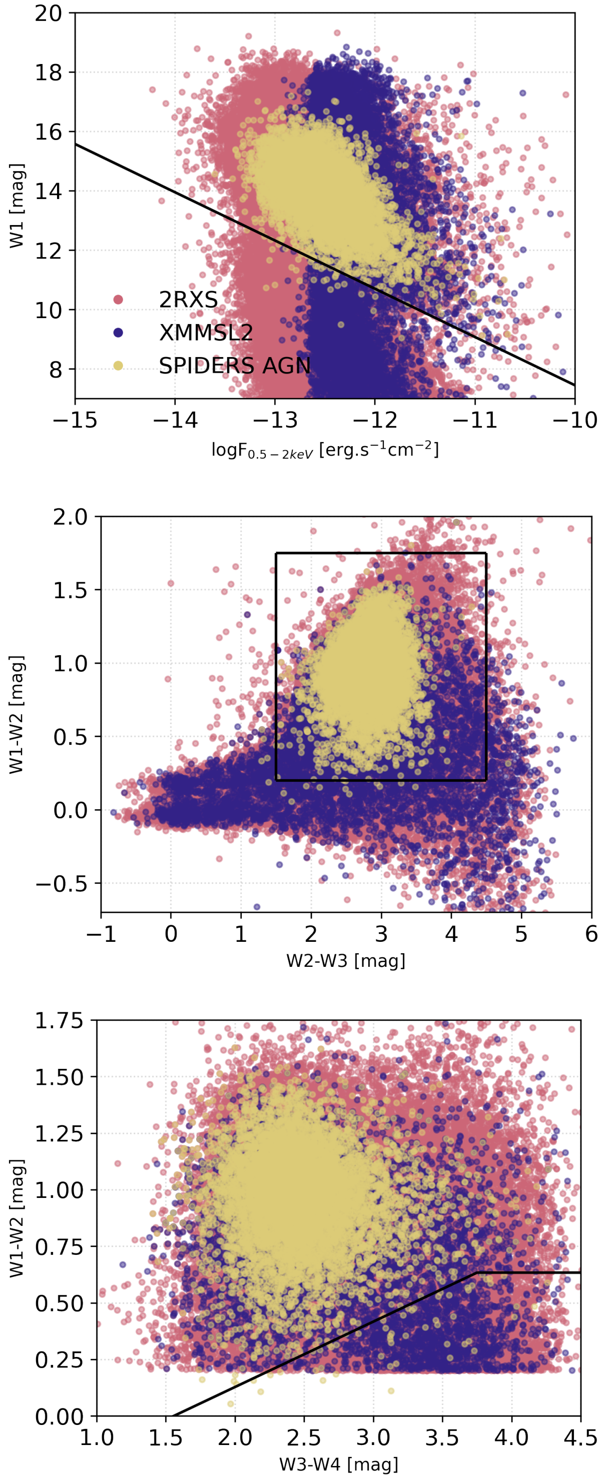


Fig. 3. *Top*: Distribution of sources in the W1 band vs soft X-ray flux parameter space for the ALLWISE counterparts to 2RXS (pink) and XMMSL2 (blue). The confirmed SPIDERS AGN are represented in yellow. The cut defined in Eq. 2 is also shown. W1–W2 magnitude plotted against the W2-W3 (*middle*) and W3-W4 (*bottom*). The black lines show the cuts applied based on the SPIDERS AGN position.

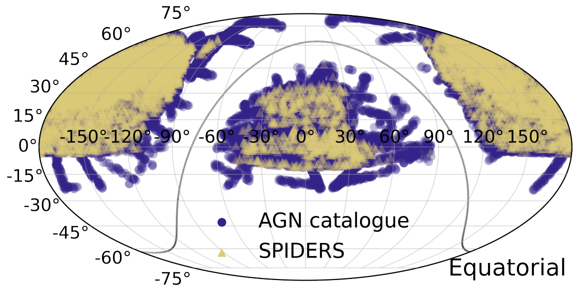


Fig. 4. Spatial distribution of sources in equatorial Mollweide projection for the for the selected AGN sample (in blue) and the SPIDERS AGN sample (yellow). The requirement for all sources to have been observed by SDSS constrain their distribution to the Northern Sky footprint. The galactic plane is shown as a gray line.

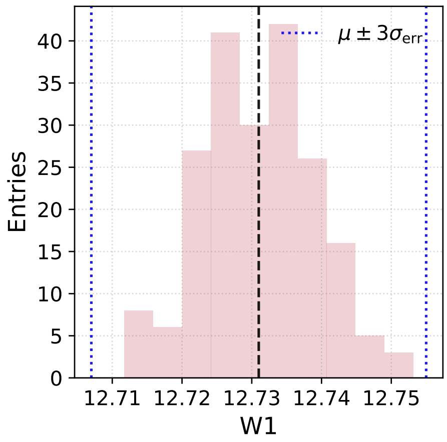


Fig. 5. Distribution of W1 input smeared by the measurement uncertainty for a single source. Each point is drawn from a normal distribution centered at the given catalogue input feature $\mu_{\rm value}$ (black dashed line) and extending to $\pm 3\sigma_{\rm err}$ (blue dotted lines), from the given photometric measurement error.

signature of the AGN with respect to the line of sight of the observer. The impact of such a suppression is wavelength-dependent, and a reliable and complete identification method of this population remains challenging and important (for a review on obscured AGN see Hickox & Alexander (2018)). Distinction between Type 1 and Type 2 AGN can be done across multiple photometric and spectroscopic observation types. The most classical way to identify AGN class is through UV-NIR spectroscopy. Type 1 AGN, have broad emission lines showing velocity dispersion >1000 km s^{-1}, while Type 2 AGN have narrow emission lines only, with velocity dispersion < 1000 km s^{-1}. However, since the purpose of this study is to characterize AGN that have not been spectroscopically observed, we must circumvent the absence of such of information.

Our sources were originally selected based on their soft X-ray flux (Salvato et al. 2011): this already skews the sample towards a majority of Type 1 sources, as soft X-rays get absorbed by the high hydrogen column density $N_{\rm H}$ around the accretion disk (Hasinger 2008), while harder X-ray are less suppressed in obscured AGN. These are also known to have stronger emission in the MIR than they do in other bands, as larger dust column reprocesses radiation from other bands. We thus expect weak UV/optical/NIR emission compared to that of the

MIR. As described in Sect. 2.4.2, we already know the AGN class for a sub-sample of sources which have a classification CLASS_BEST, such that the correlations between photometric observations and the object type can be studied. One could try to identify a single feature that best allows the distinction between obscured/unobscured AGN, the ratio of W2/W1 infrared emission for example. We develop this "classical" method in Appendix B following the method developed in Abbasi et al. (2022). However, a more judicious use of the multiwavelength information collected would be to train a classification machine learning model. Taking the 13 415 sources for which AGN type is known as a training sample, we add a new feature called "obscuration": its value is 0 for Type 1 AGN, and 1 for Type 2 AGN. We seek to characterize whether the remaining 15 533 are obscured or not.

4.1. Imbalanced classification

Of the 13 415 labeled sources, 12 629 are unobscured — including SPIDERS sources which are all Type 1 — while 786 are obscured AGN, a ratio of 16:1. This classification task is thus an imbalanced one, a frequent situation where a classifier must learn to identify a minority case, although it is trained on a dataset over-represented by a majority case, leading to bias in the reconstructed sample. While many strategies exist to mitigate this issue, such as oversampling by synthetically creating minority cases (Chawla et al. 2002), a simpler one is to undersample the majority cased by randomly selecting a sub-sample of Type 1 AGN: this way the n_1/n_2 ratio is changed and brought closer to parity.

To test the approach, we train a random forest (RF) for different n_1/n_2 ratios, with 18 features as inputs, the smeared photometric measurements (see Sect. 3). We choose to use 50% of the available dataset for training, and 50% for validation, by comparing the predicted from the true values. We call a true positive (TP) a Type 2 AGN classified as Type 2, a false positive (FP) a Type 1 classified as Type 2, a true negative (TN) a Type 1 classified as Type 1, and a false negative (FN) a Type 2 classified as Type 1. The precision of the classification is then defined as

$$P = \frac{TP}{TP + FP}, \tag{3}$$

that is, the ability of the classifier not to label as positive a sample that is negative. The recall, also called sensitivity or true positive rate (TPR), is calculated with:

$$R = \frac{TP}{TP + FN}, \tag{4}$$

that is, the ability of the classifier to find all the positive samples. For completion, we provide the definition of the false positive rate (FPR), used in the ROC (Receiving Operating Characteristic) curve:

$$FPR = \frac{FP}{TN + FP}, \tag{5}$$

such that it is a measure of finding all the negative samples. In instances of classification task on imbalanced datasets, the precision-recall curve (PRC) is more informative than the more widely used ROC (Receiving Operating Characteristic), as is detailed in Saito & Rehmsmeier (2015).

Fig. 6 shows the PRC for several models varying class ratios in the training, with the task of classifying AGN sources as Type 1 or Type 2, on a single pseudo-set. The confusion matrices for the 1:1 and 16:1 training ratios are also presented in Fig. 6. We reach the following conclusions:

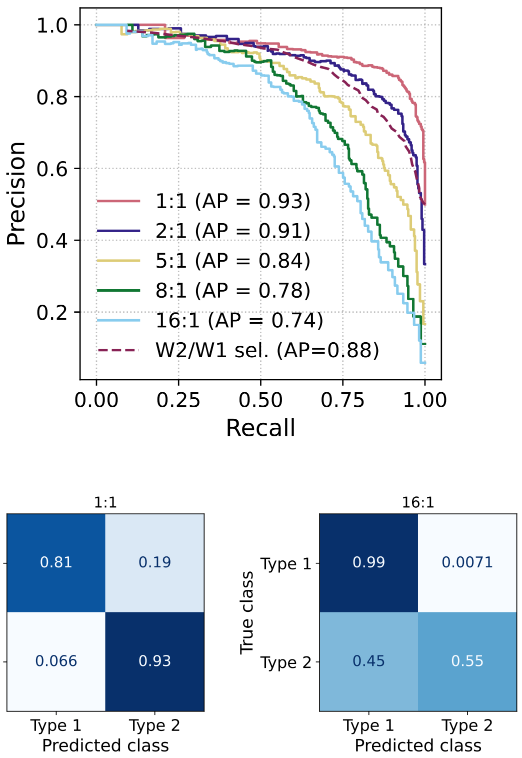


Fig. 6. *Top*: Precision recall curve for different training ratios and single feature selection: a perfect classifier would lie on the top right corner of the PRC. *Bottom*: Confusion matrix for training ratios 1:1 (*left*) and 16:1 (*right*). The undersampled (1:1) classifier is much more apt to identify Type 2 AGN, while still performing well in the identification of Type 1 AGN (91% and 87% efficiency respectively). The naturally imbalanced set, while accurately selecting Type 1 AGN 99% of the time, performs poorly in finding the rarer Type 2 AGN.

