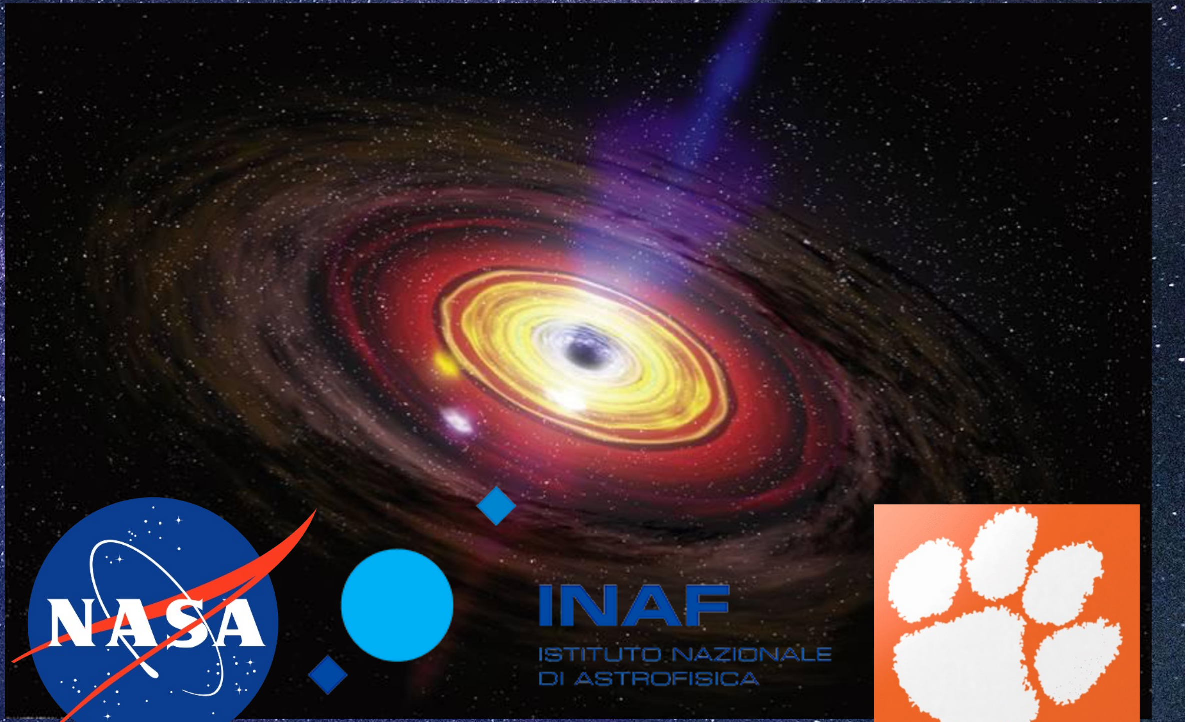
Predicting AGN Obscuration with Machine Learning

By: Ross Silver

NPP Fellow at NASA GSFC
NASA Advisor: Dr. Francesca Civano

Collaborators: Dr. N. Torres-Albà, Dr. X. Zhao, Dr. S. Marchesi, Dr. A. Pizzetti, I. Cox
Clemson Advisor: Dr. Marco Ajello

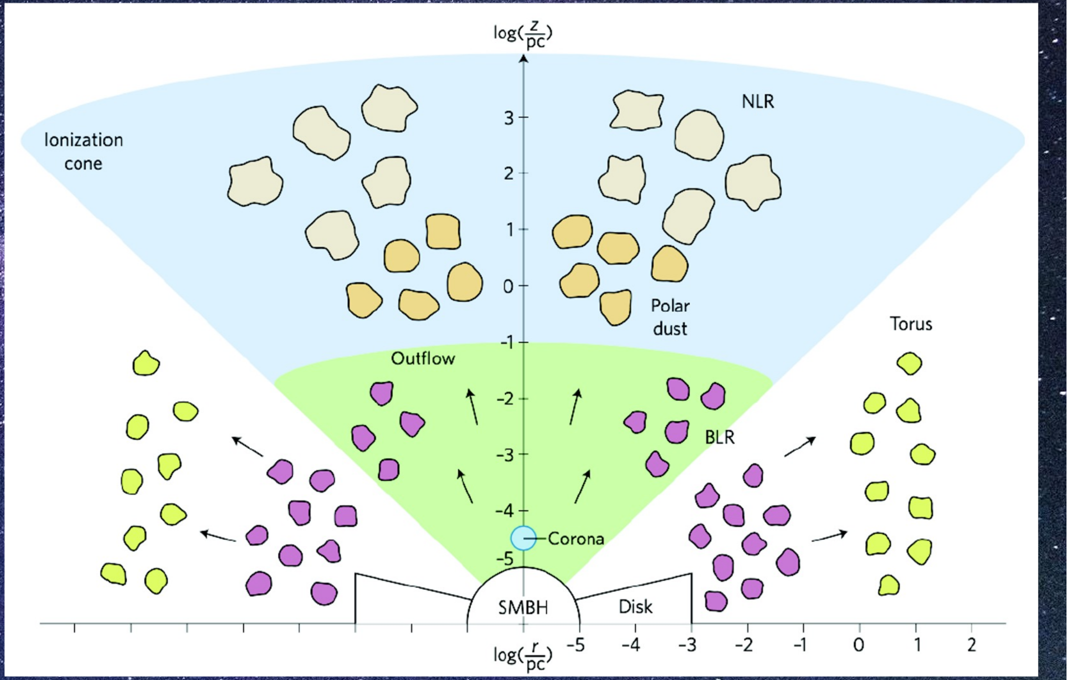
MACHINE LEARNING FOR ASTROPHYSICS
2ND EDITION
CATANIA, 8-12 JULY, 2024



INAF
ISTITUTO NAZIONALE DI ASTROFISICA

What is an AGN?

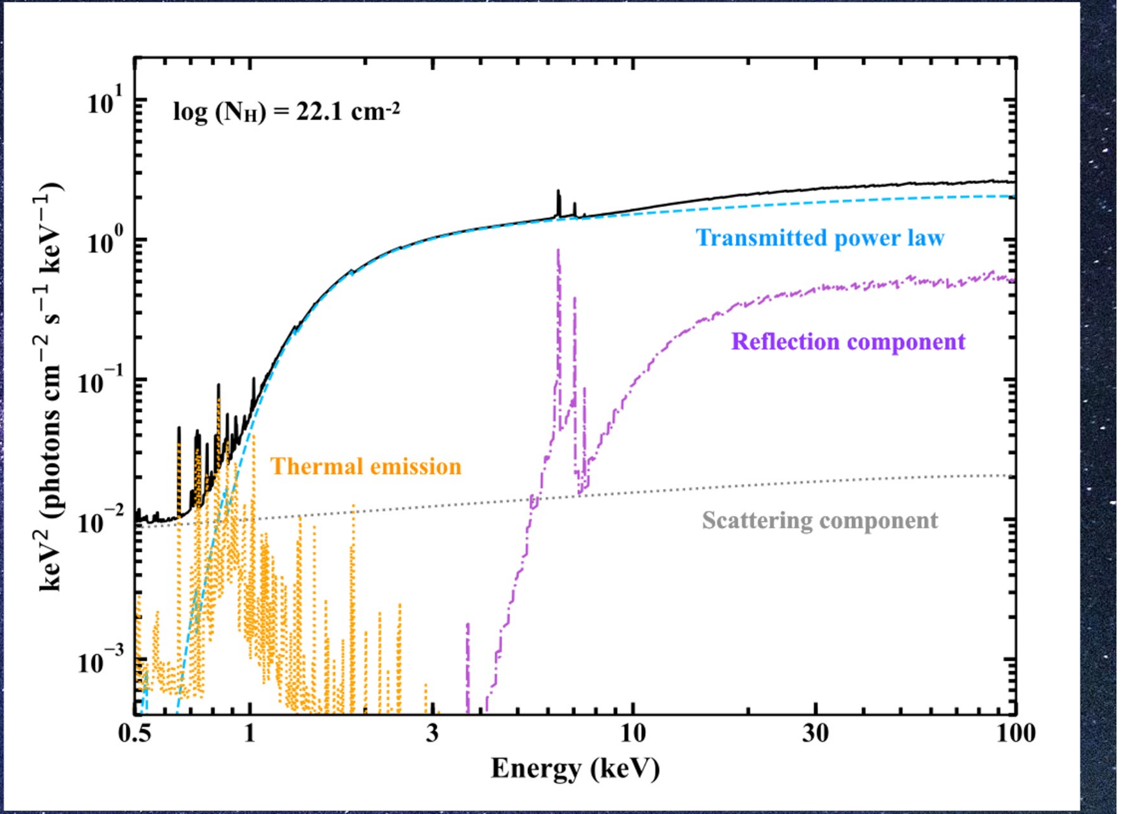
- Active Galactic Nuclei, or AGN, are supermassive blackholes (SMBH) in the center of galaxies that accrete gas onto themselves and emit radiation through the entire EM spectrum.
- Accretion disk: Gas falls onto the SMBH and emits from the IR to UV.
- Corona: Hot plasma of electrons placed near the accretion disk. Photons from the disk collide with the hot electrons and reach X-ray energies via IC scattering.
- Torus: The AGN is surrounded by gas and dust in a toroidal shape that obscures emission from the optical to the X-rays.



