A&A 675, A65 (2023)
https://doi.org/10.1051/0004-6361/202345980
© The Authors 2023

Astronomy & Astrophysics

A machine learning algorithm for reliably predicting active galactic nucleus absorbing column densities[★]

R. Silver[1], N. Torres-Albà[1], X. Zhao[2], S. Marchesi[3,1,4], A. Pizzetti[1], I. Cox[1], and M. Ajello[1]

[1] Department of Physics and Astronomy, Clemson University, Kinard Lab of Physics, Delta Epsilon Ct., Clemson, SC 29634, USA
e-mail: rmsilve@clemson.edu
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[3] Dipartimento di Fisica e Astronomia (DIFA), Università di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[4] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti, 93/3, 40129 Bologna, Italy

Received 23 January 2023 / Accepted 16 May 2023

ABSTRACT

We present a new method for predicting the line-of-sight column density ($N_{\rm H}$) values of active galactic nuclei (AGN) based on mid-infrared (MIR), soft X-ray, and hard X-ray data. We developed a multiple linear regression machine learning algorithm trained with WISE colors, *Swift*-BAT count rates, soft X-ray hardness ratios, and an MIR–soft X-ray flux ratio. Our algorithm was trained off 451 AGN from the *Swift*-BAT sample with known $N_{\rm H}$ and has the ability to accurately predict $N_{\rm H}$ values for AGN of all levels of obscuration, as evidenced by its Spearman correlation coefficient value of 0.86 and its 75% classification accuracy. This is significant as few other methods can be reliably applied to AGN with Log($N_{\rm H}$ < 22.5). It was determined that the two soft X-ray hardness ratios and the MIR–soft X-ray flux ratio were the largest contributors toward accurate $N_{\rm H}$ determinations. We applied the algorithm to 487 AGN from the BAT 150 Month catalog with no previously measured $N_{\rm H}$ values. This algorithm will continue to contribute significantly to finding Compton-thick (CT) AGN ($N_{\rm H} \geq 10^{24}\,{\rm cm}^{-2}$), thus enabling us to determine the true intrinsic fraction of CT-AGN in the local Universe and their contribution to the cosmic X-ray background.

Key words. infrared: galaxies – galaxies: active – galaxies: nuclei – X-rays: galaxies – X-rays: diffuse background – methods: data analysis

1. Introduction

Active galactic nuclei (AGN) are supermassive black holes (SMBHs) that reside in the center of nearly all massive galaxies and accrete nearby material. They are one of the most powerful source classes in the Universe, and they emit over the entire electromagnetic spectrum. It has been shown that the masses of SMBHs correlate with those of the host galaxy bulge, velocity dispersion, and luminosity (Magorrian et al. 1998; Richstone et al. 1998; Gebhardt et al. 2000; Merritt & Ferrarese 2001; Ferrarese & Ford 2005; Kormendy & Ho 2013). This trend indicates that SMBHs may determine star formation rates, due to molecular and ionized outflows (Ferrarese & Merritt 2000; Gebhardt et al. 2000; Di Matteo et al. 2005; Merloni et al. 2010; Fiore et al. 2017; Martín-Navarro et al. 2018). If true, then the cosmic evolution of SMBHs and their host galaxies are inextricably linked. Therefore, being able to study the properties of SMBHs, including the gas and dust that surrounds them, becomes crucial.

One of the best ways to study AGN through cosmic time is via the cosmic X-ray background (CXB), that is, the diffuse X-ray emission from 1 to 200−300 keV (e.g., Alexander et al. 2003; Gilli et al. 2007; Treister et al. 2009; Ueda et al. 2014; Brandt & Yang 2022). Models have shown that a significant fraction (15−20%; Gilli et al. 2007; Ananna et al. 2019) of the peak of the CXB (∼30 keV; Ajello et al. 2008) is generated by a population of AGN with large obscuring column densities, $N_{\rm H,los} \geq 10^{24}\,{\rm cm}^{-2}$, categorized as Compton-thick (CT) AGN. Additionally, population synthesis models designed to accurately describe the origins of the CXB estimate that between 20% (Ueda et al. 2014) and 50% (Ananna et al. 2019) of all AGN are CT. Nonetheless, the current fraction of observed CT-AGN is only between 5% and 10% (Burlon et al. 2011; Ricci et al. 2015), although samples limited to very small volumes ($z < 0.01$) have reached up to 20% (Torres-Albà et al. 2021).

Discovering CT-AGN is challenging because the majority of their emission, from the optical through the soft X-rays, is obscured by the surrounding dust and gas (i.e., the torus; Urry & Padovani 1995). However, the hard X-rays (>10 keV) and the mid-infrared (MIR; 3−30 μm) are able to pierce through the torus up to high column densities, making them the least biased bands for the detection of heavily obscured AGN (Treister et al. 2004; Stern et al. 2005; Alexander et al. 2008). Hard X-ray emission is created when UV light from the accretion disk interacts with hot electrons in the corona above the disk, thus Compton up-scattering into the hard X-ray band (Haardt & Maraschi 1993). Additionally, the same UV radiation is absorbed by the dust, which in turn emits thermally in the infrared (Almeida & Ricci 2017; Hönig 2019). Because of this, the emission in these two bands is expected to correlate significantly in AGN. Therefore, targeting X-rays and the MIR is the ideal way to discover new CT-AGN.

Observing and analyzing spectra of AGN in X-rays and the infrared to identify strong CT candidates is a time- and

[★] A table of data of the 451 sources used to train and test the algorithm is only available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/675/A65 resource-intensive endeavor (see, e.g., Marchesi et al. 2017; Andonie et al. 2022; Silver et al. 2022). The Burst Alert Telescope (BAT) on board *Swift* (Gehrels et al. 2004) detected 1390 sources which are listed in its 150 Month catalog[1], almost 500 more than its predecessor, the 100 Month catalog. With the addition of hundreds of sources in every catalog release, an efficient and accurate method for identifying potential heavily obscured AGN is necessary. For this reason, our team has developed a new multiple linear regression machine learning algorithm to predict the line-of-sight column density of AGN. We constructed a large sample of AGN with known $N_{\rm H}$ values and trained the algorithm using MIR data from the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010), soft X-ray data from the *Swift*-X-ray Telescope (XRT), and hard X-ray data from *Swift*-BAT to accurately predict the column density of newly identified AGN.

The remainder of this article is organized as follows: Sect. 2 discusses the creation of our sample, and Sect. 3 describes how the $N_{\rm H}$ values were determined. Section 4 details the algorithm implemented and the input parameters included. Section 5 discusses the results of our algorithm and compares our predictive capabilities to other recent methods that are based on linear data modeling, rather than on multi-parameter machine learning algorithms such as the one presented in this paper.

2. Sample selection

Our sample is taken from the 1390 sources detected in the BAT 150 Month Catalog[2] (Imam et al., in prep.). Of these 1390 sources, 568 are AGN with reliable[3] $N_{\rm H}$ determinations (see Sect. 3 for details). Asmus et al. (2015) show that the ratio of the MIR and X-ray flux can be a strong predictor of column density. Therefore, our machine learning algorithm (see Sect. 4.2) includes data from XRT and WISE, and we thus cross-matched (using the BAT counterpart coordinates) with the 2SXPS (Evans et al. 2020) and AllWISE (Cutri et al. 2013) with 5″ and 10″, respectively. For the 2SXPS and AllWISE, we found an average separation of ∼1.7″ and ∼1.8″, and a standard deviation of ∼1.1″ and ∼1.5″, respectively. This left us with a sample of 451 sources to train and test our machine learning algorithm (see Sect. 4).