1. At its optimized best performance, the ML-model is more apt to identify Type 2 from Type 1 sources than a single feature selection would (solid pink line in Fig.6) vs purple dashed line), with a selection efficiency of 91% for the ML-model, compared to 79% for the single feature selection.
2. The balance of minority/majority cases impacts the performance of the network substantially: at its worst (for a 16:1 ratio), the single feature selection scores better than the ML-model (purple solid line verse dashed line) in selecting Type 2 AGN.

4.2. Classification accuracy on the labeled set

After the study on a single pseudo-set, we settle on training a random forest classifier with a 1:1 training ratio. To propagate the measurement uncertainties in the input features and the random fluctuations inherent to a ML classifier, we make use of the N=200 pseudo-sets previously generated (see Sect. 3) to label the unknown AGN. We follow the steps outlined in Shy et al. (2022): for each simulation set, a classifier is fit to a realization of the labeled data, then used to reconstruct a realization of the unlabeled data. This way, all sources are reconstructed N times, whether they belong to the unlabeled or the validation datasets. For all performance metrics defined in Sect. 4.1 we thus obtain a posterior predictive distribution comprising of the results of each set's classification, such as the precision of Type 2 classification distribution shown Fig.

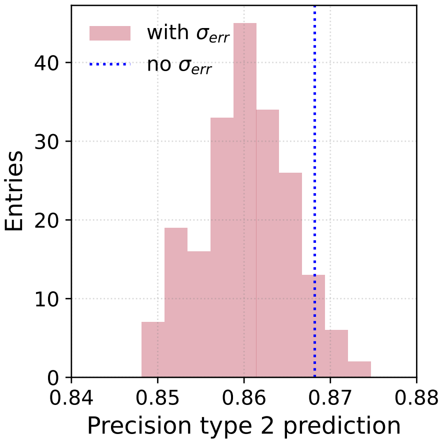


Fig. 7. Posterior distribution for the precision of Type 2 prediction from fitting the labeled dataset N=200 times. The vertical blue dotted line indicates the value obtained from a single RF reconstruction without inclusion of measurement errors, in which case the performance of the classifier is overestimated.

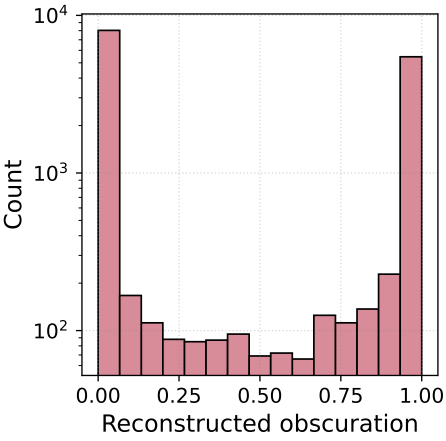


Fig. 8. Histogram of the averaged reconstructed obscuration values for all unlabeled data. While the majority of sources have an obscuration value equal to 0 or 1, a non-negligible number of them lie in the region between the two hard values: a hard classifier, softened.

AGN Type	Precision	Recall
Type 1	0.918 ± 0.007	0.851 ± 0.007
Type 2	0.861 ± 0.006	0.923 ± 0.007

Table 3. Precision and recall scores and uncertainties for Type 1 and Type 2 prediction using N=200 fits to pseudo-sets with measurement uncertainties.

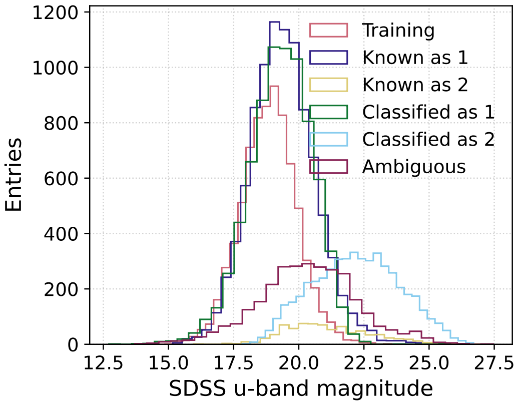


Fig. 9. SDSS u-band magnitude for all labeled and reconstructed datasets. The unlabeled sources classified as Type 2 AGN are fainter than the labeled Type 2 sources, which have made the photometric limit criteria for SDSS spectroscopic observations. In general, obscured AGN have a fainter optical spectra than unobscured ones.

7. The variation across multiple fits reflects the propagated uncertainty through all steps of the procedure. The blue dotted line indicates the precision obtained running the classifier a single time without taking into account measurement errors. For all performance metrics, ignoring uncertainties leads to an overestimation of the predictive power of a classifier (Shy et al. 2022). The final scores with uncertainties on the AGN classifier can be found in Table 3.

4.3. Softening a hard classifier

In addition to giving a more accurate view of the classifier's performance thanks to the validation set, this simulation-based method introduces nuances into the reconstruction of the unlabeled obscuration level. With each source now being reconstructed N=200 times, the reconstructed obscuration is then the relative probability of a source to be in each class. That is, while a single classifier would give a "hard" 0 (Type 1) or 1 (Type 2) prediction on unlabeled data, taking the average value of N reconstructions leads to a "softening" of the "hard" classifier: each source has now an obscuration level $\mu_{\text{obscuration}}$ between 0 and 1, which is the arithmetic mean of the N classification results, with its associated standard deviation value $\sigma_{\text{obscuration}}$. Fig. 8 shows how the unlabeled data now lies in a continuous spectrum between 0 and 1.

This softening of the classification also let us establish a custom decision threshold t on $\mu_{\text{obscuration}}$ and $\sigma_{\text{obscuration}}$ for an AGN source to be considered as eiher Type 1 or Type 2. This threshold can be set to be more or less stringent. Choosing to enhance the purity of the classification, we set t=0.8, such that an AGN source is considered of Type 2 if $\mu_{\text{obscuration}} > 0.8$ and of Type 1 if $\mu_{\text{obscuration}} < 0.2$.

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 ∼ 1.5:1, which is markedly smaller than the 16:1 ratio from the labeled dataset. The reason for such a stark discrepancy can be found in Fig. 9, which shows how the labeled dataset ("known as 1", "known as 2") is biased towards optically brighter (u-band mag < 24). This follows from the target requirements established by the various SDSS surveys prior to the spectroscopic observations of the AGN targets (Alam et al. 2015). The AGN classified as Type 2 (blue histogram) constitute the fainter end of our catalogue, too faint to have been spectroscopically followed-up. Because the RF classifier infers that fainter sources are more likely to be of Type 2, the reconstructed unlabeled catalogue naturally results in a more balanced AGN ratio.

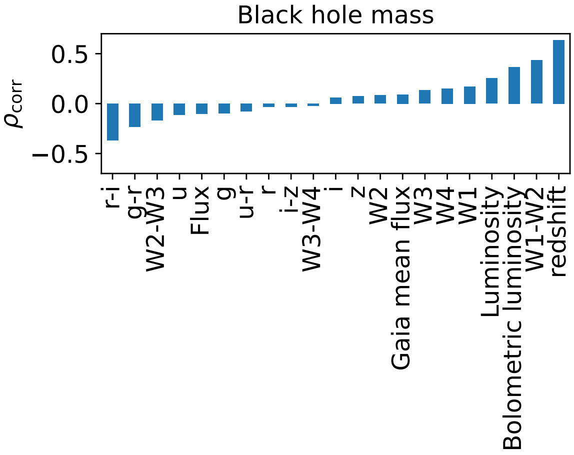


Fig. 10. Bar chart showing the Pearson correlation score of input variables and the black hole mass, from the training sample data. The redshift, luminosity, and bolometric luminosity correlations are included in this chart, since z (and thus $L_{\rm X}$) are known for almost half of the sources, and the outputs will be predicted before the black hole mass, underlining the logic behind the chain regression.

5. Machine-learning for AGN properties estimations

The following section dives into the detail of selecting a suitable machine learning model to predict the parameters of Table 2, using as inputs the features presented in Table 1. Since redshift measurements are available for almost half of the 21 364 AGN sources, 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 of the regressor.

Just as is was done in Sect. 4 for ML classification, we develop how measurement errors are taken into account in ML regression using the pseudo-sets generated, giving in the process a more complete picture of the performance and quality of the reconstruction.

For the training, we transform our target parameters: z, $L_{\rm X}$, $L_{\rm Bol}$, $M_{\rm BH}$, and $\lambda_{\rm Edd}$ are often expressed in log scale to representatively describe the span of values across several decades.

5.1. Exploratory data analysis

Before choosing and training a machine-learning algorithm on the SPIDERS sample, we explore the relationship between the input variables and the target parameters. Fig 10 shows the sorted Pearson's correlation coefficients for all inputs and one of the outputs, the black hole mass of the AGN. The relative infrared and optical color magnitudes demonstrate the highest level correlation. This is even more visually evident when one looks, once more, at the IR color-color plot in Fig. 11. The scatter plot presents the W1-W2 vs W2-W3 AllWISE colors for AGN with $\lambda_{\rm Edd}$ < 0.1 and $\lambda_{\rm Edd}$ > 0.1, the median value of $\lambda_{\rm Edd}$ in the training sample. We observe that strong accretion disks (higher $\lambda_{\rm Edd}$) are redder in both W1-W2 and W2-W3 than lower $\lambda_{\rm Edd}$ values. These clear connections between infrared photometry and spectroscopic observables, already accessible with a naive and straightforward data analysis, are encouraging indications that our goal — the estimation of AGN physical properties — is suited for a machine-learning task.

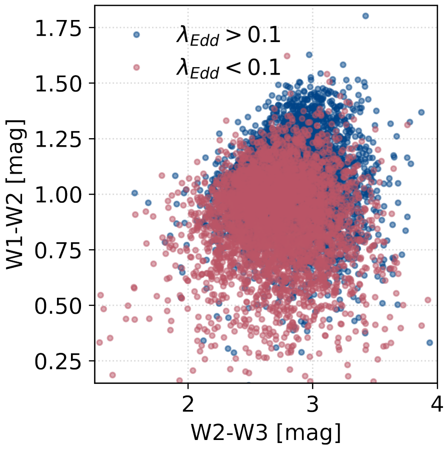


Fig. 11. W1-W2 magnitudes as a function of W2-W3 magnitudes for sources in the training sample with low (red dots) and high (blue dots) $\lambda_{\rm Edd}$. The low and high samples are separated by the median value of the $\lambda_{\rm Edd}$ distribution, 0.1.

5.2. Machine-learning model parameters

Ours is essentially a multi-dimensional linear regression task, with 18 multi-wavelength inputs, listed in Table 1, and 5 or 6 target parameters, depending on whether z is known for a source (see Table 2). Many ML applications are readily available to use for such a supervised learning task notably through the `scikit − learn` python library (Pedregosa et al. 2011). 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, $L_{\rm X}$, and so on.