Adopted from Ramos-Almeida & Ricci (2017)

Compton Thick (CT)-AGN

- The shape of the spectra depends on the N_H.
- $N_{H,los} > 1 \times 10^{22}$ cm^{-2} = obscured
- $N_{H,los} > 1 \times 10^{23}$ cm^{-2} = heavily obscured
- $N_{H,los} > 1 \times 10^{24}$ cm^{-2} = Compton-thick



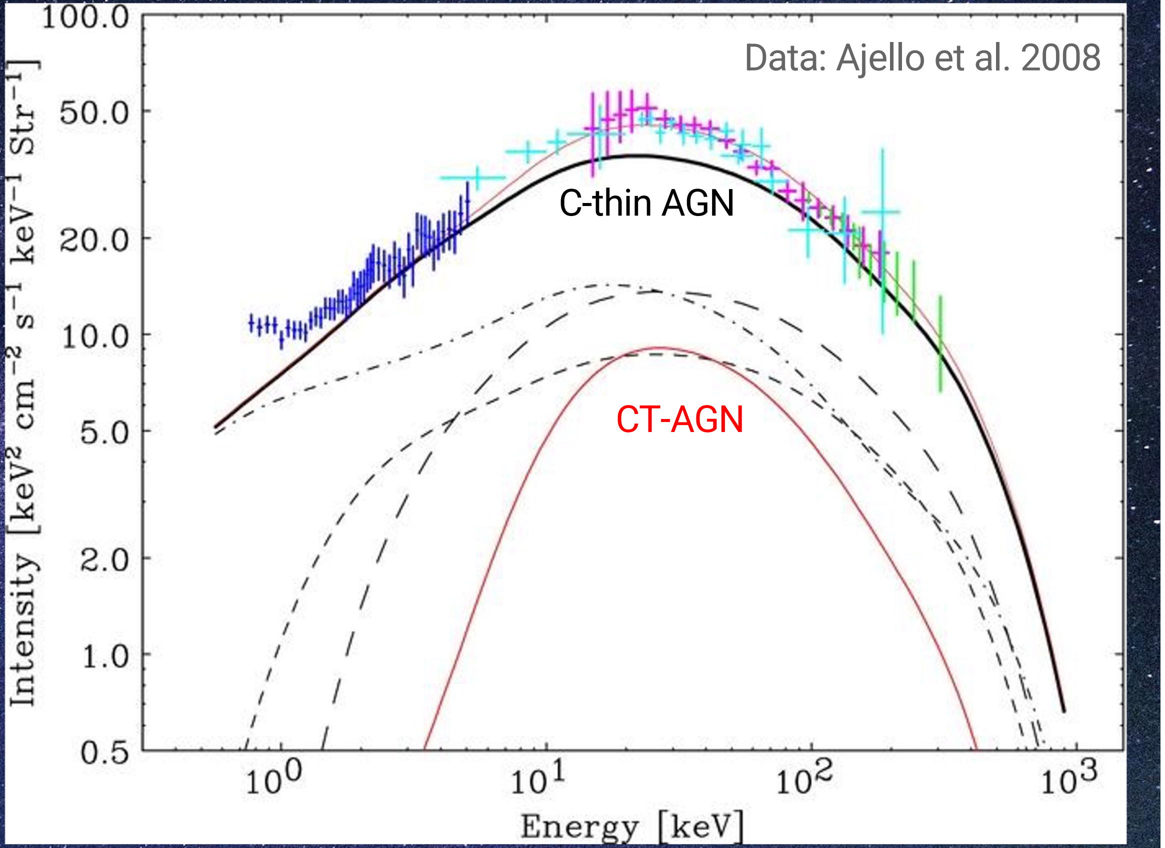
Credit: X.Zhao

Why CT-AGN Are Important

- The Cosmic X-ray Background (CXB), the diffuse X-ray emission in the 1 ~200-300 keV band, is produced mostly by AGN.
- Compton-thin AGN:

  $1 \times 10^{22}\ \mathrm{cm}^{-2} < N_{H,los} < 1 \times 10^{24}\ \mathrm{cm}^{-2}$

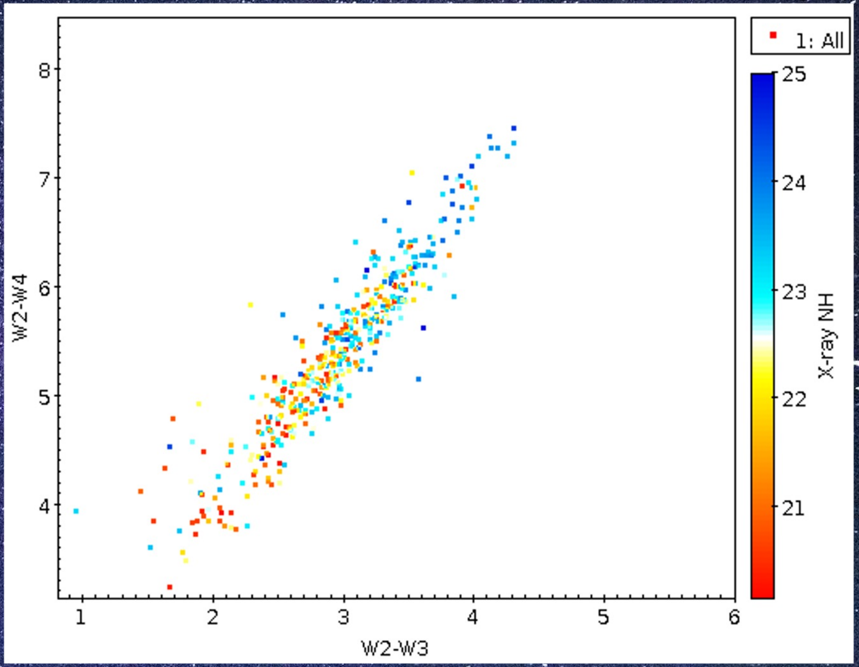
- CT-AGN are required to explain the peak ~30 keV.



Ueda et al. 2014

How Can We Measure NH? With Machine Learning!

- We used a Multiple Linear Regression method.
- The algorithm was trained using 451 AGN detected by the hard X-ray telescope *Swift*-BAT (14-150 keV) and with NH values determined through spectral fitting.
- Mid-Infrared (MIR, 3.4-22 μm): WISE Colors
- MIR - Soft X-ray Relation
- Soft X-rays (0.3-10 keV): Two Hardness Ratios (HRs)
- Hard X-rays (14-150 keV): *Swift*-BAT count rates



W2-W4 VS W2-W3

Results and Comparison with a Previous Method

- Asmus et al. 2015 (orange)

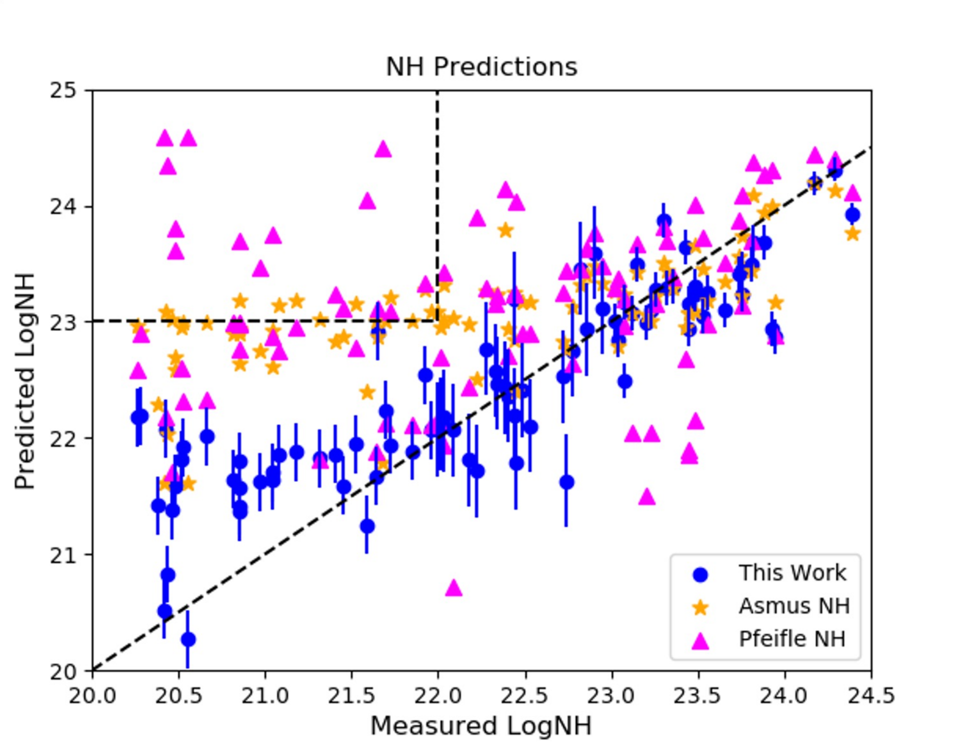
$$\log\left(\frac{N_{\mathrm{H}}}{\mathrm{cm}^{-2}}\right) = (14.37 \pm 0.11) + (0.67 \pm 0.11)\log\left(\frac{F^{\mathrm{nuc}}(12\,\mu\mathrm{m})}{F^{\mathrm{obs}}(2-10\,\mathrm{keV})}\,\frac{\mathrm{erg\,s^{-1}/cm^2}}{\mathrm{mJy}}\right).$$

- Pfeifle et al. 2022 (magenta)

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

- False positives = Real NH < 22, Predicted NH > 23

- My algorithm: 0
- Asmus: 12
- Pfeifle: 15



Future Work

- Analyze 11 sources predicted to be CT by this algorithm that were accepted in NuSTAR GO Cycles 9 and 10.
- Increase training sample with sources observed by NuSTAR (hard X-rays)
- Test different machine learning techniques to see if we can improve the predictive power of the algorithm.