3. Data analysis

The majority of the sources (361) in our sample of 451 are in the BAT 70-month catalog (Ricci et al. 2017), which provides $N_{\rm H}$ values based on spectral analysis of soft X-ray (ASCA, *Chandra*, Suzaku, *Swift*-XRT, and *XMM-Newton*) and BAT spectra. For the remaining 90 sources[4], we modeled their soft X-ray jointly with their *Swift*-BAT spectra. *XMM-Newton* data were available for 18 sources, while *Chandra* data were available for an additional 24. For the remaining 48 sources, the soft X-ray data were provided by *Swift*-XRT. As the greater part of the sources in the sample were unobscured ($\text{Log}(N_{\rm H}) < 22$) or mildly obscured ($22 < \text{Log}(N_{\rm H}) < 23$), they were sufficiently modeled with the following absorbed powerlaw:

$$\text{Model1} = \text{constant}_1 * \text{phabs} * (z\text{phabs} * z\text{powerlw}). \tag{1}$$

However, Compton-thin ($23 < \text{Log}(N_{\rm H}) < 24$) sources required a more complex model to account for the Fe Kα emission and the fraction of intrinsic emission that leaks through the torus rather than being absorbed by the obscuring material. These sources were modeled as follows:

$$\begin{aligned}\text{Model2} = \text{constant}_1 * \text{phabs} * (&z\text{phabs} * z\text{powerlw}\\ &+ z\text{gauss} + \text{constant}_2 * z\text{powerlw}),\end{aligned} \tag{2}$$

where constant$_1$ accounts for cross-normalization differences between the soft X-ray instrument and *Swift*-BAT, phabs models the galactic absorption, $z$phabs ∗ $z$powerlw is the absorbed power-law modeling the intrinsic emission, $z$gauss models the Fe Kα emission line, and constant$_2$ ∗ $z$powerlw represents the scattered emission that leaks through the torus.

When sources approach or surpass the CT limit, they require even more sophisticated modeling. These sources were modeled with physically motivated models such as `MYTorus` (Murphy & Yaqoob 2009) and `borus02` (Baloković et al. 2018), which have been described in detail in for example, Zhao et al. (2019a,b), Torres-Albà et al. (2021), Silver et al. (2022). These models are used for heavily obscured AGN because they account for the photons that interact with the dust and gas surrounding the SMBH and are reflected into the observer line of sight.

4. Machine learning

4.1. Multiple linear regression

Linear regression is one of the most commonly used machine learning techniques (see, e.g., Chen et al. 2021; Mizukoshi et al. 2022). Simply put, linear regression models the linear relationship between an explanatory variable (input parameter) and the response variable (output parameter). Since few quantities can be accurately modeled using only one explanatory variable, using numerous can improve the predictive capability of an algorithm. This is referred to as multiple linear regression, or just multiple regression for short, and is modeled as:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n, \tag{3}$$

where y_i is the response variable for every source in the sample, x_n are the explanatory variables, β_0 is the y-intercept (if necessary), and β_n are the slope coefficients corresponding to each explanatory variable. Using a large sample of sources with data for every explanatory variable and a known value for the response variable, the algorithm trains itself according to the ordinary least squares method. This method iterates through many combinations of β_n to find that which minimizes the sum of squared errors as:

$$\sum_{i=1}^{i=m} = (wX - y_i)^2, \tag{4}$$

where m is the number of sources in the training sample, w is the vector of slope coefficients, X is the matrix of explanatory variables, and y_i is the known response variable for each source.

Out of our sample of 451 sources, 80% were randomly selected to be used in our training sample, thus leaving 20% (91 sources) left for our test sample. We determined this was the optimal ratio as it used enough sources to accurately train the

[1] https://science.clemson.edu/ctagn/bat-150-month-catalog/

[2] The online version of the catalog can be found at: https://science.clemson.edu/ctagn/bat-150-month-catalog/

[3] i.e., where the uncertainties at a 90% confidence level are smaller than the $N_{\rm H}$ measurement.

[4] These sources can be found in the BAT 105-month catalog (Oh et al. 2018).

algorithm while simultaneously leaving a statistically significant sample to verify this accuracy[5]. We performed Kolmogorov–Smirnov tests (K–S tests; Karson 1968) for every parameter used in the training and test set and find that, to a 95% confidence level, they originate from the same sample. Thus, our training set is representative of the testing sample.

4.2. Parameters used

The algorithm will only be as accurate as however strong the relationship is between the chosen input parameters and the desired output parameter. We selected WISE colors, BAT count rates, soft X-ray hardness ratios (HRs), and an MIR–soft X-ray flux ratio as all have been previously shown to correlate with N_H. These parameters are described below. Their relationship with N_H is illustrated for the sources in this sample in Appendix A.

4.2.1. MIR colors

Roughly half of the intrinsic emission from the AGN is absorbed by the dusty torus (see, e.g., Almeida & Ricci 2017; Hönig 2019). As a consequence, the dust present in the torus is heated to temperatures of several hundred Kelvin, and thus radiates thermally. This emission peaks in the MIR (∼3−30 µm) and is much less prone to absorption than the optical and UV, making it a crucial tool to study obscured AGN. With the launch of WISE, we have an all-sky instrument with superb resolution (∼6″) capable of studying these obscured sources. WISE observes the entire sky in four bands, 3.4, 4.6, 12, and 22 µm ($W1$, $W2$, $W3$, and $W4$, respectively), and to date has detected nearly 750 million sources and reached flux limits of $7.1 \times 10^{-14}\,\mathrm{erg\,cm^{-2}\,s^{-1}}$. The differences between these bands have been proven to be a good predictor for different levels of obscuration. Kilerci Eser et al. (2020) used a sample of AGN from the BAT 105 Month catalog (Oh et al. 2018) to create new CT-AGN selection criteria based on MIR colors. They find that the median values for different colors have an increasing trend with N_H (see Fig. 9 of their paper). Therefore, our algorithm includes six WISE colors: $W1 - W2$, $W1 - W3$, $W1 - W4$, $W2 - W3$, $W2 - W4$, and $W3 - W4$.

4.2.2. MIR–X-ray flux ratio

As the X-ray and MIR emission are both reprocessed from the same material, it is expected that a correlation exists between them. Asmus et al. (2015) show the trend between the observed 12 µm flux and the 2−10 keV flux (see Fig. 1). Moreover, a shift is evident in the trend based on obscuration. As source obscuration increases, the observed 2−10 keV flux decreases, thus causing the source to fall to the left of the predicted trend. This is evidenced in the figure by Seyfert 2 galaxies (red squares) falling to the left of traditionally unobscured Seyfert 1 galaxies (blue circles). Moreover, the confirmed CT-AGN (black stars) fall well to the left of even Seyfert 2 galaxies, suggesting an extremely suppressed observed X-ray flux. As a result of this trend, Asmus et al. (2015) used the log ratio of the 12 µm flux density and the 2−10 keV flux to predict the column density of an AGN. We included this parameter in our algorithm, using the 12 µm flux density measurement from WISE and the 2−10 kev flux from *Swift*-XRT.

4.2.3. Soft X-ray hardness ratios

Soft X-rays (0.3−10 keV) are very susceptible to changes in column density, as evidenced in Fig. 2. It can be seen that the 0.3−10 keV emission is far more suppressed in a source with $\mathrm{Log}(N_\mathrm{H}) = 24$ compared to a source with $\mathrm{Log}(N_\mathrm{H}) = 23$. Therefore, the ratio between the counts in different energy bands, or HRs, covering this energy band are highly dependent on column density. For this reason, we included two HRs from the latest *Swift*-XRT point source catalog, the 2SXPS (Evans et al. 2020); $(M - S)/(M + S)$ and $(H - M)/(H + M)$ where S, M, and H correspond to the 0.3−1, 1−2, and 2−10 keV bands.