5.2.1. Selection of ML-model

We detail in this section our non-exhaustive search for the most suited estimator. All supervised learning ML models essentially learn a mapping of inputs to outputs given an example of such a map. There exists however a plethora of model types one could choose from. For instance, linear models expect the output to be a linear combination of the features: certain regressors simplify the model by introducing penalty coefficients that will minimize (in the case of ridge regression) or reduce (for Lasso regression) the input parameters, if some are found to contribute less to the learning. This procedure, called regularization, aims to reduce the overall error in the validation dataset. On the other hand, Support Vector Regression (SVR), an application of the kernel-based Support Vector Machines (Cortes & Vapnik 1995), allows to tune the tolerance ϵ to such errors, while introducing non-linearity parametrization through hyperplane fits to the data. Non-linearity is also a feature of neural networks, e.g, in a multilayer perceptron (MLP), via an activation function connecting the neural layers. We compare these different ML models in the next section.

5.2.2. Model evaluation

To determine the best model, performance metrics are defined based on the validation dataset, where $y_{\rm true}$ and $y_{\rm reco}$ are known.

Model features	Type	Properties	Notes
Model type	Support Vector Regression	Kernel function="rbf", C=1, ϵ=0.001	Training a chain regressor connects the non-independent target parameter to one another
Data scaling	Max-min normalization	Inputs and outputs are scaled	Aids the model to learn the problem
Validation method	K-fold cross-validation	k=10 with data shuffle	Ensures the model gets trained on every single data point

Table 4. Properties of the final SVR ML-model properties chosen to be trained on.

To reduce the variance of these metrics, we use the popular K-fold cross validation method (Stone 1974). It offers an alternative to a standard split of the training sample into train/test datasets, by dividing the dataset into k randomly shuffled groups, and using every single one of them as test set at least once, while training on the $k-1$ groups left. This way, it insures a non-biased evaluation of the model, as each sample (in our case, each AGN source), will be used as a validation point once, and as a training set $k-1$ times: the following metrics are thus calculated for a sample size $N = 7616$. As is common for regression problems, we use the R^2 score, also called coefficient of determination, for each parameter to assess the performance of each model. This coefficient is calculated as:

$$R^2 = 1 - \frac{\sum_i (y_{i_{\mathrm{true}}} - y_{i_{\mathrm{reco}}})^2}{\sum_i (y_{i_{\mathrm{true}}} - \overline{y_{\mathrm{true}}})^2}, \tag{6}$$

with the numerator being the residual sum of squares, and the denominator the total sum of squares. For a perfect regressor, we have $R^2 = 1$. Fig. 12 presents the target-by-target comparison between the two linear models (Ridge and Lasso Regression), a Support Vector Regression model, and a multi-layered perceptron (MLP) deep neural network model[8]. We stress that for this test, none of the model parameters have been tuned, except for the SVR. Although the SVR slightly underperforms in predicting the first target parameters, namely the soft X-ray luminosity L_{X}, as is given by the lower R^2 compared to the other models, the model is markedly better at predicting M_{BH}, L_{Edd} and λ_{Edd}. This trend is also confirmed in the more difficult case of unknown z, represented by open circles in Fig. 12. From a pragmatic aspect, the runtime speed and low number of tuning parameters of the SVR regressor were clear advantages compared to the vast phase space of MLP neural networks.

Once the performance of the SVR model has been assessed, we complete a grid search over its hyperparameters C and ϵ, the regularization and margin of errors, respectively, to find their optimal values. We repeat this procedure for the no-redshift case, $ML_{\mathrm{wo/z}}$. This step is important as the the final result can quite vary between default and optimized parameters. We summarize the final parameters of the machine learning model to be trained in Table 4.

5.2.3. Regression metrics on N pseudo-sets

Just as it was done for the classification task, 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. Here again, we adopt an iterative method, where, for each i-th training sample T_i the SVR model is fitted, values of T_i are predicted (for performance evaluation studies). This i-th fitted SVR is then used to reconstruct the target parameters in C_i, i-th pseudo-set of the unreconstructed catalogue. The process is repeated N times.

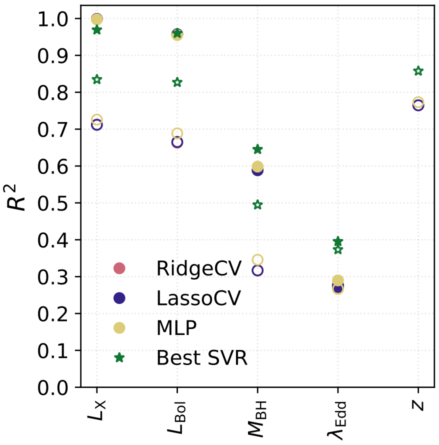


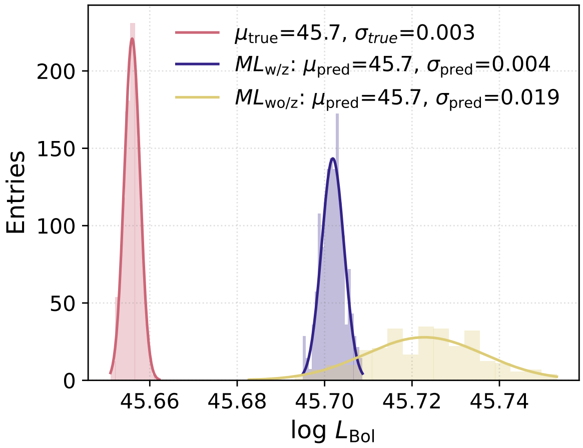
Fig. 12. Comparison of R^2 for ML-models tested on all target parameters. Full circles represent $ML_{\mathrm{w/z}}$ and open circles $ML_{\mathrm{wo/z}}$, the learning done for sources with unknown redshift. The SVR algorithm performs best on crucial variables (M_{BH} and λ_{Edd}).

The top panel of Fig. 13 displays the true and predicted distributions for the bolometric luminosity of a single source in the training sample. μ_{true} and σ_{true} are given by the SPIDERS-AGN catalogue for all target parameters, and all sources, while μ_{pred} and σ_{pred} are obtained by fitting a gaussian function to the N=200 reconstructed values for each source. From these, we construct the "pseudo-pull" distribution $\mu_{\mathrm{true}} - \mu_{\mathrm{pred}}$. We note how, in this single source, the $ML_{\mathrm{w/z}}$ (purple) and $ML_{\mathrm{wo/z}}$ (yellow) recontructed values of L_{Bol} are multiple σ_{pred} apart. This is also clear from the bottom panel of Fig. 13, which shows the distribution of all pulls for the reconstructed L_{Bol}. Here, the difference in the quality of reconstruction between $ML_{\mathrm{w/z}}$ and $ML_{\mathrm{wo/z}}$ is apparent in the greater smearing of the pull distribution.

The precision of the reconstruction is also derived from the normalized median absolute deviation $\sigma_{\mathrm{NMAD}} = 1.48\times$ median(| $\Delta\mu$ - median($\Delta\mu$) | /μ_{true}, expressed in %, with $\Delta\mu = \mu_{\mathrm{true}} - \mu_{\mathrm{pred}}$. It is a robust measure of the deviation, one that is insensitive to outliers.

Finally, the last metric we establish is the contamination level of the reconstruction: for each bin in μ_{true}, we look at the μ_{pred} distribution and fit a gaussian PDF (see Fig. 14). We then define the contamination to be the overlapping area between the lowest and highest true intervals reconstructed PDF, represented by the hashed area in Fig. 14. In the case of the λ_{Edd} parameter, it is the

[8] As default parameters, the MLP has one hidden layer with 100 neurons, use the "relu" activation function and the "adam" optimizer. More details can be found `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html`



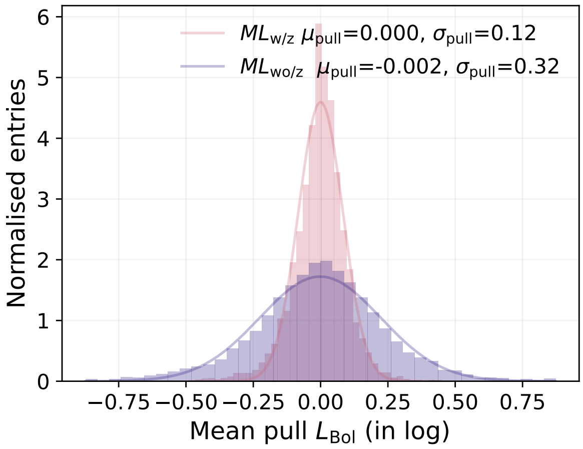


Fig. 13. *Top*: True (red), and predicted distributions reconstructed N times with $ML_{w/z}$ (purple) and $ML_{wo/z}$ (yellow) of the bolometric luminosity for a training source. The true value is represented by a normal distribution by taking into account the measurement error σ_{true} and assuming it to be gaussian. *Bottom*: Mean pull distribution for the L_{Bol} for all training sources, taking all $\mu_{true} - \mu_{pred}$ values for $ML_{w/z}$ (red) and $ML_{wo/z}$ (purple).

measure of how often our ML-model mistakes a low accretion-rate AGN with a high accretion-rate AGN, and vice versa. This is a useful measure, in the event that the regressor lacks the precision to carry on single source studies, but provides enough information to look at features of a larger population, as it does for λ_{Edd}: even though the mean of the binned reconstructed values do not correspond to the center of the true bins, the scaling relation between lower and higher λ_{Edd} is preserved.

All performance metrics described in this section are presented in Table 5 for the $ML_{w/z}$ and $ML_{wo/z}$ cases. In the following section, we discuss in greater details the ability of our ML-model to predict the physical parameters of AGN cores.

5.3. Prediction performance

Fig.15 and Fig. 16 summarizes the performance of the SVR for the case with ($ML_{w/z}$) and without redshift ($ML_{wo/z}$) respectively. Each bin of the response matrix is normalized to the true bins (by column). The matrix elements represent the probability for an AGN with target parameter P_{true} to be reconstructed with

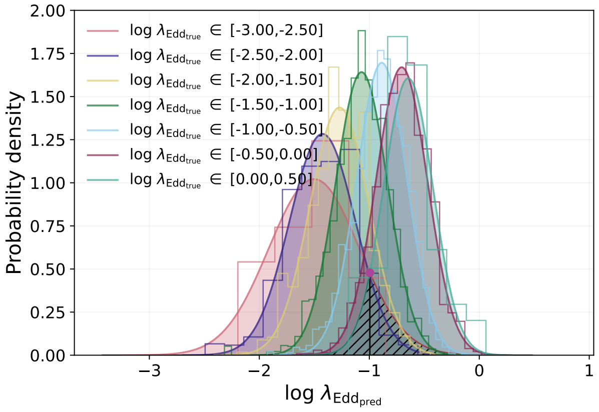


Fig. 14. Distributions of $\lambda_{Edd_{pred}}$ for various bins in $\lambda_{Edd_{true}}$, in log space. Gaussian PDFs are fitted and shown over their respective histograms. The hatched area corresponds to the contamination between the lowest and highest range PDFs: the values quoted in Table 5 were calculated by dividing the overlapping (hatched area) with the integral of the highest range PDF.

a value P_{pred}. The error on the reconstructed value σ_{pred} for each source is used as weight to the histogram.