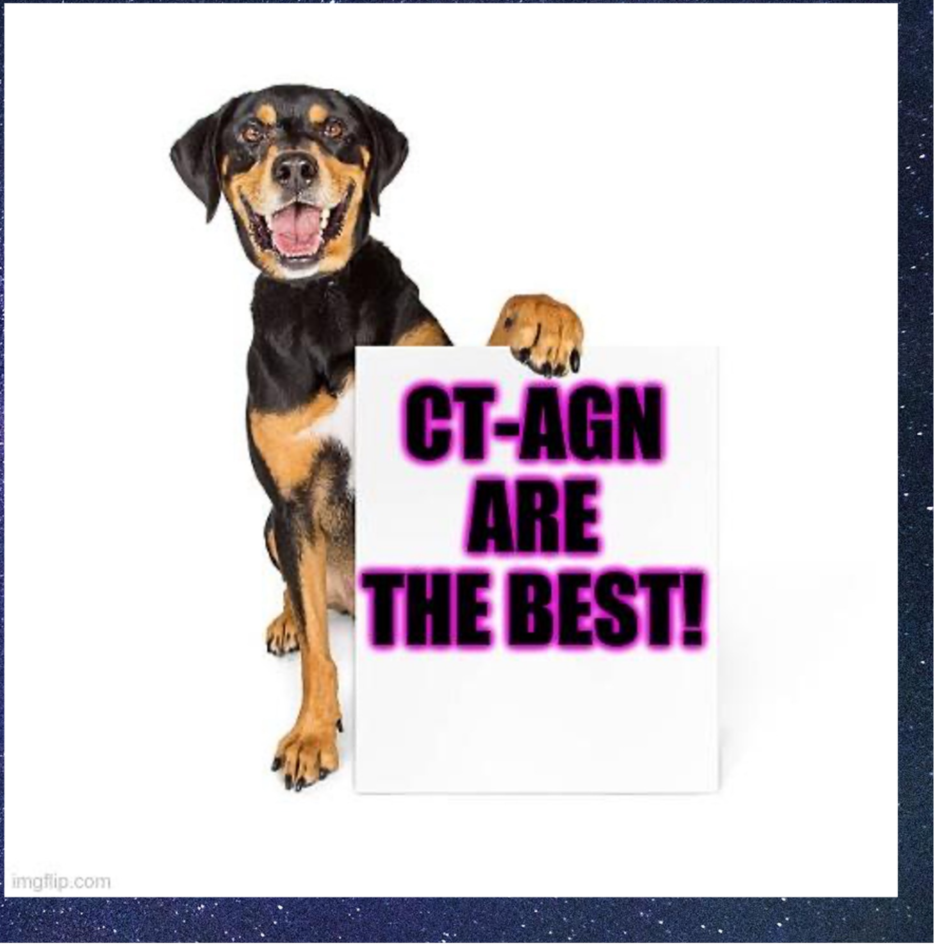
Thank You for Listening!

The Clemson-INAF team page

My Paper

Take-home Points

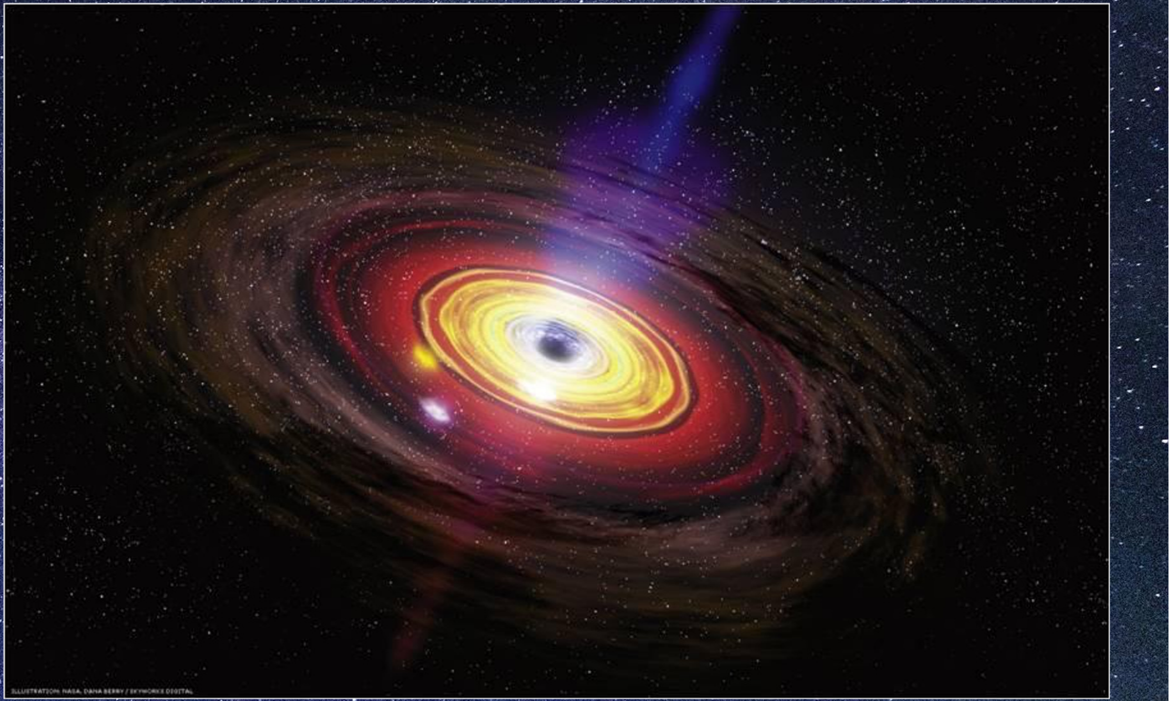
- CT-AGN are crucial to understanding the CXB and are believed to make up a much larger fraction of AGN than those found from observations.
- Our new algorithm has improved upon previous methods in predicting N_H of AGN, particularly for unobscured sources.
- This will help us find new CT AGN in the future.



Extra Slides

What is an AGN?

- Active Galactic Nuclei, or AGN, are supermassive blackholes (SMBH) in the center of galaxies that accrete gas onto themselves and emit radiation through the entire EM spectrum.



Simulation from NASA’s Dana Berry

AGN Emission in the MIR

- UV light from the accretion disk is absorbed by the dust grains in the torus, which are heated up to ~300 K.
- The dust then emits thermally, peaking in the MIR.

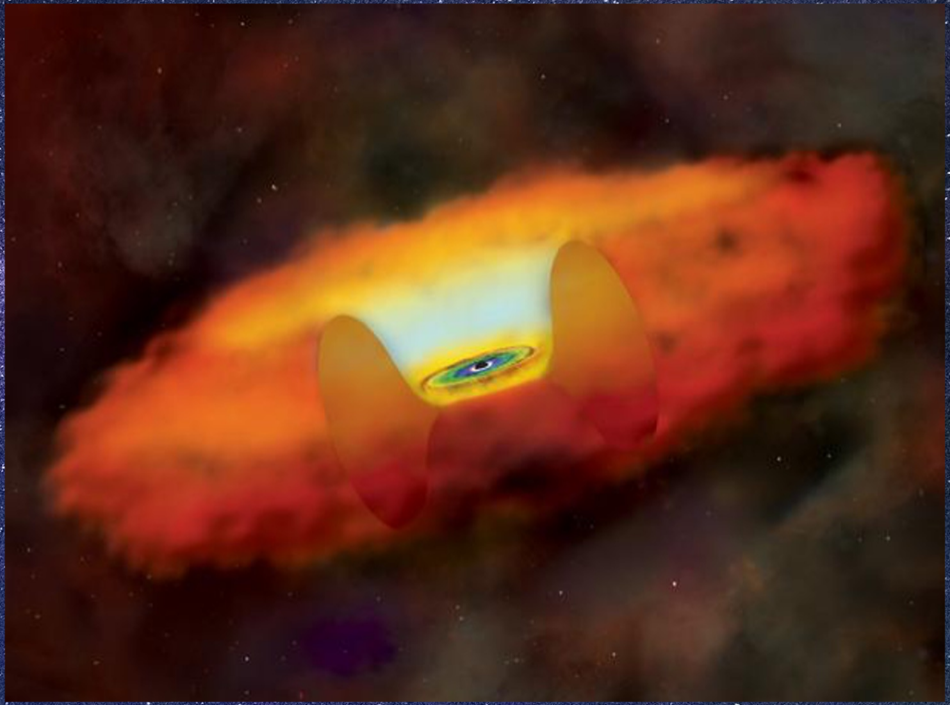
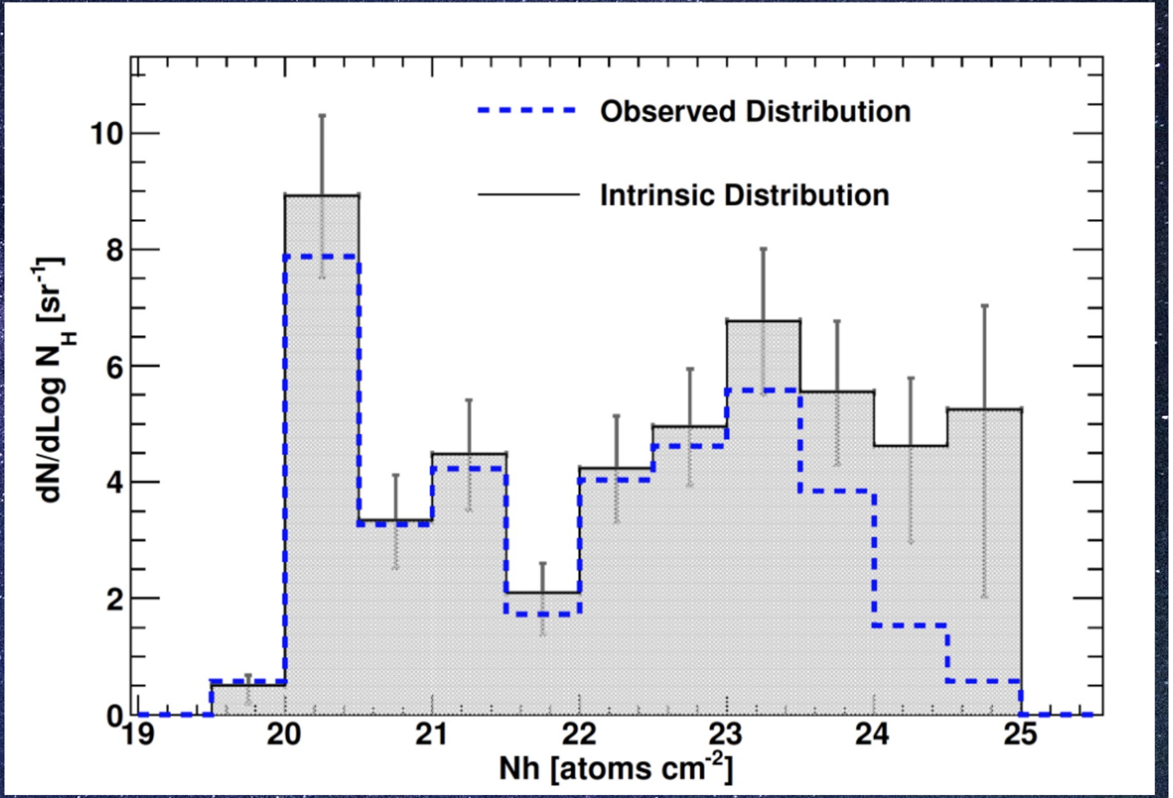


Illustration: NASA/CXC/M.Weiss

Where are they?