4.2.4. Hard X-ray count rates

While significantly less affected than soft X-rays, hard X-rays do display an increased curvature with higher column densities. Koss et al. (2016) analyzed sources with *Swift*-BAT data and found a correlation between the spectral curvature and column density. Using simulated data of CT-AGN, they generated the following equation:

$$\mathrm{SC_{BAT}} = \frac{-3.42 \times A - 0.82 \times B + 1.65 \times C + 3.58 \times D}{\mathrm{Total\,Rate}}, \tag{5}$$

where A, B, C, and D refer to the 14−20 keV, 20−24 keV, 24−35 keV, and 35−50 keV bands, while the total rate is the 14−50 keV band. As plotted in Fig. 3, an increase in this value (calculated via Eq. (5)) was linked to an increase in line-of-sight column density. Two different models that calculate N_H are plotted and all agree that the spectral curvature value increases with N_H. However, we note that this method is only valid up to $N_\mathrm{H} = 4 \times 10^{24}\,\mathrm{cm^{-2}}$.

We used this principle to improve our algorithm. Whereas Koss et al. (2016) only included data up to 50 keV, we found our algorithm performed better when including data up to 150 keV. Because of this, we included BAT count rates for nine different energy bands in our algorithm: 14−20 keV, 20−24 keV, 24−34 keV, 34−45 keV, 45−60 keV, 60−85 keV, 85−110 keV, 110−150 keV, and 14−150 keV. Including each band as a parameter accounts for both the curvature in the spectrum while also serving as a proxy for the BAT flux. For this reason, we elected to include every band instead of just the SC value.

4.3. Missing data

A significant problem in machine learning applications is how to deal with sources that do not possess all the required data. For our training set, we followed the procedure of other works and did not include any sources with missing data (Chang et al. 2021; Dainotti et al. 2021; Narendra et al. 2022). However, valuable predictions can still be made on sources with missing data (see Sect. 6). We artificially replaced data in our training sample according to three different methods to see which would provide the most accurate results. First, we replaced data entries with the mean and median values of the sample (see Luo et al. 2020; Joffre et al. 2022). Second, we used the "large negative value" technique (as described in Farrell et al. 2015; Yang et al. 2022). Lastly, we followed Wenzl et al. (2021) and replaced the missing values with a flux upper limit equal to the limit of the

[5] We note that neural networks are another commonly used machine learning algorithm (see, e.g., Finke et al. 2021; Chainakun et al. 2022; Zubovas et al. 2022). We applied this technique to our data set and after finding the optimal configuration, yielded very similar results to those generated by our linear regression model. For this reason, we optioned to present the results from the comparatively simpler linear regression model in this paper.

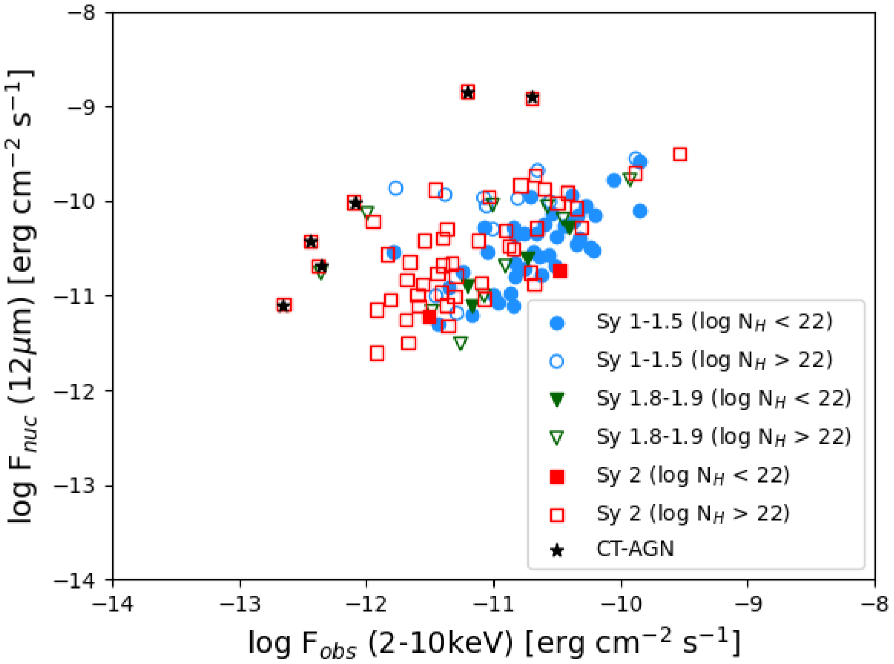

Fig. 1. Relationship between the observed 12 μm flux and the observed 2−10 keV flux, adapted from Fig. 1 of Asmus et al. (2015). The closed shapes represent unobscured AGN (LogN_H < 22) while open shapes represent obscured AGN (LogN_H > 22). The blue circles are Seyfert (Sy) 1−1.5 galaxies; green triangles are Sy 1.8−1.9 galaxies; and red squares are Sy 2 galaxies. The black stars represent confirmed CT-AGN. The obscured sources fall to the left of the trend, signifying that the ratio between these two quantities can be a tracer of obscuration.

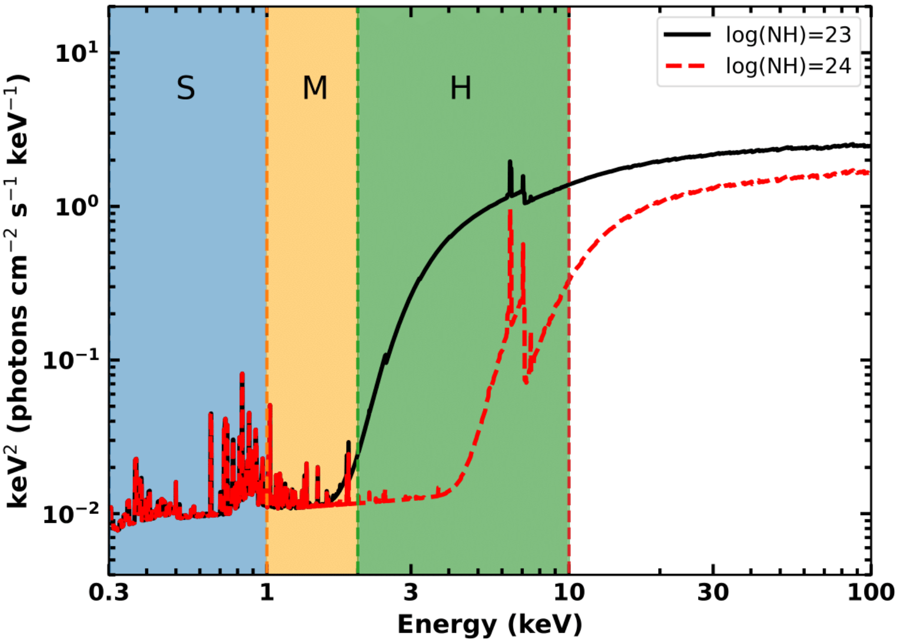

Fig. 2. Simulated X-ray spectra of an AGN with line-of-sight column density Log(N_H) = 23 (solid black line) and Log(N_H) = 24 (dotted red line). The vertical regions denoted by S (blue), M (orange), and H (green) represent the different bands used in our two HRs and they correspond to the 0.3−1, 1−2, and 2−10 keV bands, respectively. The two spectra show extreme differences in the soft X-rays, particularly in the 2−10 keV band. Thus, HRs targeting this band are helpful in determining the column density of AGN.

instrument. We found this last method to be the most accurate. Indeed, if sources are not detected by all-sky instruments such as WISE or BAT, it is because they are below the flux limit and thus replacing them with median values of the sample will not be indicative of their intrinsic properties. Therefore, we treated each case as such:

- If a source is missing BAT data, one should use the count rates of the source (4PBC J1022.1+5123) in the sample of 451 with the lowest total count rate in the 14−150 keV band.
- If a source is missing WISE data, one should use the median colors of the 10 dimmest sources in our sample, all with $W1$ > 14 (1.064, 4.2725, 6.7995, 3.1725, 5.825, 2.6835).