5.3.1. Prediction of z

Determining redshifts through spectroscopic or photometric means has always been a primary goal of large AGN surveys. The overall performance of the SVR in predicting the redshift up to z ~ 2.5 can be seen on the top left matrix of Fig. 16. To better evaluate the accuracy of | Δz | $/(1 + z_{true})$ ($\Delta z = z_{pred} - z_{true}$), we modify our formula of σ_{NMAD} slightly to match the same estimator found in the literature (Brammer et al. 2008; Luo et al. 2010), such that σ_{NMAD}=1.48× median(| Δz- median(Δz) | $/(1 + z_{true})$. An outlier is defined as having | Δz | $/(1 + z_{true}) > 0.15$ (see Fig. 17). For redshifts reconstructed with the best parameter SVR, we find the rate of outlier to be 3.48% and σ_{NMAD}=0.071 (accuracy of 7.1%), performing just as well as the estimation of photometric redshifts using AGN SDE templates (Luo et al. 2010).

The regressor's reconstruction worsens as we move into higher z: this is simply because the SPIDERS dataset provides few sample for the supervised learning to train on, as the distribution of z decreases sharply (see Fig. 19). The ML reconstruction of redshifts proves to be a very reliable estimator for a crucial parameter for AGN studies, and one that could be adapted to different depths given an appropriate training dataset.

5.3.2. L_X and L_{Bol}

Naturally, once the model has an estimation for z, it can easily find the appropriate regression for L_X, given that the X-ray flux is one of the model's inputs. As hoped for, the X-ray luminosity is predicted with high accuracy and precision: when z is known, the error on the estimate is log(L_X/erg s^{-1}) ~ 0.05, and rises to ~ 0.33 when z is reconstructed. The bolometric luminosity L_{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 σ_{err} = 0.12 (0.31) for $ML_{w/z}$ ($ML_{wo/z}$). 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

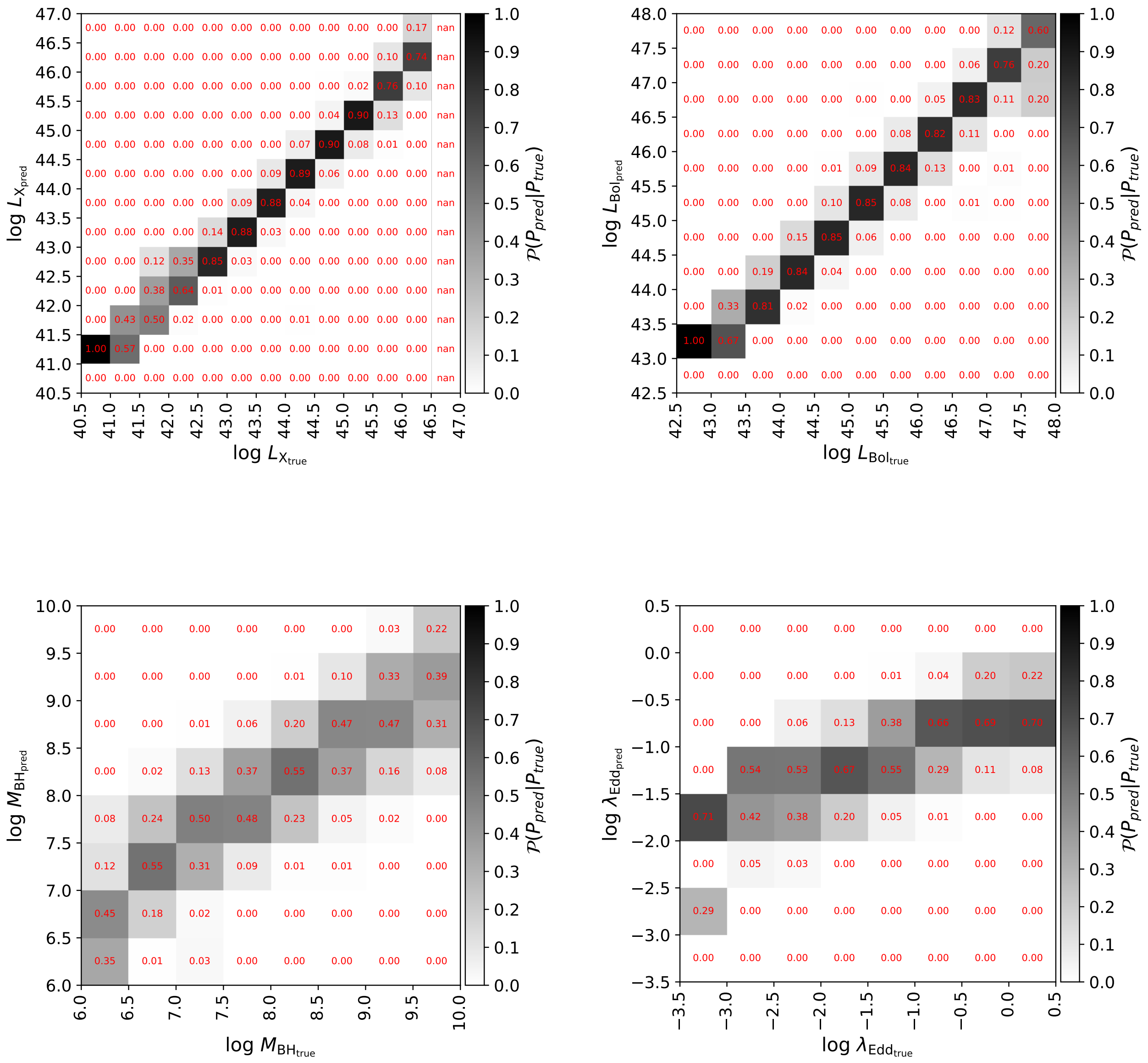


Fig. 15. Normalized performance matrices for ML-estimator with known redshift as an input ($ML_{\mathrm{w/z}}$). The true and reconstructed parameters are plotted on the x and y axis, respectively. The error on the reconstruction is used as a weight to the histogram.

tends to be overestimated for low values of the true parameter, while they are underestimated for high values.

5.3.3. Prediction of M_{BH} and λ_{Edd}

When it comes to estimations of M_{BH}, the knowledge of z of the source is the most determining factor for the performance, as a visual comparison of Figs 15 and 16 reveals. The R^2 score is 0.65 for $ML_{\mathrm{w/z}}$ and 0.49 for $ML_{\mathrm{wo/z}}$, and the width of the pull σ_{pull} goes from 0.54 to 0.66 in units of $\log(M_{\odot})$. The metric which shows the greatest discrepancy is the contamination level: $\sim 2\%$ for $ML_{\mathrm{w/z}}$ and $\sim 8\%$ $ML_{\mathrm{wo/z}}$.

The Eddington ratio is the hardest parameter to predict, since the errors from the previous predictions get propagated and compounded. It is also the characteristic for which the prior knowledge of z has the least impact, although the reconstruction of both L_{Bol} and M_{BH} is markedly better in the $ML_{\mathrm{w/z}}$. To understand why that is, we can study the correlations between the predicted parameters in the form of the mean pull values μ_{pull}. Fig. 18 presents the correlations of pulls between all parameters, for $ML_{\mathrm{w/z}}$ (red) and $ML_{\mathrm{wo/z}}$ (blue). In the case where the redshift z is the first predicted parameter in the chain regression ($ML_{\mathrm{wo/z}}$), all subsequent parameters remain more or less strongly correlated to one another, as the Pearson's correlation score attest to. That is, if a previous parameter is poorly reconstructed, the subsequent one will be as well. On the other hand, no such correlations between the pulls is found in the case of $ML_{\mathrm{w/z}}$, where L_{X} is the first estimated parameter.

Considering that the reconstruction quality $\sigma_{\lambda_{\mathrm{Edd}}}$ depends on $\sim \frac{\sigma_{\mathrm{M_{BH}}}}{\sigma_{\mathrm{L_{Bol}}}}$, and propagating the errors of the ratio gives

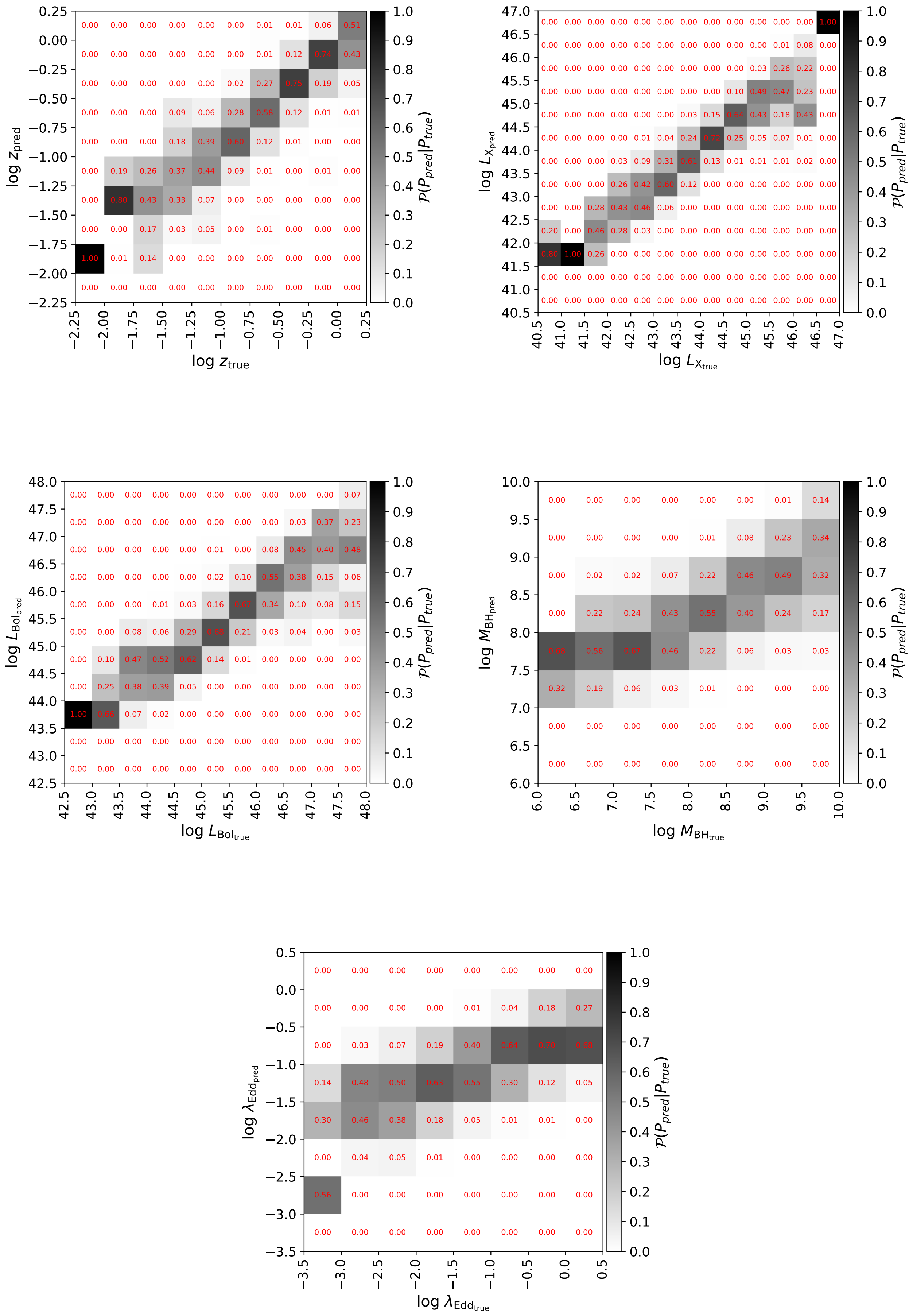


Fig. 16. Normalized performance matrices for the ML-estimator without a known redshift as an input ($ML_{\rm wo/z}$). The matrix for the reconstructed z is added to the variables already presented in Fig. 15.