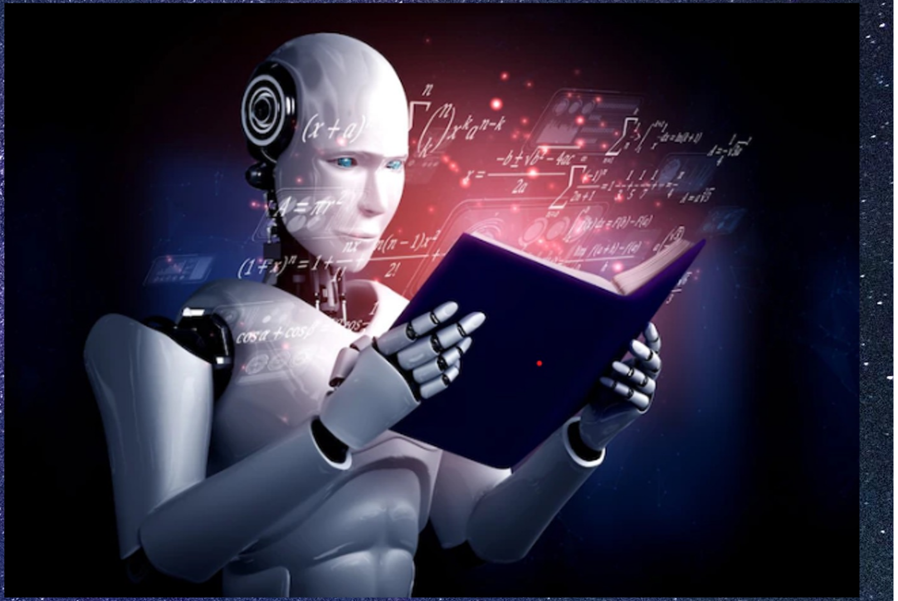
- Population synthesis models predict between 20% (Ueda et al. 2014) and 50% (Ananna et al. 2019) of AGN are CT while only 5-10% of observed AGN are CT (in the local universe).



Burlon et al. 2011

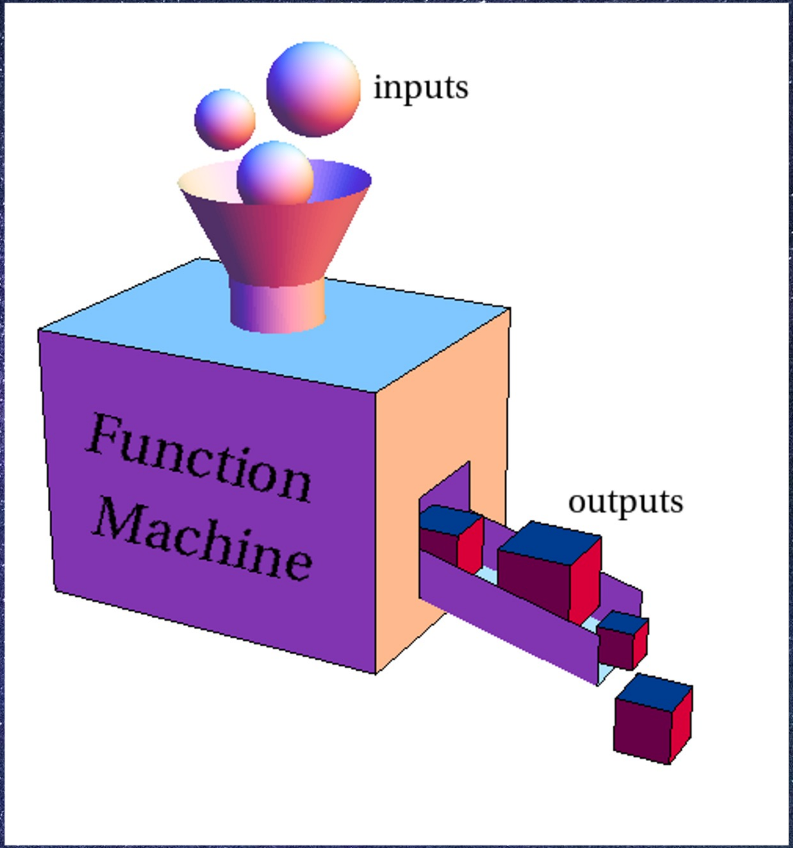
Machine Learning Algorithms: How Do They Work?

- Target Value: Some quantity that you are trying to predict.
- Input Parameters: Data used to predict the target value.
- Training Set: A sample of data with all input parameters and known target values used to teach the algorithm.
- Testing set: A sample of data with all input parameters and known target values used to test the accuracy of the algorithm.



Input Parameters Selected

- Mid-Infrared (MIR, 3.4-22 μm): WISE Colors
- MIR - Soft X-ray Relation
- Soft X-rays (0.3-10 keV): Two Hardness Ratios (HRs)
- Hard X-rays (14-150 keV): *Swift*-BAT count rates



WISE Colors

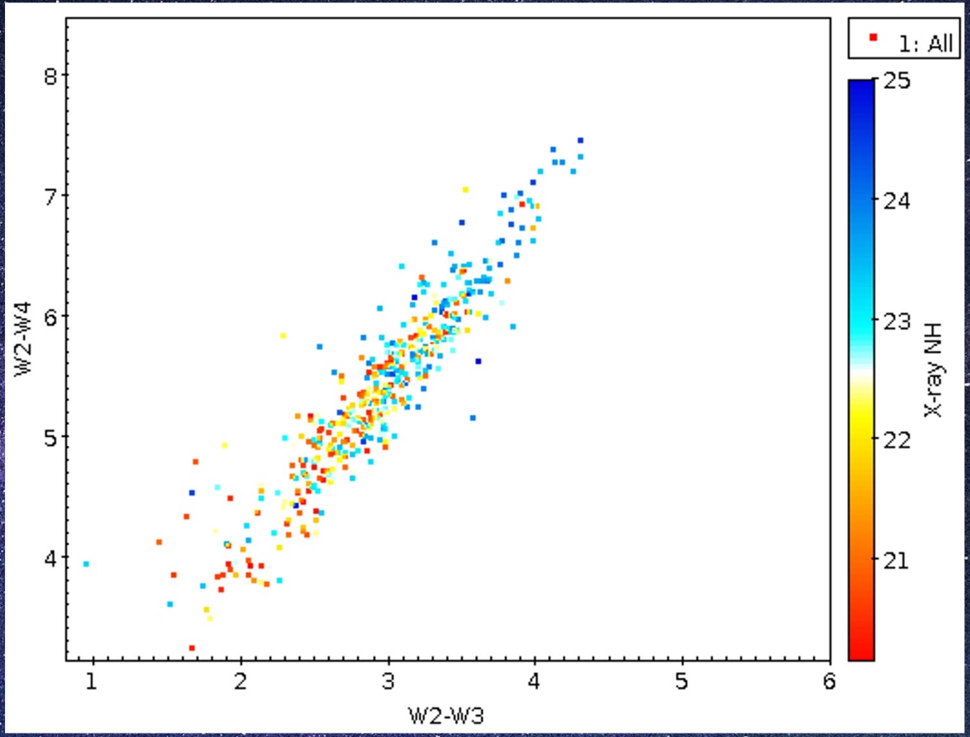
- Six Colors:
- W1-W2
- W1-W3
- W1-W4
- W2-W3
- W2-W4
- W3-W4