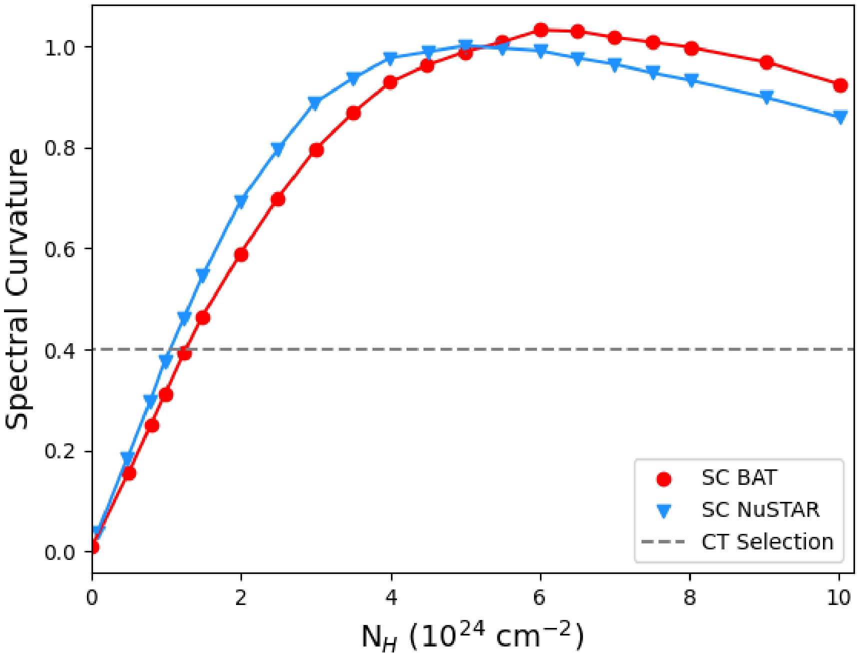

Fig. 3. Spectral curvature value based on different column densities, displaying some of the configurations adapted from Fig. 3 of Koss et al. (2016). Each curve represents how the SC value changes with N_H based on different input parameters used in the `MYTorus` model simulations. The red circles show the curve when the SC equation is calibrated to *Swift*-BAT data, while the blue triangles show the different SC equation when calibrated to *NuSTAR* data. The dashed gray line indicates the cutoff for CT-AGN determined by Koss et al. (2016). Both lines illustrate how the increased curvature of the hard X-rays of AGN is related to an increase in column density.

- If a source is missing XRT data, one should search for counterparts detected by either *Chandra* or *XMM-Newton*. Use the 2–10 keV fluxes and listed HRs that most closely overlap with the bands used by XRT[6]. It should be noted that these replacements can produce a reasonable prediction for N_H; however, it is less accurate than what is possible with XRT data as that is what was used to train the algorithm[7]. If neither *Chandra* nor *XMM-Newton* are available, then reliable N_H measurements cannot be found as the soft X-rays were found to be the most valuable parameter (see Sect. 5).

This procedure was applied to missing data in the sample described in Sect. 6.

5. Results

Figure 4 shows the X-ray-confirmed N_H values for the 91 sources plotted against the predictions by our algorithm (blue circles). We used the Spearman rank correlation coefficient to measure the strength of the correlation between the two sets of N_H values. Our algorithm yielded a Spearman coefficient of 0.86, signifying that it performs very well in recreating the true N_H values of these sources. Moreover, of the 31 heavily obscured ($\mathrm{Log}(N_\mathrm{H}) \geq 23$) sources in our test sample, our algorithm correctly predicted 25 of them (80%) with $\mathrm{Log}(N_\mathrm{H}) > 23$ and 30 (97%) with $\mathrm{Log}(N_\mathrm{H}) > 22.80$. We note that we are currently unable to distinguish this obscuration as being caused by the nuclei or the host galaxy, particularly for edge-on galaxies.

In order to determine which input parameters were most impactfull in training our algorithm, we used the percent difference of the Spearman correlation coefficient when all parameters were used (0.86) and the coefficient when only the parameter listed was excluded. The larger this difference, the worse our algorithm performed without including said input parameter. Since removing one WISE color or BAT count rate had little effect, we grouped the parameters as follows: all six WISE colors; WISE colors + the MIR-X-ray flux ratio (MIR); the MIR-X-ray flux ratio; the two XRT HRs; the two HRs + the MIR-X-ray flux ratio (Soft X-ray); and the BAT count rates. Since the MIR-X-ray flux ratio includes both information from the infrared and the X-rays, we included separate categories without it to determine which wavelength influenced our algorithm the most. Figure 5 shows that the three parameters using soft X-ray data (the two XRT HRs and the MIR-X-ray flux ratio) were the largest contributors toward our algorithm producing accurate results. We note that while the BAT count rates appear to have a negligible impact on the entire sample, removing the BAT information results in much lower accuracy at the highly-obscured and CT ranges.

5.1. Comparison with previous methods

5.1.1. Asmus et al. (2015)

Utilizing MIR data alongside soft X-ray data of a sample of 152 AGN, Asmus et al. (2015) determined a relation to predict line-of-sight column densities. The relation is as follows:

$$\log\left(\frac{N_\mathrm{H}}{\mathrm{cm}^{-2}}\right) = (14.37 \pm 0.11) + (0.67 \pm 0.11) \times \log\left(\frac{F^\mathrm{nuc}(12\,\mu\mathrm{m})}{F^\mathrm{obs}(2{-}10\,\mathrm{keV})}\frac{\mathrm{erg\,s^{-1}\,cm^{-2}}}{\mathrm{mJy}}\right). \quad (6)$$

Using the WISE 12 μm and XRT 2–10 keV fluxes, we plotted the N_H values predicted by the Asmus relation for our test sample of

[6] We performed simulations to convert the HRs from *Chandra* and *XMM-Newton* to the equivalent value expected in the XRT bands and found no improvement in the results.

[7] We plan to calibrate this algorithm to *Chandra* and *XMM-Newton* data in the future to improve our algorithm's predictive capabilities.