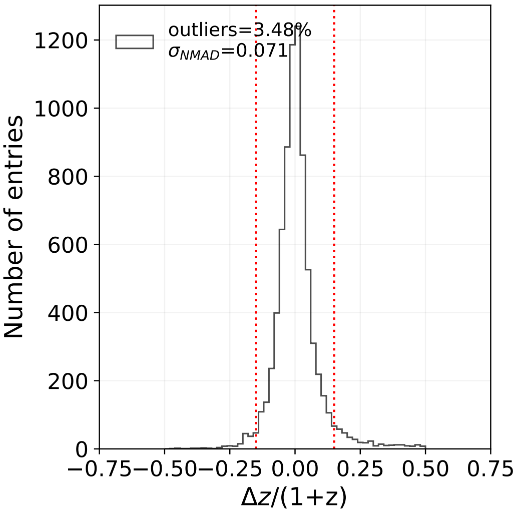


Fig. 17. Distribution of the predicted redshift $z_{\rm pred}$ accuracy derived with the true spectroscopically measured redshift $z_{\rm true}$. The dashed red lines represent the limit beyond which a prediction is counted as an outlier.

$\sim \sqrt{\sigma^2_{\rm M_{BH}} + \sigma^2_{\rm L_{Bol}} - 2\rho_{\rm M_{BH}-L_{Bol}}}$. Hence, $\sigma_{\lambda_{\rm Edd}}$ decreases with increasing $\rho_{\rm M_{BH}-L_{Bol}}$ value, which is higher for $ML_{\rm wo/z}$ (0.42 vs 0.12). That is, the covariance term decreases the uncertainty, which explains why the compounded parameter $\lambda_{\rm Edd}$ appears not to be better reconstructed with $ML_{\rm w/z}$ from $ML_{\rm wo/z}$

6. Reconstruction of the full catalogue

In this following section, we take a closer look at the estimated AGN physical parameters for 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 [9]. 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. The distribution of z for the 9944 sources in the full catalogue is overlaid on top of that of the training sample in Fig 19: the median is $\bar{z}$ = 0.52, and one can observe the range of z follows a similar trend, with slightly more AGN sources in $z > 1$ range.

6.1. Reconstruction quality for Type 1 and Type 2 AGN

Although the ML-model was trained on a sample of Type 1 AGN only, we have reconstructed the 5362 sources in our catalogue identified as Type 2 AGN, using the classifier and criteria presented in Sect. 4. Fig. 20 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. Sources reconstructed with $ML_{\rm w/z}$ have smaller uncertainties, following the results from the training sample (see Sect. 5), and the SVR has a harder time estimating parameters for Type 2 AGN. The shape of the distribution informs that 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.

Considering what was already shown in Fig. 9, we know that many AGN classified as Type 2 fall into the faint end of the optical magnitude distribution (eg: SDSS u-band mag > 22). Fig. 21 shows the uncertainty $\sigma_{\rm reco}$ on z as a function of the u-band magnitude: the fainter the source, the more outside of the bounds of u magnitude the ML model has trained on, the greater the uncertainty on the reconstructed parameter. This is yet another information one gains by propagating the input uncertainties and reconstructing each source iteratively: the difficulty the regressor encounters when estimating points outside of its known range is translated in the spread of the posterior distribution for all output parameters.

6.1.1. Control sample: reconstructing z for Type 2 AGN with known z

For the full catalogue, the only control parameter to verify the accuracy of the regressor's reconstruction for Type 2 AGN is the redshift, which will determine the successful reconstruction of subsequent parameters down the chain. This will also help to determine the range of reasonable extrapolation as a function of the input parameters. We use the 1416 Type 2 AGN that have a redshift information, and estimate $z_{\rm reco}$ using $ML_{\rm wo_z}$. We repeat the analysis done in Sect. 5.3.1 by calculating the | Δz | $/(1+z_{\rm true})$ distribution. We obtain an outlier rate of 7.3% and an accuracy of 8.2%, both under the 10% limit. Fig. 22 shows the accuracy for sources above a defined threshold in u-band magnitude (purple, top axis) and $\sigma_{\rm reco}$ (pink, bottom axis), respectively. The quality of the z reconstruction is in part driven by the optical faintness of the sample (and how far from the training range it lies), and in order to select a "purer" sample, one that is well reconstructed, a selection based on $\sigma_{\rm reco}$ is equivalent to one in optical brightness.

As an additional check, we also reconstruct subsequent parameters $L_{\rm Bol}$, $M_{\rm BH}$ and $\lambda_{\rm Edd}$, to further verify the soundness of extrapolating the ML regressor from Type 1 AGN to Type 2 AGN. The blue distribution in Fig. 20 shows the posterior distribution for the black hole mass parameter for Type 2 AGN with z information, reconstructed with $ML_{\rm wo_z}$. The blue curve follows the expected PDF for non-noisy reconstruction, a confirmation of the merit of the ML reconstruction for this sub-sample. This, and the good reconstruction of z for Type 2 AGN (with known z) can be taken as an assurance that within a range, the regressor is capable of estimating certain parameters for obscured sources.

6.1.2. Removal of outliers

As a last step, we remove sources in the final sample for which the reconstructed values lie too far beyond the phase space of the ML training. We require that $z_{\rm reco} < 4$ and $-5 < \log \lambda_{\rm Edd_{reco}} < 2$. 149 sources are removed after this selection. As already mentioned in Sect. 6.1, the 4457 Type 2 AGN without redshift z are given N/A values for $L_{\rm X}$, $L_{\rm Bol}$, $M_{\rm BH}$ or $\lambda_{\rm Edd}$, since the ML regressor is incapable of estimating these parameters. The reconstructed redshift and associated errors is however added to the catalogue.

6.2. Type 1 AGN and population studies

We now focus on the AGN classified as Type 1, as that population follows the training dataset more closely. Fig. 23 presents

[9] In the training set, these were called $\mu_{\rm pred}$ and $\sigma_{\rm pred}$, see top panel of Fig. 13

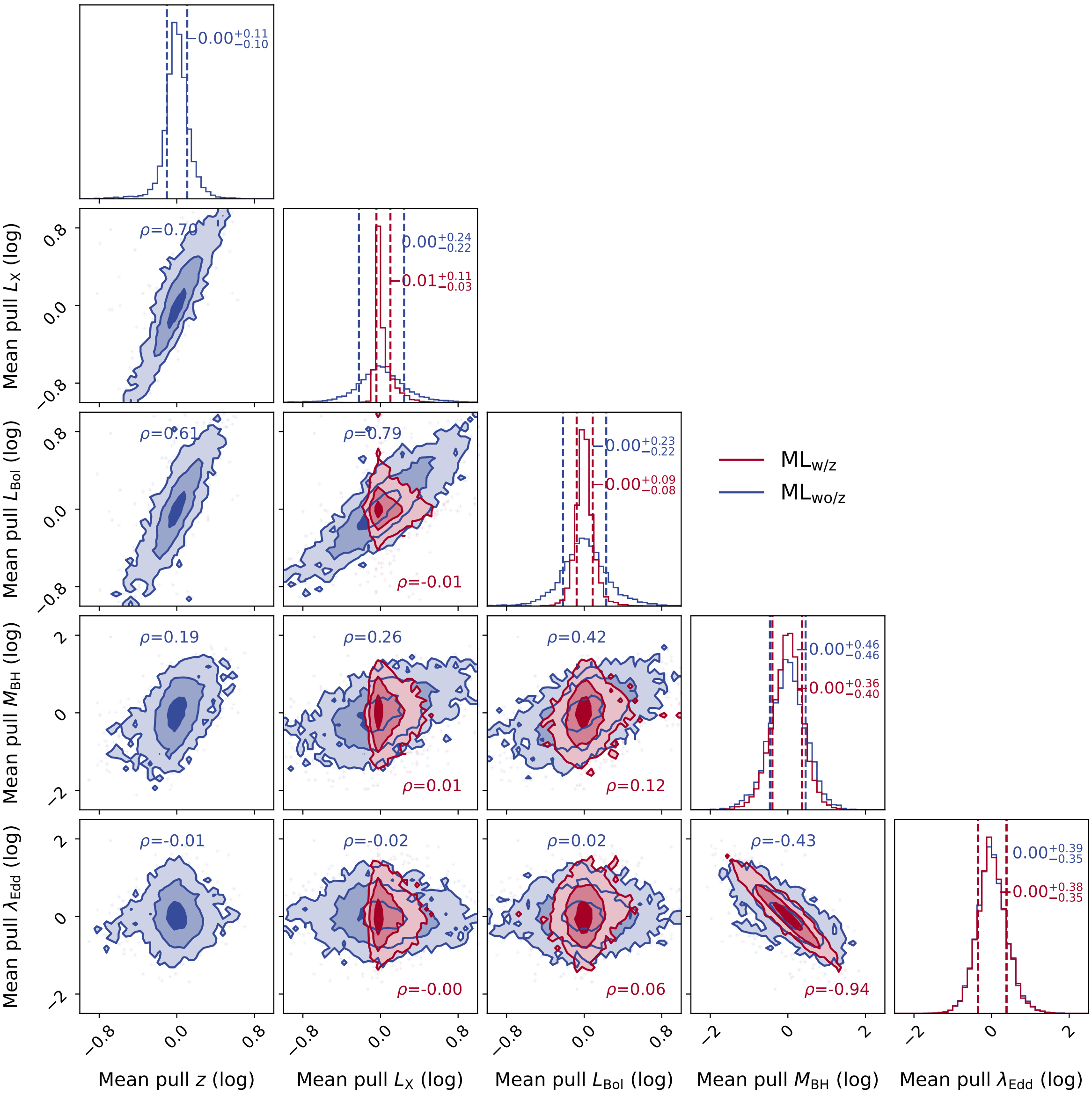


Fig. 18. Correlation between pull values for all parameters for $ML_{\mathrm{w/z}}$ (red) and $ML_{\mathrm{wo/z}}$ (blue). The quality of reconstruction, represented by the mean pull value, is more correlated between the variables in $ML_{\mathrm{wo/z}}$ than it is for $ML_{\mathrm{w/z}}$. The vertical dashed lines in the histograms indicate the 0.16 and 0.84 quantiles of the distributions, and the numbers show the respective medians and 0.16 and 0.84 quantiles.