W1 = 3.4 μm

W2 = 4.6 μm

W3 = 12 μm

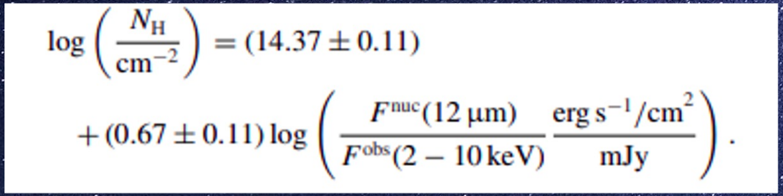
W4 = 22 μm



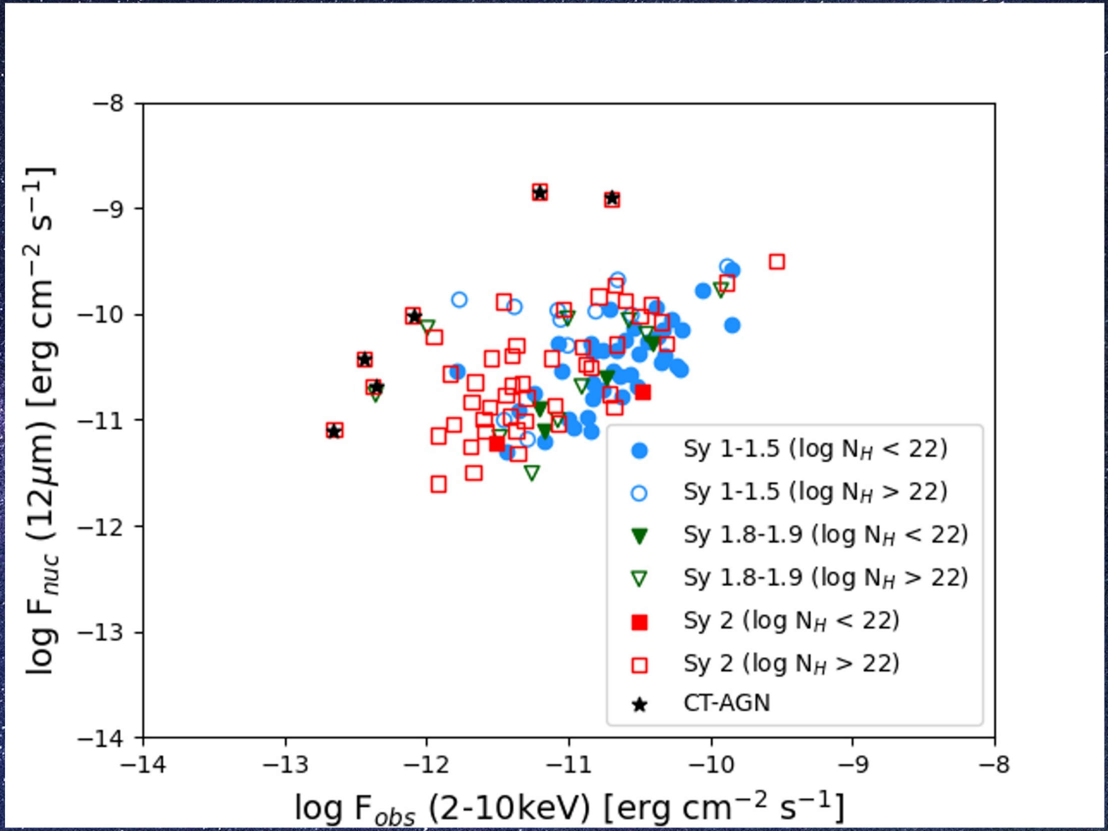
W2-W4 VS W2-W3

MIR- X-ray Relation

- There exists a well documented trend between the MIR and soft X-ray flux. The more obscured the source is, the farther off the line it will fall.



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



Asmus et al. 2015

Hardness Ratios

Soft X-rays are prone to absorption. Therefore, we use two ratios from the 2SXPS catalog (*Swift*-XRT) covering three X-ray bands:

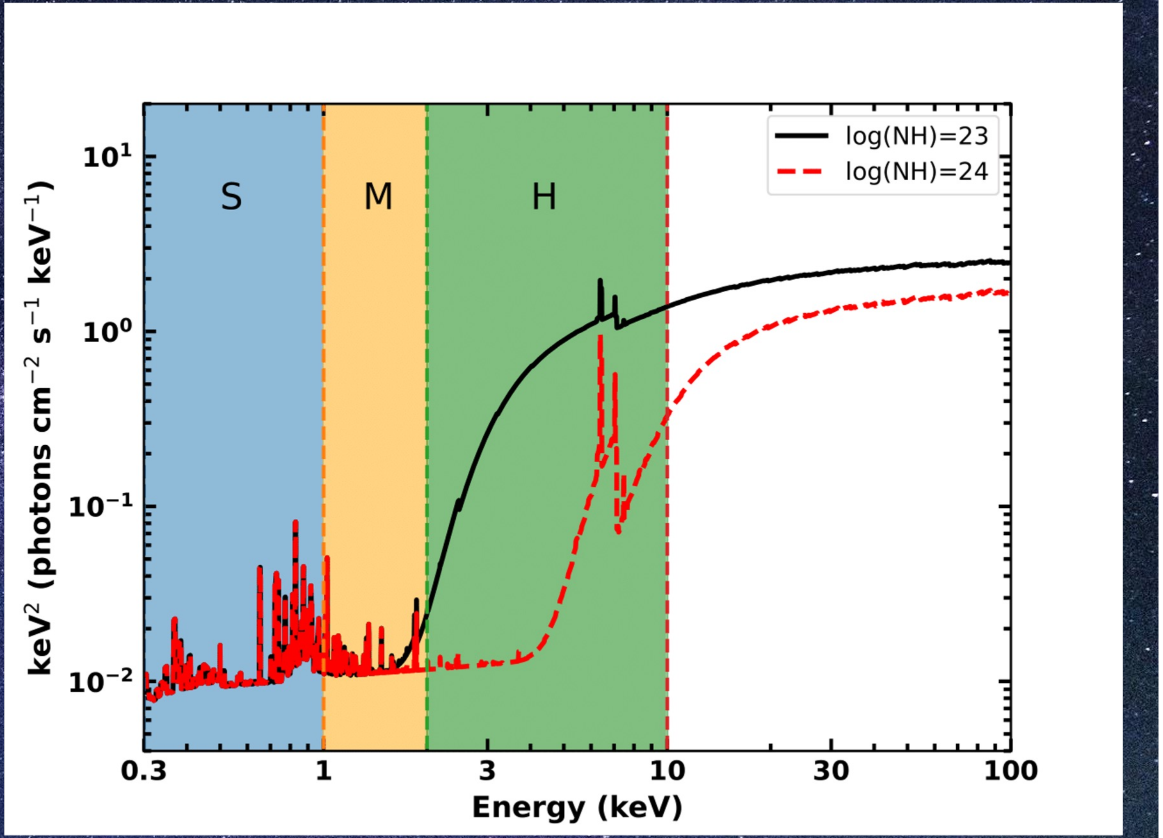
(M-S) / (M+S)

(H-M) / (H+M)

S = 0.3 - 1 keV

M = 1 - 2 keV

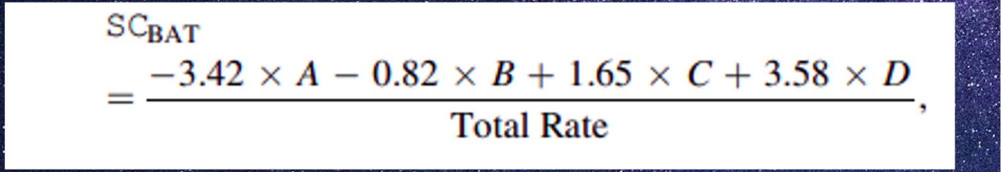
H = 2 - 10 keV



Simulations by X. Zhao

Swift-BAT Spectral Curvature

- BAT observes the hard X-ray sky from 14-150 keV.
- While less affected than soft X-rays, hard X-rays (> 10 keV) display increased curvature with NH.

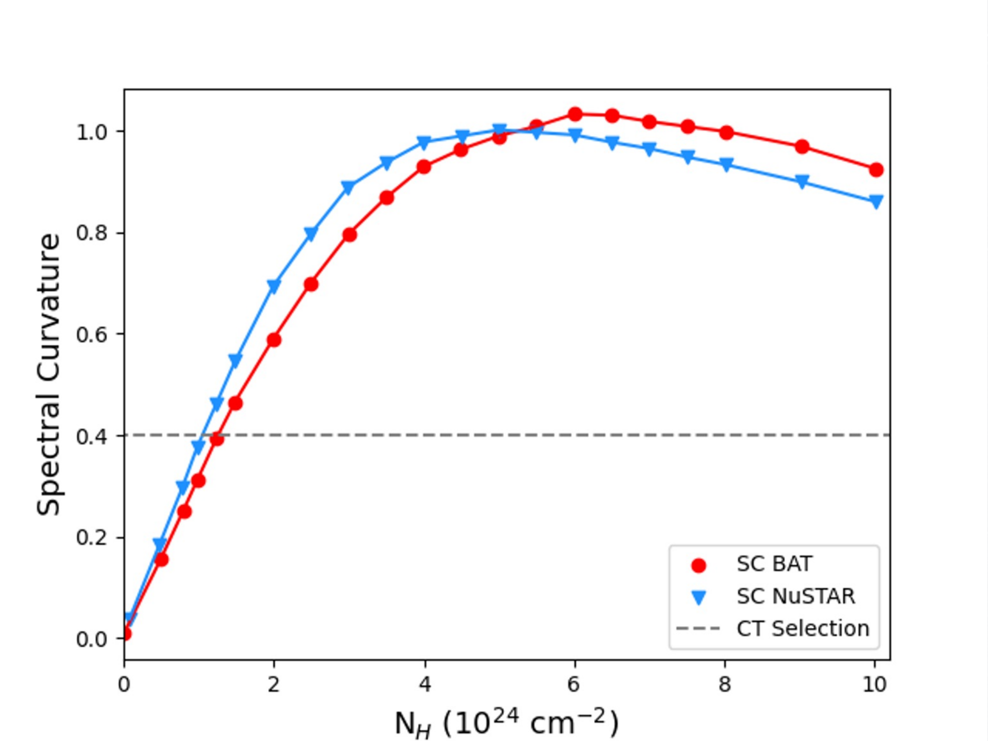


$$\mathrm{SC_{BAT}} = \frac{-3.42 \times A - 0.82 \times B + 1.65 \times C + 3.58 \times D}{\text{Total Rate}},$$