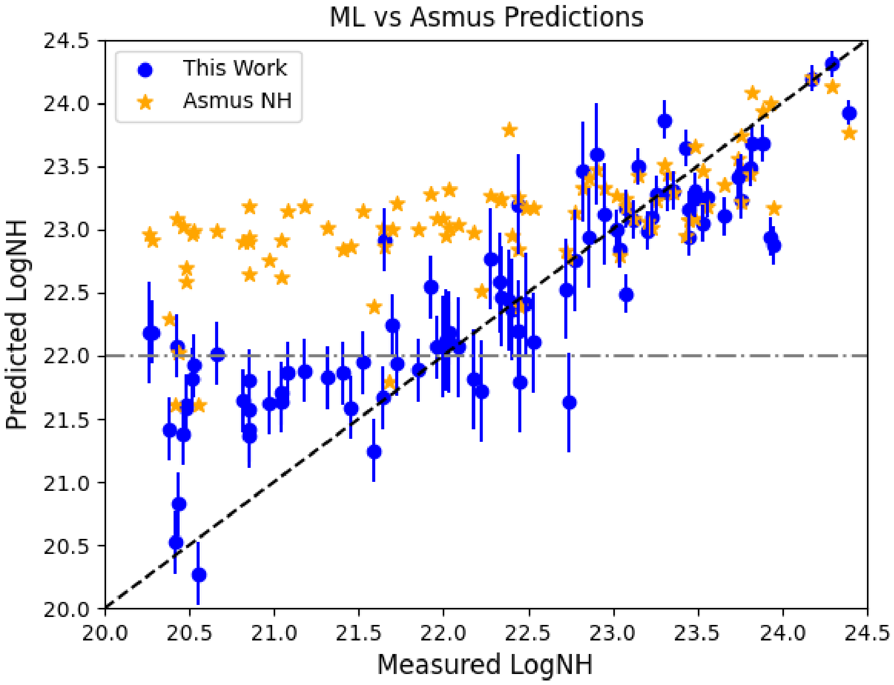

Fig. 4. ML vs. previous methods. The x-axis shows the "true" line-of-sight Log(N_{H}) values, as determined by spectral fitting. The y-axis shows the Log(N_{H}) values predicted by our machine learning algorithm (blue circles) and those predicted by the Asmus et al. (2015) equation (orange stars). Our algorithm shows superior predictive capabilities, particularly for lower levels of obscuration (Log(N_{H}) < 23), where our algorithm does not incorrectly classify unobscured sources as heavily obscured as displayed by the dash-dotted gray line. The dotted black line represents the one-to-one ratio between the true and predicted N_{H} values. The uncertainties were calculated statistically based on the different classifications listed in Table 1. We determine the error that needs to be added or subtracted to the predicted NH values in order to achieve a 90% classification accuracy in each bin. Thus, each of the four classification bins have different uncertainties. No errors are included on the orange points for readability purposes.

91 sources in Fig. 4. While these results show a good trend for heavily obscured sources, below Log(N_{H}) = 23, our machine learning algorithm performs far better. This is quantified by the lower Spearman correlation coefficient of 0.65 for the Asmus predictions and the real N_{H} values. The lack of predictive capability below 10^{23} cm^{-2} affects the whole range of possible N_{H} values. This is because, a priori, one does not know the "true" N_{H} of the source, and if choosing a source with Log(N_{H}) < 23, the relation will confidently place it as being heavily obscured. Therefore, sources with Log(N_{H}) < 23 can actually have any value of true Log(N_{H}) between 20 and 23.

5.1.2. Pfeifle et al. (2022)

Pfeifle et al. (2022) improved upon the work of Asmus et al. (2015) by creating a new relationship based on the ratio of the 2–10 keV and 12 µm luminosities. Using 456 AGN detected in the 70-month BAT catalog (Ricci et al. 2017) that also possess infrared data, their team determined the following relation:

$$\log\left(\frac{N_{\mathrm{H}}}{\mathrm{cm}^{-2}}\right) = 20 + \left(1.61^{+0.33}_{-0.31}\right) \times \log\left(\left|\frac{\log\left(\frac{L_{\mathrm{X,Obs.}}}{L_{12\,\mu\mathrm{m}}}\right) + \left(0.34^{+0.06}_{-0.06}\right)}{\left(-0.003^{+0.002}_{-0.005}\right)}\right|\right). \quad (7)$$

With this relation, we predicted the N_{H} values for our sample of 91 test sources as seen as the magenta triangles in Fig. 6. Pfeifle et al. (2022) claims that their method is most accurate when applied to sources with Log(N_{H}) > 22.5, which is confirmed here. While their method is accurate for heavily obscured sources, it is far less predictive than our algorithm for AGN with Log(N_{H}) < 22.5. Just as with the Asmus relation, this represents a significant drawback when selecting sources as we do not know whether or not the true Log(N_{H}) > 22.5. In total, it has a Spearman correlation coefficient of 0.27.

We note, however, that the Asmus and Pfeifle relations and the one presented here have different advantages and disadvantages, which make them extremely complementary. While our algorithm performs better overall, particularly in differentiating between obscured and unobscured sources, the Asmus et al. (2015) and Pfeifle et al. (2022) relations have an advantage in that they only require one parameter each. This inherently makes them more applicable to a larger sample of sources.

5.1.3. Koss et al. (2016)

Koss et al. (2016) developed a method for identifying new CT-AGN using weighted averages of different *Swift*-BAT bands. It was determined that an $\mathrm{SC_{BAT}} > 0.40$ would identify a CT-AGN candidate. We applied this formula to our 91 test sources and found 14 that would be considered CT. These sources are plotted as red squares in Fig. 7, overlapped on our machine learning predictions. We note that this method does show promise, as 8 of the 14 sources are heavily obscured, with Log(N_{H}) > 23. However, 6 sources (43%) predicted as CT have true Log(N_{H}) < 23, including two that are unobscured (Log(N_{H}) < 22). Additionally, of the 14 predicted as CT, only 2 (14%) truly are. Our machine learning algorithm does not misclassify any unobscured sources as CT and performs more accurately throughout all column density ranges. Moreover, both sources predicted as CT by our algorithm are truly CT.

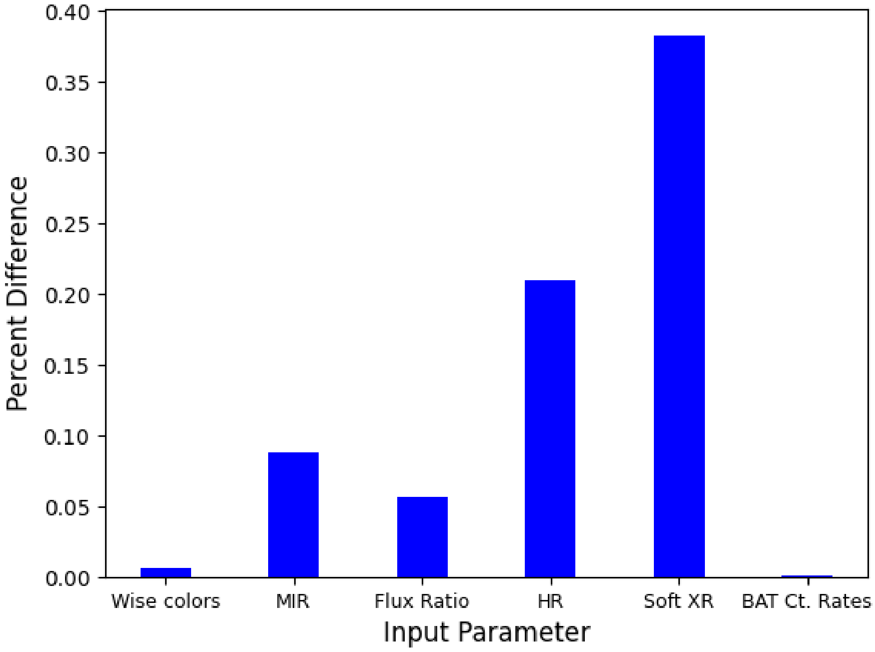

Fig. 5. Percent difference between the Spearman correlation coefficient including all parameters and the coefficient when the listed parameter is excluded. The larger the difference, the worse the fit without that parameter (i.e. the higher the importance of that parameter). Therefore, the soft X-ray-related parameters are the highest contributors to the predictive capability of the algorithm. "MIR" refers to the WISE colors and the MIR-X-ray flux ratio. "HR" represents the two X-ray HRs. "Soft XR" refers to the two X-ray HRs and the MIR-X-ray flux ratio.