Predicted parameter	Unit	z input?	R^2	$\mu_{\mathrm{pull}} \pm \sigma_{\mathrm{pull}}$	σ_{NMAD}	Contamination
z	log	$ML_{\mathrm{wo/z}}$	0.86	0.003 ± 0.150	7.07%	0.00%
L_{X}	erg.s^1(log)	$ML_{\mathrm{w/z}}$	0.97	-0.012 ± 0.049	5.81%	0.0%
		$ML_{\mathrm{wo/z}}$	0.83	0.008 ± 0.334	22.5%	0.43%
L_{Bol}	erg.s^1(log)	$ML_{\mathrm{w/z}}$	0.96	-0.000 ± 0.117	7.71%	0.0%
		$ML_{\mathrm{wo/z}}$	0.83	0.001 ± 0.318	21.5%	1.94%
M_{BH}	log($M_{\odot}$)	$ML_{\mathrm{w/z}}$	0.65	-0.009 ± 0.546	37.3%	1.49%
		$ML_{\mathrm{wo/z}}$	0.49	-0.001 ± 0.661	45.0%	7.53%
λ_{Edd}	log	$ML_{\mathrm{w/z}}$	0.40	0.008 ± 0.524	35.8%	13.4%
		$ML_{\mathrm{wo/z}}$	0.37	0.008 ± 0.532	36.7%	5.96%

Table 5. Results of the best-tuned SVR estimators, $ML_{\mathrm{w/z}}$ and $ML_{\mathrm{wo/z}}$, for the performance metrics presented in Sect. 5.2.2 for all target parameters.

the bolometric luminosity versus BH mass (top panel) for the Type 1 AGN presented in this work and the SPIDERS AGN: the reconstructed sample is well bounded by the Eddington limit. The bottom panel of Fig. 23 shows the Eddington ratio as a function of the redshift for the same subsamples. As the response matrices in Fig. 15 and Fig. 16 have shown, the ML-model is not very apt to reconstruct extreme cases, in the lower and upper tails of the target parameter distribution. That is, it will overestimate low values and underestimate higher ones: this is indeed visible in the bottom panel of Fig. 23, where the reconstructed samples occupy a smaller region of the log λ_{Edd} space than the SPIDERS AGN do.

The top panel of Fig 24 presents the AGN number source density over a wide range of redshifts for several bins in bolo-

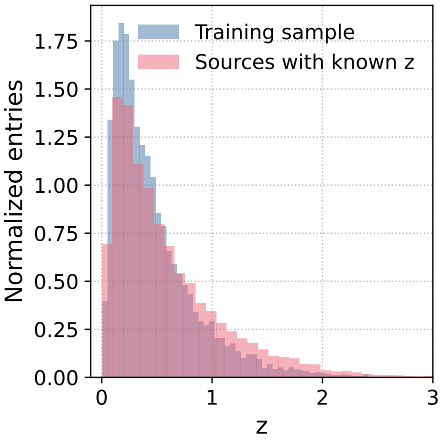


Fig. 19. Distribution of z for the training sample coming from the SPIDERS AGN catalogue (blue) and the subsample of AGN sources in the reconstruction sample for which z is known (red).

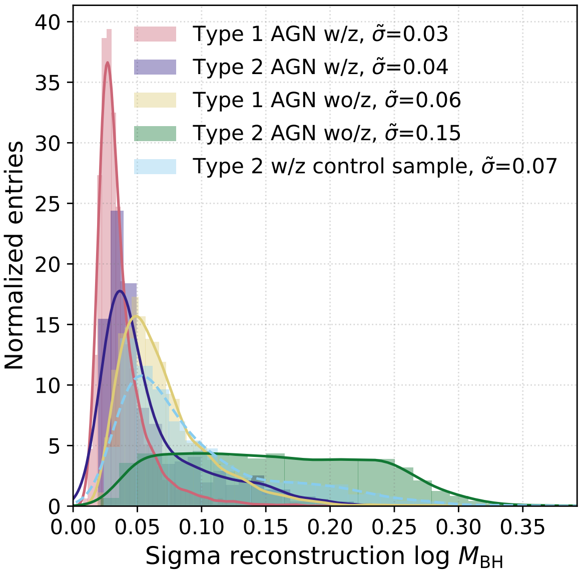


Fig. 20. Uncertainty $\sigma_{\rm reco}$ on the log of the reconstructed black hole mass $M_{\rm BH}$ for Type 1 and 2 AGN, with and without z information (estimated with $ML_{\rm w/z}$ and $ML_{\rm wo/z}$ respectively). The $\tilde{\sigma}$ values correspond to the median of their distributions. The 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.

metric 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 Brandt & Alexander (2015)). The bottom panel of Fig 24 shows the black hole masses of these sources, using the same binning in $L_{\rm Bol}$. Although the tails of the distributions are not well represented in the reconstructed sample when matched to their SPIDERS AGN counterpart, importantly, 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.

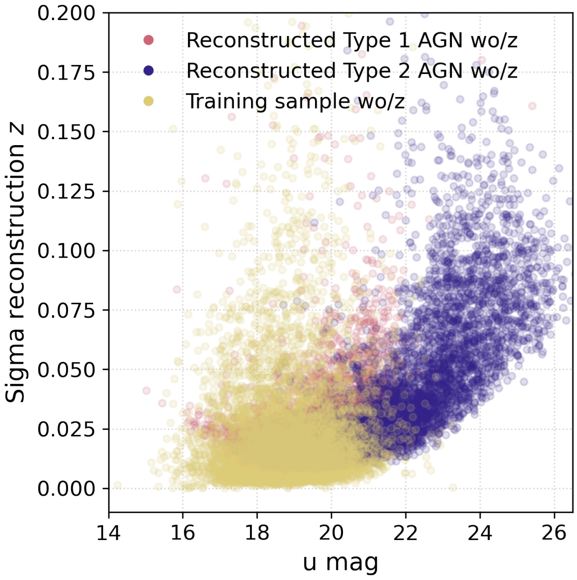


Fig. 21. Uncertainty of the reconstructed value on z as a function of the SDSS u-band magnitude for the training sample (true), and reconstructed Type 1 and 2 AGN. The fainter – and outside of the bounds of the training sample range — the greater the $\sigma_{\rm reco}$ of the N iterations is.

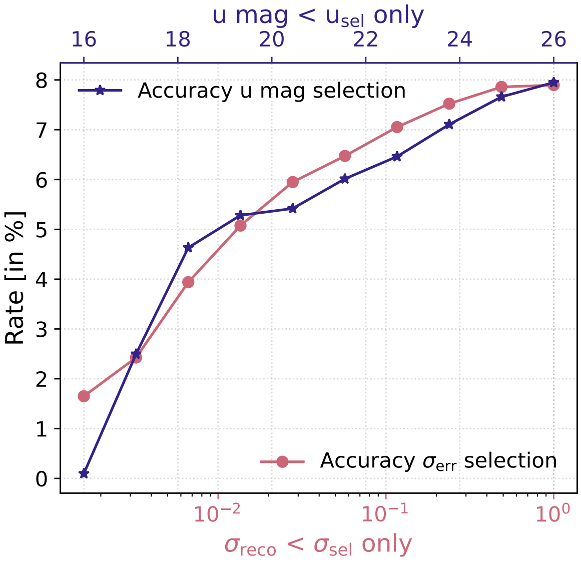
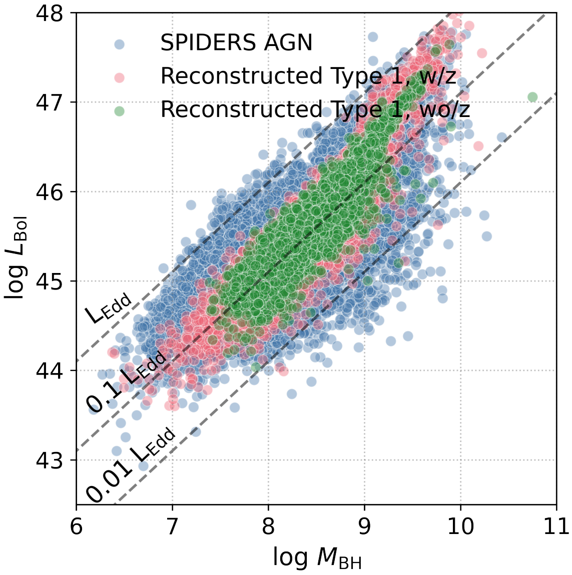


Fig. 22. Accuracy for Type 2 AGN z reconstuction for sources above the indicated threshold in $\sigma_{\rm reco}$ value (pink) and the SDSS u-band magnitude value (purple). The outlier rate is driven by faint sources outside of the training range. A quality selection in $\sigma_{\rm reco}$ is equivalent to one in the optical brightness.

Deriving a luminosity function from these AGN would require to correct the number density for detection and selection efficiencies and biases, a non-trivial task considering the many catalogues used to build our sample (Schulze et al. 2015; Weigel et al. 2017; Ananna et al. 2022): this is beyond the scope of the paper.