A = 14 - 20 keV C = 24 - 35
keV

B = 20 - 24 keV D = 35 - 50
keV

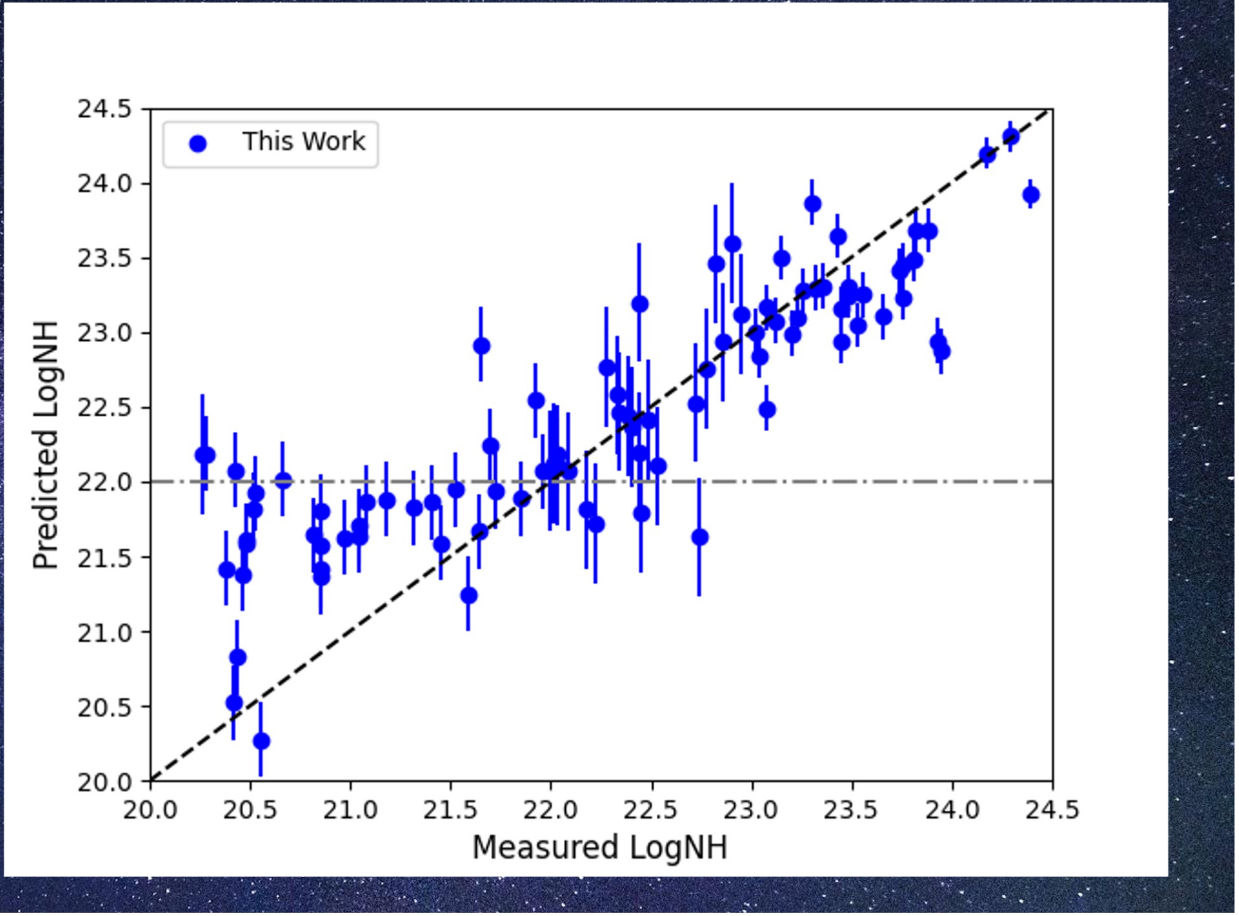
Total Rate: 14 - 50 keV



Koss et al. 2016

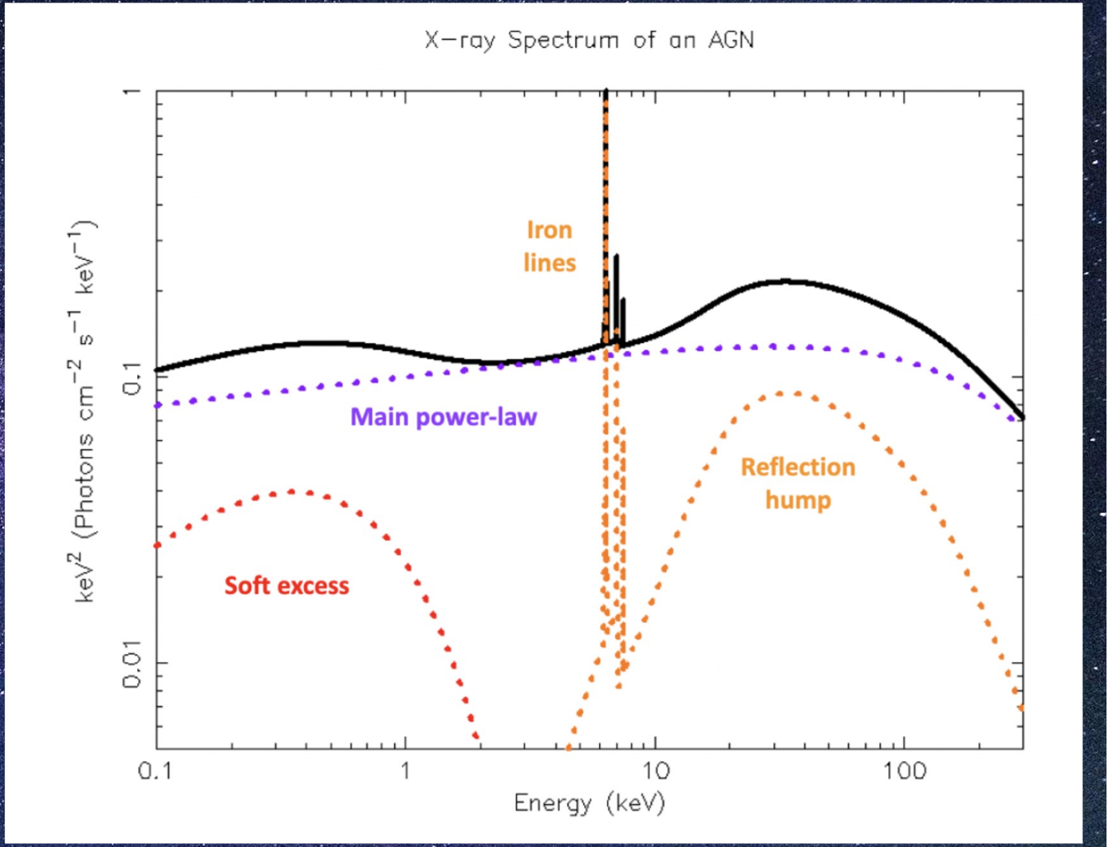
Results

- ML NH predictions vs X-ray measured NH values.
- Spearman correlation coefficient: 0.86



Components of the AGN X-ray Spectra

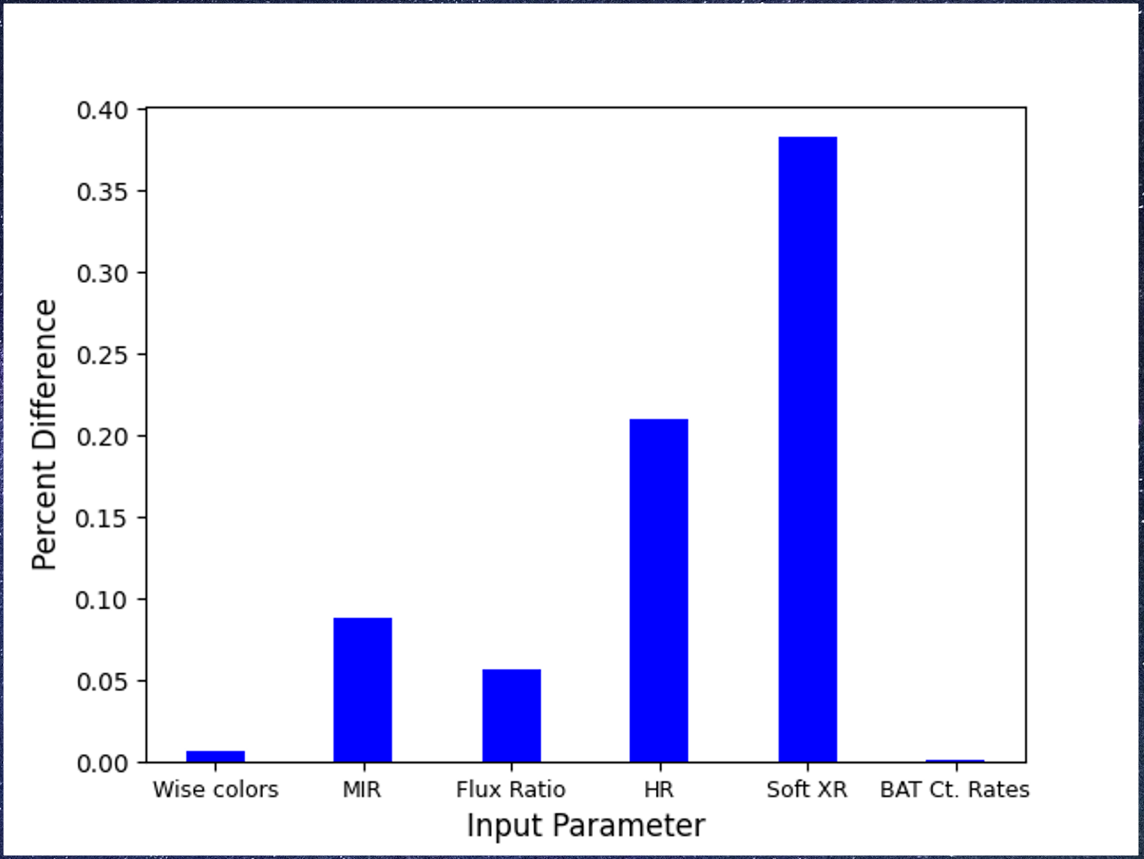
- Line-of-sight power law: Intrinsic AGN emission from the corona with a high-energy cutoff at energies between 300-500 keV.
- Reflection hump: Reprocessed intrinsic emission reflected off the torus.
- Fluorescent lines: Caused by the reflection, the predominant line being the Fe Kα at 6.4 keV.