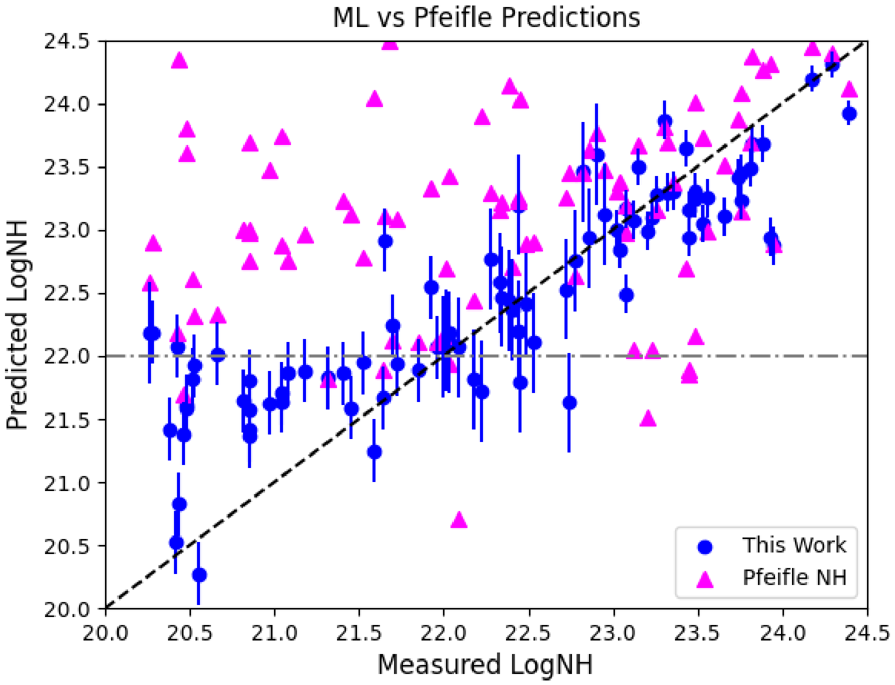

Fig. 6. ML vs. previous methods. As in Fig. 4, the x-axis shows the true line-of-sight Log($N_{\rm H}$) values determined by spectral fitting while the y-axis shows the Log($N_{\rm H}$) values predicted by our machine learning algorithm (blue circles) and by the Pfeifle et al. (2022) relation (magenta triangles). The errors from our algorithm were calculated statistically as described in the caption of Fig. 4. No errors are included on the magenta points for readability purposes.

5.2. N_H classifications

Table 1 displays how many of the 91 test sources fall into each of these four categories: CT (Log($N_{\rm H}$) $>$ 24), Compton-thin (23 $<$ Log($N_{\rm H}$) $<$ 24), obscured (22 $<$ Log($N_{\rm H}$) $<$ 23), and unobscured (Log($N_{\rm H}$) $<$ 22). As can be seen, our machine learning algorithm performs the best overall, correctly classifying ~75% of the sources. This is particularly true for sources with Log($N_{\rm H}$) $<$ 23, for which our algorithm has a 73% accuracy, while the other two applicable methods are both only 18% accurate.

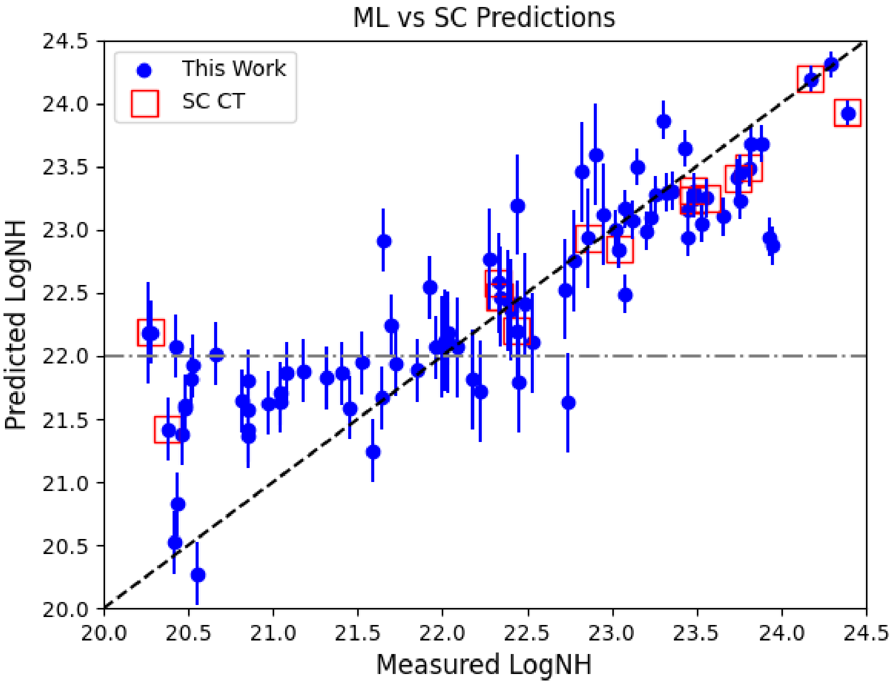

Fig. 7. ML vs. previous methods. As in Fig. 4, the *x*-axis shows the true line-of-sight Log(N_H) values determined via spectral fitting while the *y*-axis shows the Log(N_H) values predicted by our machine learning algorithm (blue circles). The red squares represent the 14 sources that were predicted to be CT based on the SC method introduced in Koss et al. (2016). Two of these sources (14%) have true N_H values $<10^{22}$ cm^{-2}. Our algorithm does not misclassify any unobscured sources as Compton-thin, let alone CT. The errors from our algorithm were calculated statistically as described in the caption of Fig. 4.

Table 1. Classification of the test sources.

Classification	Real number	This work	Asmus et al. (2015)	Pfeifle et al. (2022)	Koss et al. (2016)
Compton-thick	3	2	2	3	2
Compton-thin	28	22	25	13	...
Obscured	24	16	8	7	...
Unobscured	36	28	3	4	...
Total	91	68 (75%)	38 (42%)	27 (30%)	...

Notes. We have split the 91 test sources into four classifications based on their X-ray-measured column densities: CT (Log(N_H) > 24), Compton-thin ($23 <$ Log(N_H) < 24), obscured ($22 <$ Log(N_H) < 23), and unobscured (Log(N_H) < 22). The numbers of sources correctly classified for each of the four methods mentioned in this paper are shown below.

6. Application of the method

The BAT 150 Month catalog has 1390 sources, while only 451 were used to train and test the algorithm. This leaves our algorithm the potential to predict N_H for 939 sources depending on the availability of their data. We found 276 sources with WISE, XRT, and BAT data, and without a confirmed N_H measurement. The histogram of the predicted N_H values is shown in the left panel of Fig. 8. All 6 of the potential CT-AGN have already been accepted for observations by *Chandra* (Proposal ID 23700077, PI: Silver) and *NuSTAR* (Proposal ID 9093, PI: Silver). Additionally, we found 211 sources that lacked only one data set between XRT or WISE. These missing data were handled according to the procedure laid out in Sect. 4.3. The resulting N_H predictions can be seen in the right panel of Fig. 8. We are currently in an ongoing campaign to obtain soft X-ray data for the remaining 452 AGN from the BAT 150 Month catalog.

7. Conclusions

In this work, we present a new machine learning algorithm that predicts the line-of-sight column density of AGN, thus enabling us to discover new CT-AGN candidates. Using MIR data from WISE, soft X-ray data from *Swift*-XRT and hard X-ray data from *Swift*-BAT, our machine learning algorithm has proven its ability to accurately reproduce the N_H values of our 91-source test sample, correctly classifying 75% of sources based on their obscuration. Moreover, our algorithm has shown a superior ability to predict the column density of AGN with Log(N_H) $<$ 22.5 when compared with previously published methods. In the future, this algorithm will be used to: (1) identify promising CT-AGN candidates and (2) efficiently determine N_H values of large samples of sources (such as the *Chandra* and *XMM-Newton* source catalogs) in an effort to determine the obscuration distribution of the entire AGN population across cosmic time.