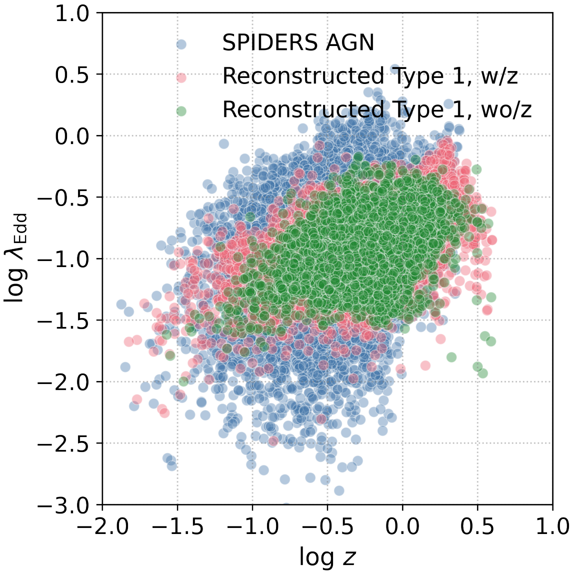
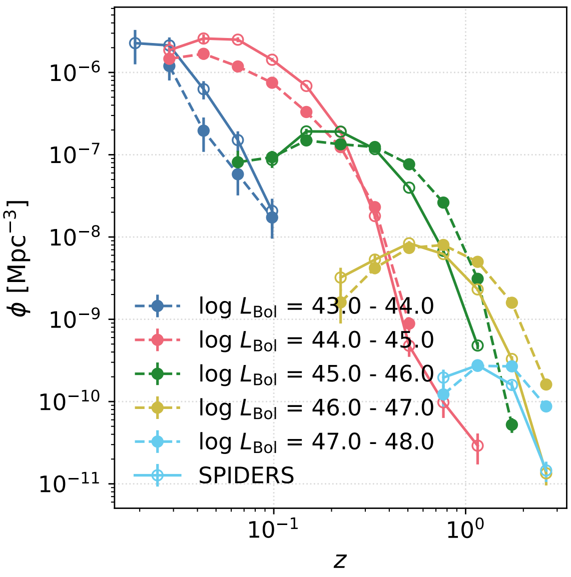


Fig. 23. (*Top*) Scatter plot of $L_{\rm Bol}$ as a function of $M_{\rm BH}$ for the SPIDERS AGN (blue dots), and reconstructed AGN classified as Type 1. AGN with known z measurement are shown in red dots, and those with reconstructed z are represented in green. (*Bottom*) Same three samples for the $\lambda_{\rm Edd}$ vs z distribution.



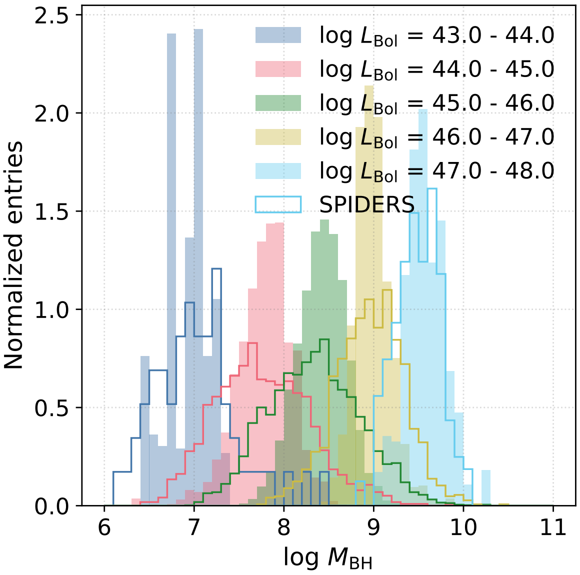


Fig. 24. (*Top*) AGN downsizing: comoving number density vs. redshift for Type 1 AGN from this work's catalogue (full circles) and the SPIDERS AGN catalogue (open circles) for different bins of $L_{\rm Bol}$ in units of log(erg s^{-1}). (*Bottom*) Distribution of $M_{\rm BH}$ for the same bins of bolometric 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.

7. Summary and conclusions

We release the first photometry based estimates of bolometric luminosity $L_{\rm Bol}$, black hole mass $M_{\rm BH}$, and Eddington ratio $\lambda_{\rm Edd}$ for 16 907 sources ranging over 6 dex in luminosity and up to z=2.5 in redshift. For 11 420 of these sources the redshift was previously determined spectroscopically and is used in the estimation of the remaining parameters, as well as for the verification of the redshift estimate. For 11 404 sources without a previously known redshift, the reconstructed z is provided. An uncertainty is given for all estimated parameters, thanks to a simulation based technique which incorporates measurement errors in the fit and reconstruction of the ML regressor. In addition, we have demonstrated how ML classification tools can help identify obscured AGN, a crucial challenge in the field (Hickox & Alexander 2018).

While the addition of ~15 000 Type 1 AGN sources from this catalogue might not dramatically improve our knowledge of the luminosity function, considering that the 8000 SPIDERS AGN sources were measured with greater accuracy in the same phase space, 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. AGN have been favoured to be strong cosmic ray emitters (Murase 2022; Murase & Stecker 2022), with the recent discovery showing the nearby obscured AGN NGC 1068 to be a steady source of neutrinos (ICECUBE COLLABORATION et al. 2022). Searches for a cumulative signal from different AGN populations, such as Abbasi et al. (2022), can help characterize which sources contribute most to the flux of neutrinos, based on their accretion parameters. By strategically targeting a subset of sources observed spectroscopically, we would able to train similar ML-algorithms and reconstruct a larger sample of photometrically measured AGN. The method can easily be expanded to other cosmic demographics — higher z for instance— granted a corresponding dataset is provided to train a ML algorithm. In this work, we were limited by demanding that sources had been observed with SDSS photometry: this constrained the coverage to a quarter of the full sky. A natural next step would be to expand the optical sky coverage by cross-matching sources with the Pan-STARRS 3π survey Flewelling et al. (2020), and recover most AGN identified by infrared and X-ray telescopes. Finally, eROSITA has been scanning the full-sky with unprecedented sensitivity in the soft (0.2–2.3 keV) and hard (2.3–8 keV) bands (Merloni et al. 2012; Predehl et al. 2021). Incorporating this dataset will also offer new understanding of obscured AGN, as harder X-ray photons are transparent to obscuring dust.

Acknowledgements. The author wishes to thank Iftach Sadeh, Anna Franckowiak, Sjoert Van Velzen, Jannis Necker, Simone Garrapa and Jonas Sinapius for their fruitful comments and support. Timo Karg is to be thanked for helping with various technicalities.

References

Abbasi, R., Ackermann, M., Adams, J., et al. 2022, Phys. Rev. D, 106, 022005

Ahumada, R. 2020, The Astrophysical Journal Supplement Series, 21

Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, ApJS, 219, 12

Ananna, T. T., Weigel, A. K., Trakhtenbrot, B., et al. 2022, ApJS, 261, 9

Arenou, F., Luri, X., Babusiaux, C., et al. 2017, A&A, 599, A50

Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, ApJ, 772, 26

Blanton, M. R. 2017, The Astronomical Journal, 35

Boller, T., Freyberg, M. J., Trümper, J., et al. 2016, A&A, 588, A103

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, Astrophysical Journal, 686, 1503

Brandt, W. N. & Alexander, D. M. 2015, Astron Astrophys Rev, 23, 1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321

Clerc, N., Merloni, A., Zhang, Y.-Y., et al. 2016, Mon. Not. R. Astron. Soc., 463, 4490

Coffey, D., Salvato, M., Merloni, A., et al. 2019, A&A, 625, A123

Comparat, J., Merloni, A., Dwelly, T., et al. 2020, A&A, 636, A97

Cortes, C. & Vapnik, V. 1995, Mach Learn, 20, 273

Cutri, R. M., Wright, E. L., Conrow, T., et al. 2021, VizieR Online Data Catalog, II/328, aDS Bibcode: 2014yCat.2328....0C

Dwelly, T., Salvato, M., Merloni, A., et al. 2017, Monthly Notices of the Royal Astronomical Society, 469, 1065

Feigelson, E. D., de Souza, R. S., Ishida, E. E., & Babu, G. J. 2021, Annual Review of Statistics and Its Application, 8, 493, _eprint: https://doi.org/10.1146/annurev-statistics-042720-112045

Flewelling, H. A., Magnier, E. A., Chambers, K. C., et al. 2020, ApJS, 251, 7

Fotopoulou, S. & Paltani, S. 2018, A&A, 619, A14

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, AJ, 157, 98

Hasinger, G. 2008, A&A, 490, 905

Hickox, R. C. & Alexander, D. M. 2018, Annual Review of Astronomy and Astrophysics, 56, 625, _eprint: https://doi.org/10.1146/annurev-astro-081817-051803

ICECUBE COLLABORATION, Abbasi, R., Ackermann, M., et al. 2022, Science, 378, 538, publisher: American Association for the Advancement of Science

Kormendy, J. & Ho, L. C. 2013, Annu. Rev. Astron. Astrophys., 51, 511, arXiv:1304.7762 [astro-ph]

Luo, B., Brandt, W. N., Xue, Y. Q., et al. 2010, ApJS, 187, 560

Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8

Mainzer, A., Bauer, J., Grav, T., et al. 2011, ApJ, 731, 53

Merloni, A., Predehl, P., Becker, W., et al. 2012, eROSITA Science Book: Mapping the Structure of the Energetic Universe, arXiv:1209.3114 [astro-ph]

Murase, K. 2022, Science, publisher: American Association for the Advancement of Science

Murase, K. & Stecker, F. W. 2022, arXiv:2202.03381 [astro-ph, physics:hep-ph], arXiv: 2202.03381

Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, Astron Astrophys Rev, 25, 2

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, MACHINE LEARNING IN PYTHON, 6

Predehl, P., Andritschke, R., Arefiev, V., et al. 2021, A&A, 647, A1

Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502

Saito, T. & Rehmsmeier, M. 2015, PLoS ONE, 10, e0118432

Salvato, M., Buchner, J., Budavári, T., et al. 2018, Monthly Notices of the Royal Astronomical Society, 473, 4937

Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61

Saxton, R. D., Read, A. M., Esquej, P., et al. 2008, A&A, 480, 611

Schulze, A., Bongiorno, A., Gavignaud, I., et al. 2015, Monthly Notices of the Royal Astronomical Society, 447, 2085

Shy, S., Tak, H., Feigelson, E. D., Timlin, J. D., & Babu, G. J. 2022, AJ, 164, 6

Stevens, G., Fotopoulou, S., Bremer, M., & Ray, O. 2021, JOSS, 6, 3635

Stone, M. 1974, Journal of the Royal Statistical Society: Series B (Methodological), 36, 111, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1974.tb00994.x

Trümper, J. 1982, Advances in Space Research, 2, 241

Voges, W., Aschenbach, B., Boller, T., et al. 2000, International Astronomical Union Circular, 7432, 3, aDS Bibcode: 2000IAUC.7432....3V

Véron-Cetty, M.-P. & Véron, P. 2010, A&A, 518, A10

Weigel, A. K., Schawinski, K., Caplar, N., et al. 2017, ApJ, 845, 134

Wolf, J., Salvato, M., Coffey, D., et al. 2020, Monthly Notices of the Royal Astronomical Society, 492, 3580

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, The Astronomical Journal, 140, 1868

Appendix A: Catalogue column description

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, $L_{\rm Bol}$, $M_{\rm BH}$, $L_{\rm Edd}$ and $\lambda_{\rm Edd}$ with associated reconstruction uncertainties. For 4457 sources, of Type 2 AGN without previous z information, entries for L_X, $L_{\rm Bol}$, $M_{\rm BH}$, $L_{\rm Edd}$ and $\lambda_{\rm Edd}$ are left blank.