Credit: N. Torres-Albà

Feature Importance

- Parameters using soft X-rays were the most valuable in training our algorithm.
- "MIR" = 6 WISE colors + MIR - X-ray flux ratio
- "Soft XR" = 2 HRs + MIR - X-ray flux ratio



Training the Algorithm

- Started with 1390 AGN from BAT 150 Month catalog.
- 568 had reliable NH measurements.
- 451 had XRT and WISE data.

Total Sources Used	Training Set	Testing Set
451	360 (80%)	91 (20%)

Spectral Fitting

- 361 in our sample are from the BAT 70-month catalog (Ricci et al. 2017), which provides NH values based on spectral analysis of soft X-ray (ASCA, Chandra, Suzaku, Swift-XRT, and XMM-Newton) and BAT spectra.
- Of the remaining 90, 18 had XMM, 24 had Chandra, and 48 had XRT.
- Depending on the level of obscuration, they were modeled with:
  - An absorbed power law
  - Absorbed power law + Gaussian line + scattered emission
  - Physically motivated models like MYTorus or Borus

NH Classifications

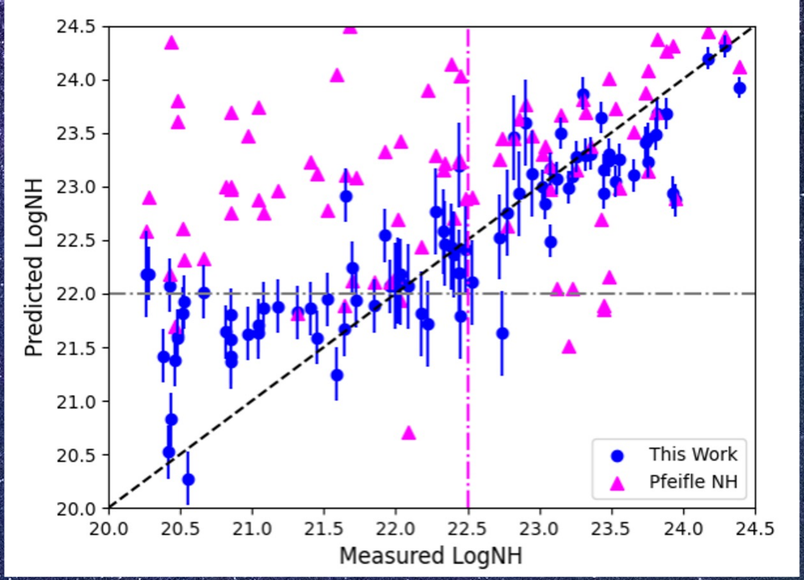
Classification	Real Number	My Work	Asmus15	Pfeifle22
CT	3	2	2	3
C-thin	28	22	25	13
Obscured	24	16	8	7
Unobscured	36	28	3	4
Total:	91	68 (75%)	38 (42%)	27 (30%)

CT: Log(NH) > 24; C-thin: 23 < Log(NH) < 24;
Obscured: 22 < Log(NH) < 23; Unobscured: Log(NH) < 22

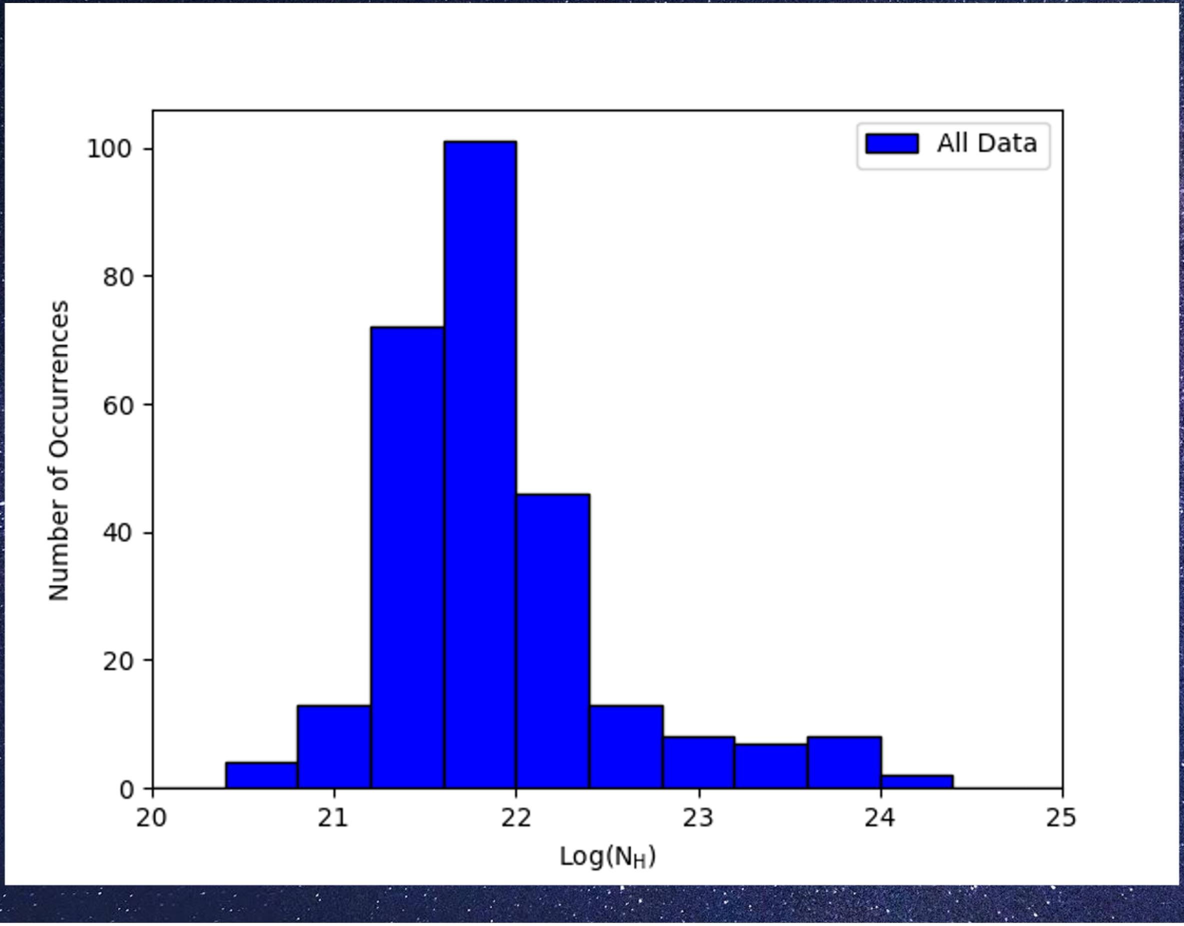
Obscured vs Unobscured

Classification	Real Number	My Work	Asmus15	Pfeifle22
Unobscured	36	28	3	4
% Correct:		77%	8%	11%

- Our algorithm is superior at identifying unobscured AGN.