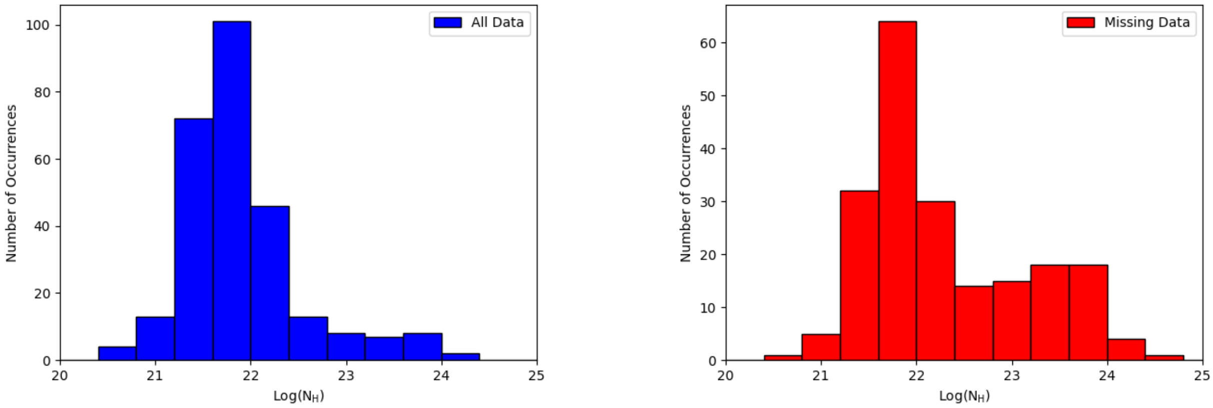

Fig. 8. Left: N_H predictions for the 276 sources in the BAT 150 Month catalog that possess all the necessary WISE, XRT, and BAT data. Right: N_H predictions for the 211 sources in the BAT 150 Month catalog that are missing either WISE or XRT data.

Acknowledgements. The authors thank the anonymous referee for their detailed and helpful comments which greatly improved the paper. We acknowledge NASA funding under grant 80NSSC21K0016.

References

Ajello, M., Greiner, J., Sato, G., et al. 2008, ApJ, 689, 666
Alexander, D. M., Bauer, F. E., Brandt, W. N., et al. 2003, AJ, 126, 539
Alexander, D. M., Chary, R. R., Pope, A., et al. 2008, ApJ, 687, 835
Almeida, C., & Ricci, C. 2017, Nat. Astron., 1, 679
Ananna, T. T., Treister, E., Urry, C. M., et al. 2019, ApJ, 871, 240
Andonie, C., Alexander, D. M., Rosario, D., et al. 2022, MNRAS, 517, 2577
Asmus, D., Gandhi, P., Hönig, S. F., Smette, A., & Duschl, W. J. 2015, MNRAS, 454, 766
Baloković, M., Brightman, M., Harrison, F. A., et al. 2018, ApJ, 854, 42
Brandt, W. N., & Yang, G. 2022, Handbook of X-ray and Gamma-ray Astrophysics, eds. C. Bambi &amp; A. Santangelo (Springer Living Reference Work), 78
Burlon, D., Ajello, M., Greiner, J., et al. 2011, ApJ, 728, 58
Chainakun, P., Fongkaew, I., Hancock, S., & Young, A. J. 2022, MNRAS, 513, 648
Chang, Y.-Y., Hsieh, B.-C., Wang, W.-H., et al. 2021, ApJ, 920, 68
Chen, Y., Gu, Q., Fan, J., et al. 2021, ApJ, 913, 93
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, VizieR Online Data Catalog: II/328
Dainotti, M. G., Bogdan, M., Narendra, A., et al. 2021, ApJ, 920, 118
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604
Evans, P. A., Page, K. L., Osborne, J. P., et al. 2020, ApJS, 247, 54
Farrell, S. A., Murphy, T., & Lo, K. K. 2015, ApJ, 813, 28
Ferrarese, L., & Ford, H. 2005, Space Sci. Rev., 116, 523
Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9
Finke, T., Krämer, M., & Manconi, S. 2021, MNRAS, 507, 4061
Fiore, F., Feruglio, C., Shankar, F., et al. 2017, A&A, 601, A143
Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJ, 539, L13
Gehrels, N., Chincarini, G., Giommi, P., et al. 2004, ApJ, 611, 1005
Gilli, R., Comastri, A., & Hasinger, G. 2007, A&A, 463, 79
Haardt, F., & Maraschi, L. 1993, ApJ, 413, 507
Hönig, S. F. 2019, ApJ, 884, 171
Joffre, S., Silver, R., Rajagopal, M., et al. 2022, ApJ, 940, 139
Karson, M. 1968, J. Am. Stat. Assoc., 63, 1047
Kilerci Eser, E., Goto, T., Güver, T., Tuncer, A., & Ataş, O. H. 2020, MNRAS, 494, 5793
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Koss, M. J., Assef, R., Baloković, M., et al. 2016, ApJ, 825, 85
Luo, S., Leung, A. P., Hui, C. Y., & Li, K. L. 2020, MNRAS, 492, 5377
Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, AJ, 115, 2285
Marchesi, S., Ajello, M., Comastri, A., et al. 2017, ApJ, 836, 116
Martín-Navarro, I., Brodie, J. P., Romanowsky, A. J., Ruiz-Lara, T., & van de Ven, G. 2018, Nature, 553, 307
Merloni, A., Bongiorno, A., Bolzonella, M., et al. 2010, ApJ, 708, 137
Merritt, D., & Ferrarese, L. 2001, ApJ, 547, 140
Mizukoshi, S., Minezaki, T., Tsunetsugu, S., et al. 2022, MNRAS, 516, 2876
Murphy, K. D., & Yaqoob, T. 2009, MNRAS, 397, 1549
Narendra, A., Gibson, S. J., Dainotti, M. G., et al. 2022, ApJS, 259, 55
Oh, K., Koss, M., Markwardt, C. B., et al. 2018, ApJS, 235, 4
Pfeifle, R. W., Ricci, C., Boorman, P. G., et al. 2022, ApJS, 261, 3
Ricci, C., Ueda, Y., Koss, M. J., et al. 2015, ApJ, 815, L13
Ricci, C., Trakhtenbrot, B., Koss, M. J., et al. 2017, ApJS, 233, 17
Richstone, D., Ajhar, E. A., Bender, R., et al. 1998, Nature, 385, A14
Silver, R., Torres-Albà, N., Zhao, X., et al. 2022, ApJ, 932, 43
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Torres-Albà, N., Marchesi, S., Zhao, X., et al. 2021, ApJ, 922, 252
Treister, E., Urry, C. M., Chatzichristou, E., et al. 2004, ApJ, 616, 123
Treister, E., Urry, C. M., & Virani, S. 2009, ApJ, 696, 110
Ueda, Y., Akiyama, M., Hasinger, G., Miyaji, T., & Watson, M. G. 2014, ApJ, 786, 104
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
Wenzl, L., Schindler, J.-T., Fan, X., et al. 2021, AJ, 162, 72
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Yang, H., Hare, J., Kargaltsev, O., et al. 2022, ApJ, 941, 104
Zhao, X., Marchesi, S., & Ajello, M. 2019a, ApJ, 871, 182
Zhao, X., Marchesi, S., Ajello, M., et al. 2019b, ApJ, 870, 60
Zubovas, K., Bialopetravičius, J., & Kazlauskaitė, M. 2022, MNRAS, 515, 1705

Appendix A: Input parameter trends with N_H

This section showcases the capability of each of the chosen input parameters to successfully predict line-of-sight column density. Each plot shows a parameter space between two input parameters with a color gradient representing the known N_H, for the 451 AGN in our sample. Here, blue represents higher N_H values while red represents unobscured AGN. We note that we only include a plot for one of the BAT count rates (14–20 keV) as every other count rate shows a similar trend when plotted against HR2.