In addition, the 7616 SPIDERS sources used in the training sample are also included. A description of the catalogue's columns is given below. Features providing X-ray, IR, optical information were taken from the references listed in Table 1.
Column 1- **x-ray detection**: Flag indicating whether the X-ray source was detected in the 2RXS or XMMSL2 survey (Boller et al. 2016; Saxton et al. 2008).
Column 2- **name**: X-ray identifier from 2RXS or XMMSL2 survey .
Column 3-4- **RA, DEC**: Right ascension and declination of the X-ray detection (J2000) in degrees.
Column 5-6- **Flux, Flux error**: X-ray flux and error converted to the 0.5-2 keV band in $\log_{10}$(erg cm^{-2}s^{-1}).
Column 7- **ALLW_ID**: WISE All-Sky Release Catalogue name (Cutri et al. 2021)
Column 8-9- **ALLW_RA, ALLW_DEC**: J2000 AllWISE Right Ascension and Declination, in degrees.
Column 10-13- **W[1234]**: AllWISE Vega magnitude in the W1, W2, W3, W4 bands.
Column 14-17- **W[1234] error**: AllWISE Vega magnitude errors in the W1, W2, W3, W4 bands.
Column 18-19- **Gaia mean flux, Gaia mean flux error**: *Gaia* mean flux and error in units of e s^{-1}.
Column 20- **CLASS**: Broad spectral classification computed by the SDSS-DR16 spectroscopic pipeline.
Column 21- **SUBCLASS**: Detailed spectral classification computed by the SDSS-DR16 spectroscopic pipeline.
Column 22-26- **psfMag_[ugriz]**: Point spread function magnitude of the optical counterpart to the IR source in the *ugriz* band (mag, AB).
Column 27-31- **psfMagErr_[ugriz]**: Uncertainties on the PSF magnitude in the *ugriz* band (mag, AB).
Column 32- **z**: Redshift of the source, and uncertainty.
Column 33- **z error**: Uncertainty on the redshift.
Column 34- **Luminosity**: X-ray luminosity in the 0.5-2 keV band in $\log_{10}$(erg.s^{-1}).
Column 35- **Luminosity error**: Uncertainty on the X-ray luminosity in the 0.5-2 keV band in $\log_{10}$(erg.s^{-1}).
Column 36- **Bolometric luminosity**: Bolometric luminosity in $\log_{10}$(erg.s^{-1}).
Column 37- **Bolometric luminosity error**: Uncertainty on the bolometric luminosity in $\log_{10}$(erg.s^{-1}).
Column 38 **Black hole mass**: BH mass in $\log_{10}(M_\odot$
).
Column 39 **Black hole mass error**: Uncertainty on the BH mass in $\log_{10}(M_\odot$
).
Column 40- **Eddington luminosity**: Eddington luminosity in $\log_{10}$(erg.s^{-1}).
Column 41- **Eddington luminosity error**: Uncertainty on the Eddington luminosity in $\log_{10}$(erg.s^{-1}).
Column 42- **Eddington ratio**: Eddington ratio in log.
Column 43- **Eddington ratio error**: Uncertainty on the Eddington ratio in log.
Column 44- **Reconstructed**: Flag indicating whether the source and values from columns 32-43 come from this work's ML reconstruction (flag==1) or from the SPIDERS AGN spectroscopic catalogue (flag==0) (Coffey et al. 2019).
Column 45- **known z**: Flag indicating whether the redshift values from column 32-33 come from this work's ML reconstruction (flag==0) or from the previous spectroscopic visual derived redshift (flag==1) (Dwelly et al. 2017; Véron-Cetty & Véron 2010).
Column 46- **obscuration**: Value between 0 and 1 indicating whether the source is obscured (obscuration ∼ 1) or not (obscuration ∼ 0), from the ML classifier presented in Sect. 4.
Column 47- **obscuration error**: Uncertainty on the obscuration value.

Appendix B: Feature selection of Type 2 AGN

Using the known SDSS classification for the subsample of 9535 AGN, we can distinguish Type 2 from Type 1 galaxies in the W2/W1 space (see top panel of Fig. B.1, following the method outlined in Abbasi et al. (2022). We can use these two distributions to define an "Obscuration" PDF as:

$$\text{Obscuration(W2/W1)} = \frac{\mathcal{P}(\text{Type2})}{\mathcal{P}(\text{Type2}) + \mathcal{P}(\text{Type1})} \tag{B.1}$$

where $\mathcal{P}$(Type1) and $\mathcal{P}$(Type2) are the probabilities of an AGN being of Type 1 and Type 2 respectively, according to the normalized histograms of Fig. B.1

By applying Equation B.1 to these two distributions, we obtain the Obscuration PDF shown in the bottom panel of Fig. B.1, with the blue line representing a sigmoid fit to the data points. Using this fitted function, we obtain the distribution of SDSS-classified Type 1 and Type 2 AGN as a function of the derived Obscuration PDF (top panel of Fig. B.2. By scanning through Obscuration PDF values between 0 and 1, we can calculate the precision and recall for each threshold , based on the definitions given in Sect. 4. This then gives us the PR curve presented in the bottom panel of Fig. B.2. Using this single feature classification, we reach a selection efficiency of 79.7% and a contamination rate of 21.9%, for an optimized Obscuration PDF threshold of 0.42.

Appendix C: Handling of null entries

The supervised ML algorithm cannot accept null entries for any of the features. This is true for the training sample, as well as the catalogue to be reconstructed. As explained in Sect. 2.4, demanding SDSS photometry observation for all AGN sources represents the greatest cut to the data, however, some features still remain incomplete. Instead of indiscriminately removing these data points, we look for correlations between fully complete features and partially incomplete ones: we fit function which describes the relationship in that parameter space in order to derive dummy values. For instance, 726 sources in the SPIDERS catalogue have AllWISE W4 magnitudes but with missing error measurements. This number is 7846 for the same feature in the full catalogue compiled. The top panel of Fig. C.1 shows the W4 magnitude as a function of the W4 error for AGN with complete information in the SPIDERS sample. An exponential function is fit and W4 error are interpolated for sources which are missing entries (yellow points). Similarly 297 SPIDERS and 6382 full catalogue sources have no entries for the Gaia mean flux and errors. The soft X-ray flux $F_{0.5-2\rm keV}$ is then used to derive a value

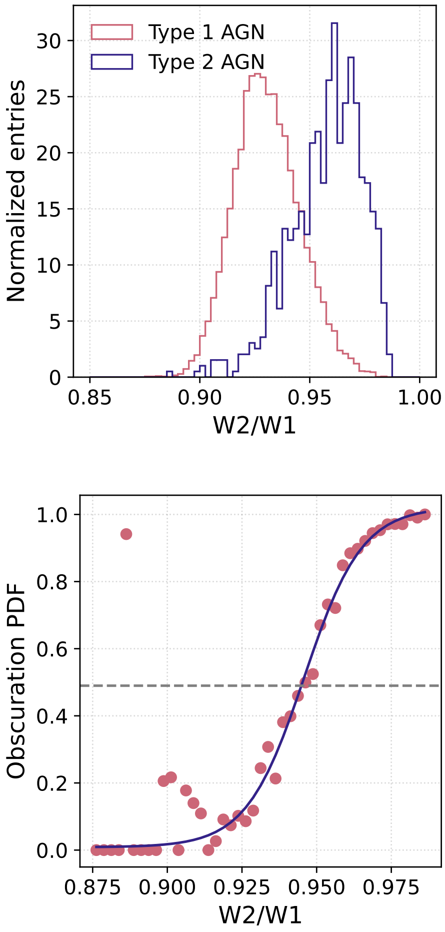


Fig. B.1. Obscuration PDF definition. (*Top*) distribution of the SDSS sources classified as Type 1 and Type 2 AGN as a function of W2/W1 magnitudes. (*Bottom*) Obscuration PDF derived from the left figure and Eq. B.1 from a sigmoid fit to the points. The dashed gray line represents the cut threshold which defines whether an AGN is of Type 1 or 2. It was chosen by doing a grid search over Obscuration PDF values and choosing the value giving the best F1-score.

for the Gaia mean flux using a log-log fit (bottom panel of Fig. C.1. The relationship between the Gaia mean flux and its associated error itself is then used to complete the error column.

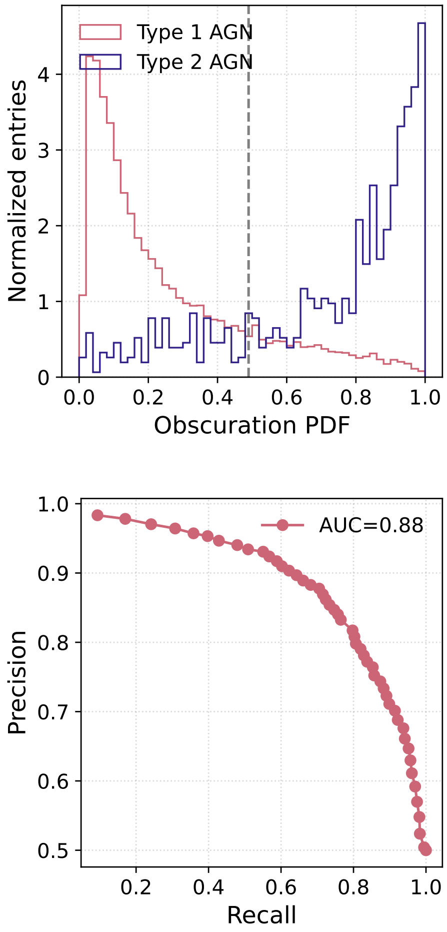


Fig. B.2. Obscuration PDF precision recall curve definition. (*Top*) Distrubution of SDSS defined Type 1 and 2 AGN as a function of their derived Obscuration PDF. The vertical gray dashed line represents the threshold value giving the optimal classification perfomance. (*Bottom*) Precision recall curve derived from the left distribution, by scanning through values between 0 and 1 and calculating the precision and recall. The optimized cut value corresponds to the point in the PRC where the distance between precision and recall is maximum.

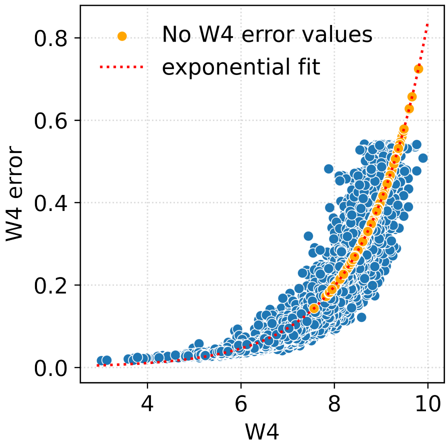


Fig. C.1. Handling of null entries: W4 error values for points missing one are estimated (yellow points) using an exponential fit to the W4 error vs W4 plane.