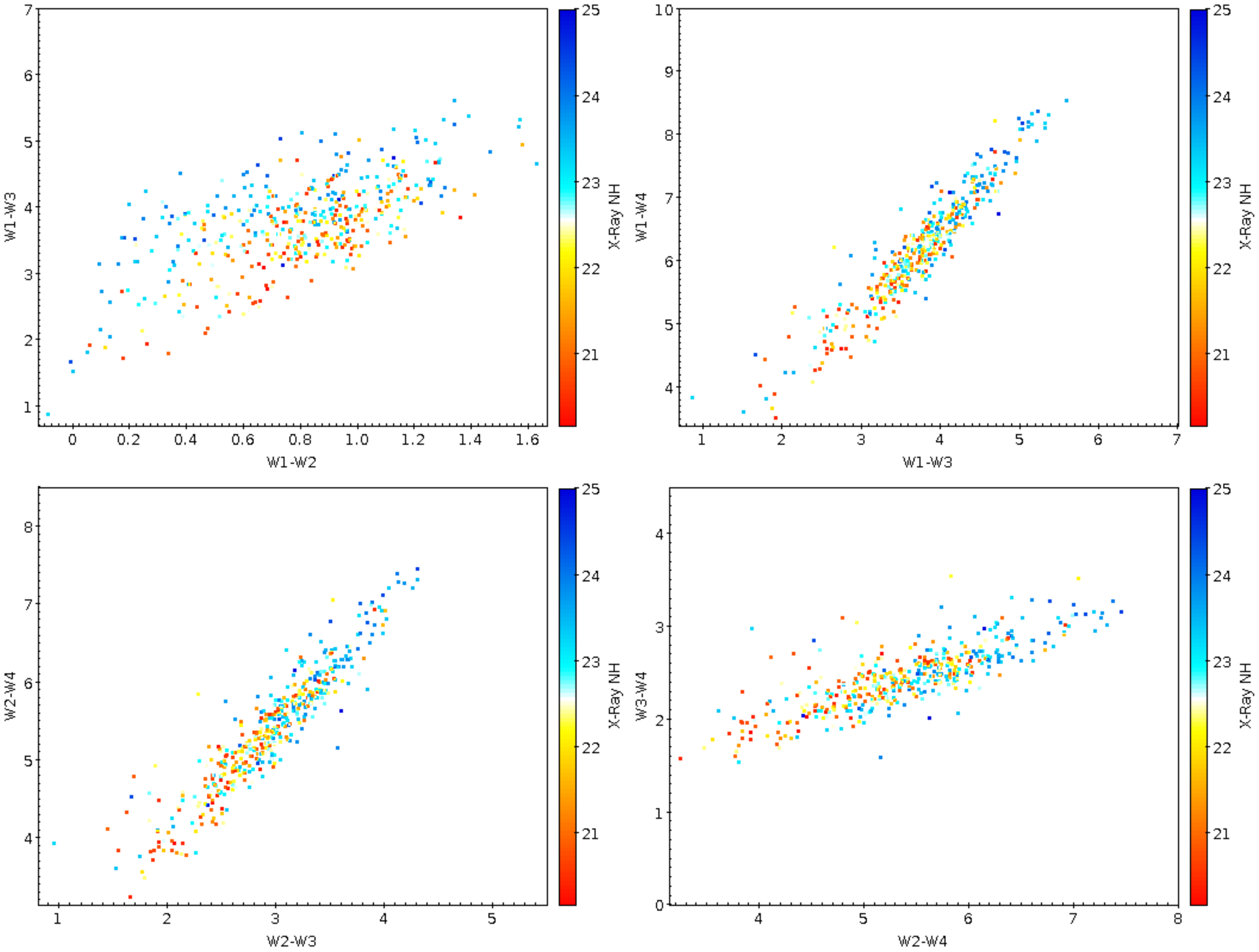

Fig. A.1. WISE color parameter spaces, showing the ability to predict N_H. Blue points represent heavily obscured (and CT) AGN while red points represent unobscured AGN. Top left: W1-W3 vs W1-W2. Top right: W1-W4 vs W1-W3. Bottom left: W2-W4 vs W2-W3. Bottom right: W3-W4 vs W2-W4.

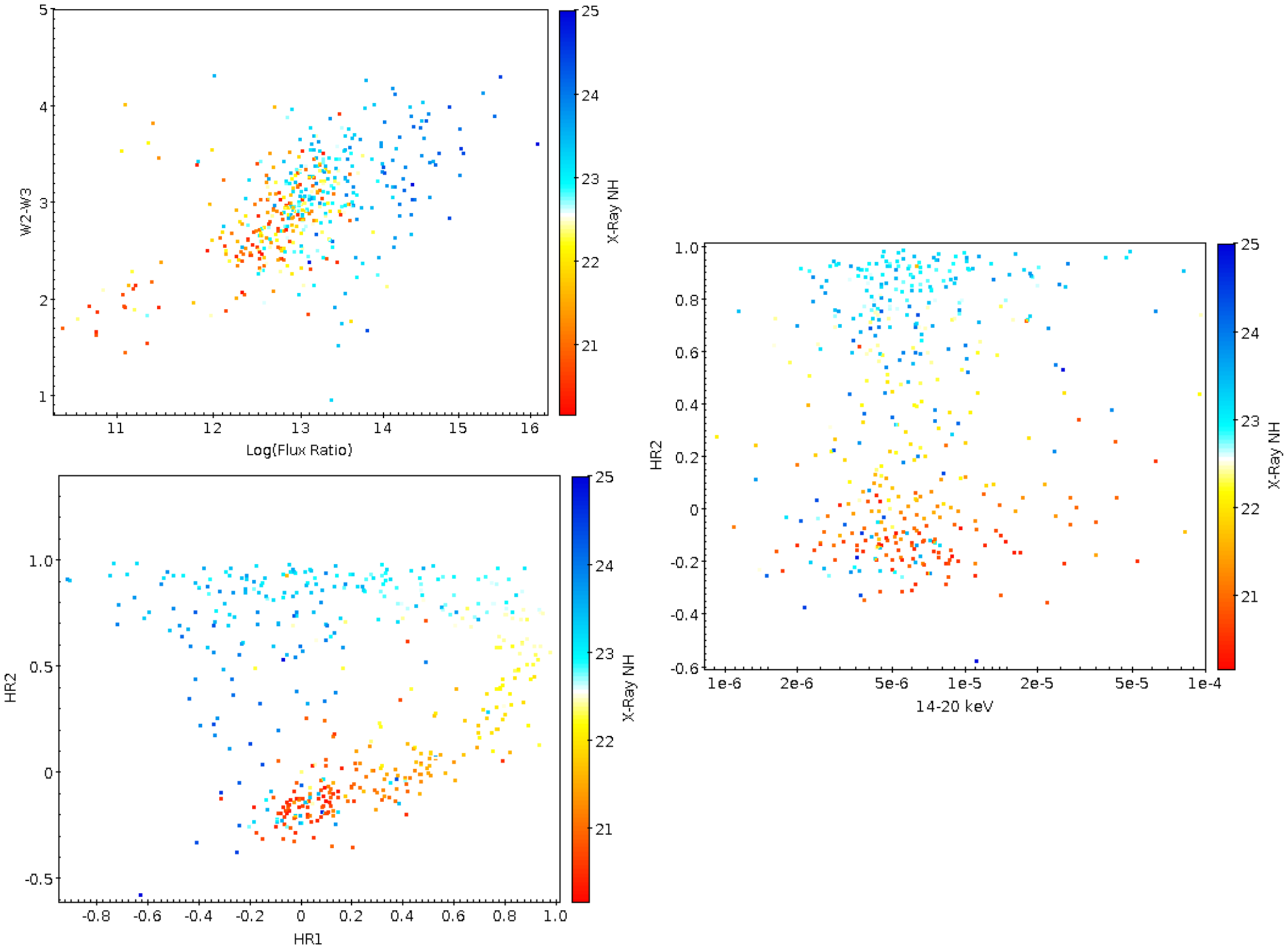

Fig. A.2. Parameter spaces between different input parameters, displaying how they can be used to predict column density. Top left: W2-W3 plotted against the MIR and X-ray flux ratio. Right: HR2 vs the 14–20 keV BAT count rate. Bottom left: HR2 vs HR1.