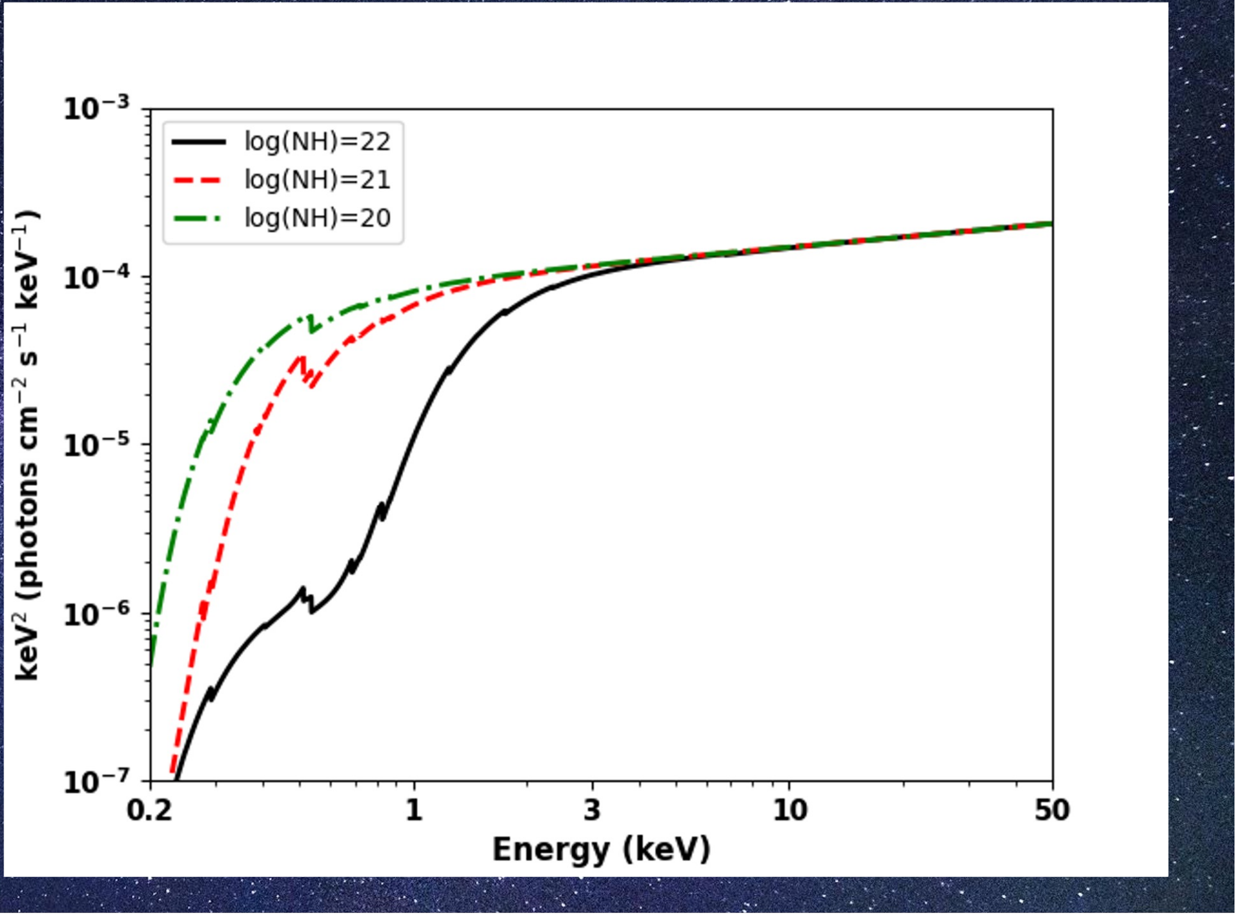


Results for BAT 150 Sources



Results

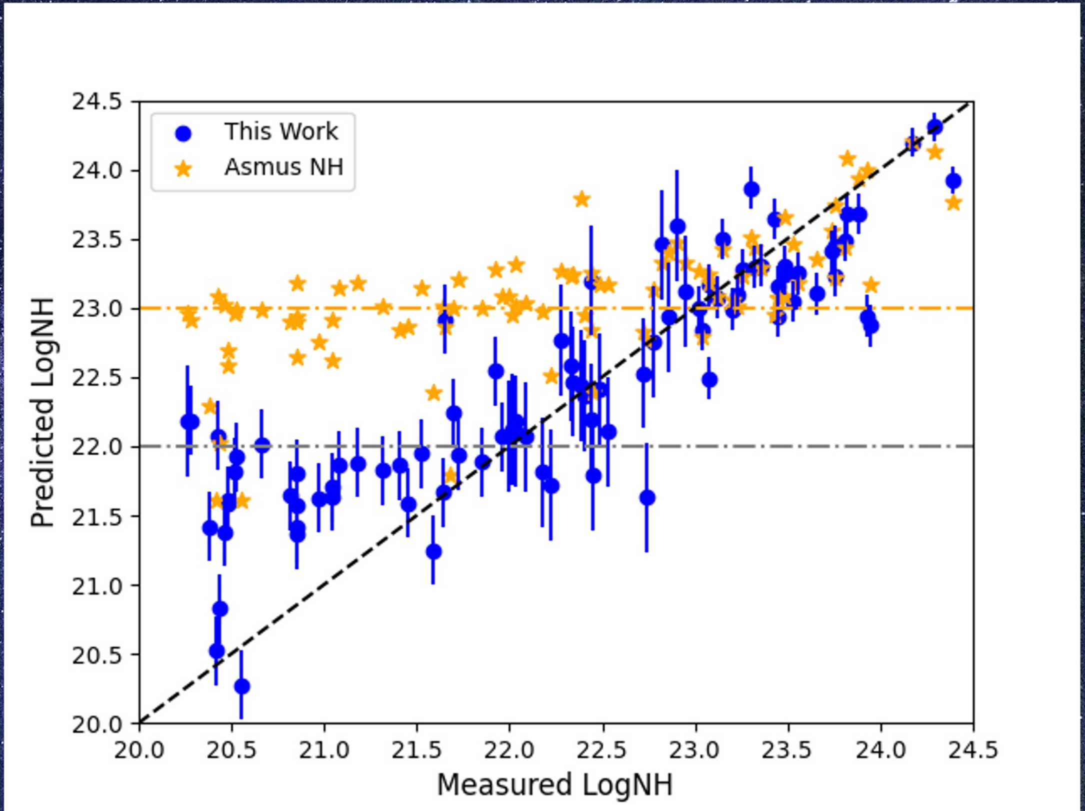
- ML NH predictions vs X-ray measured NH values.
- Spearman correlation coefficient: 0.86



Comparison with Previous Methods

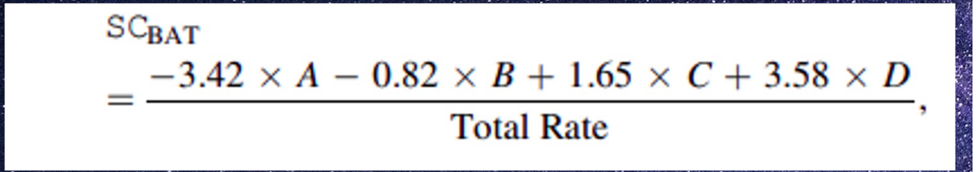
- Asmus et al. 2015
- Spearman: 0.65

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

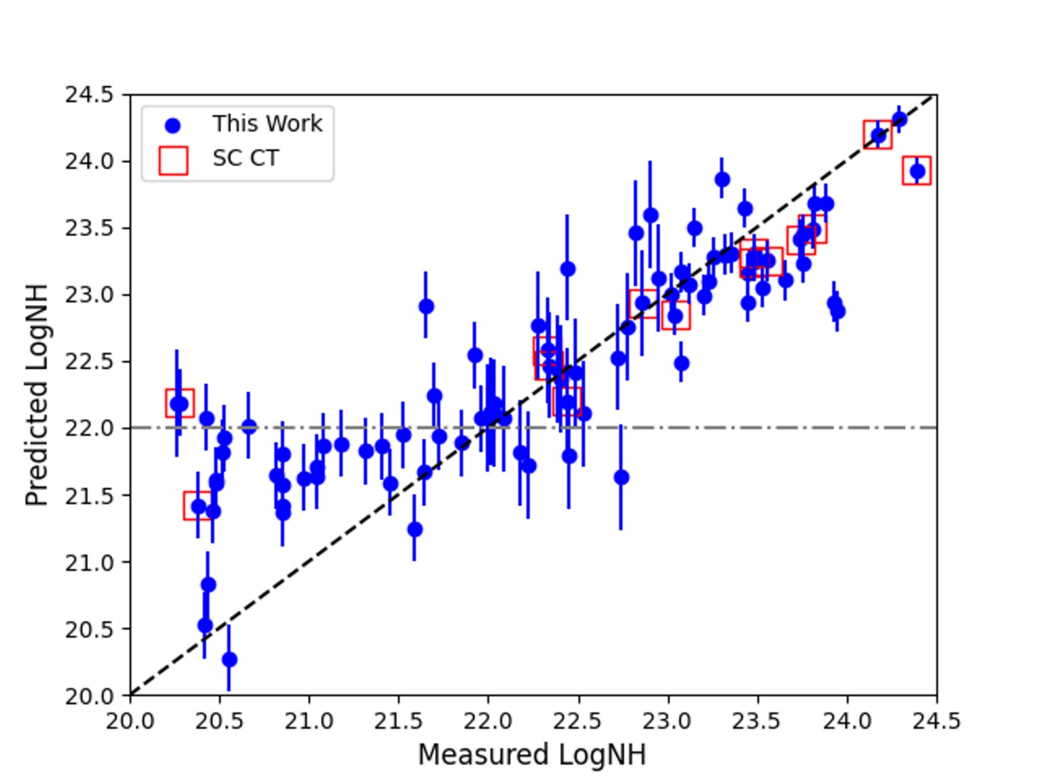


Comparison with previous Methods

- 14 sources predicted to be CT according to Koss et al. 2016.
- Only 2 are.

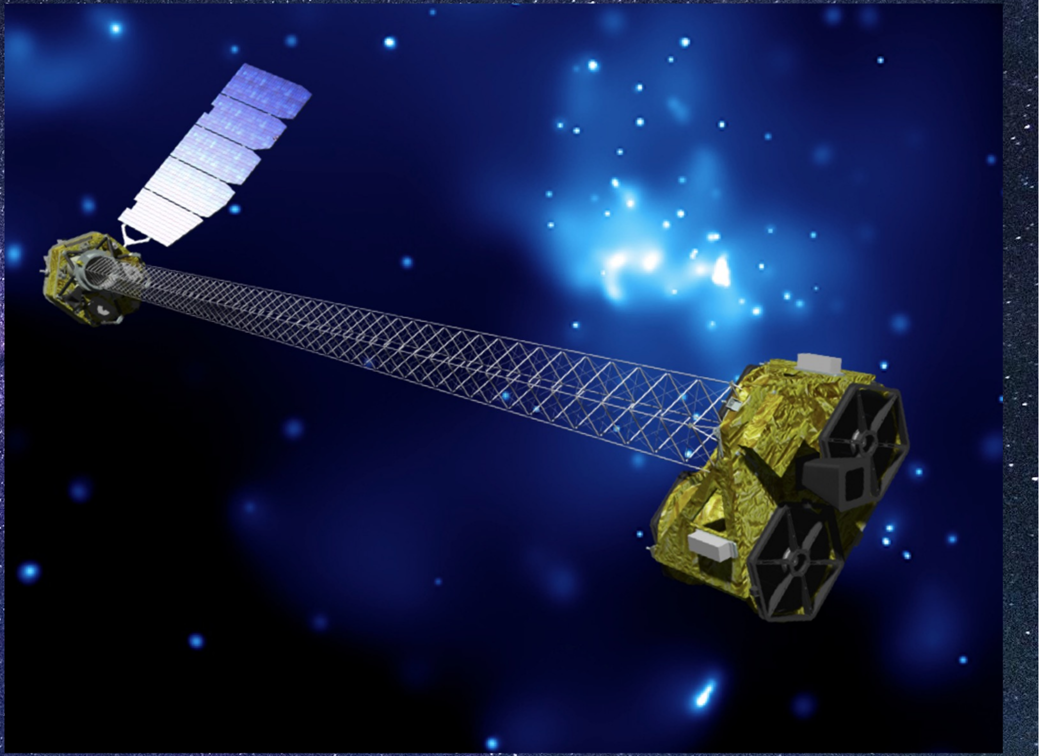


$$\mathrm{SC_{BAT}} = \frac{-3.42 \times A - 0.82 \times B + 1.65 \times C + 3.58 \times D}{\text{Total Rate}},$$

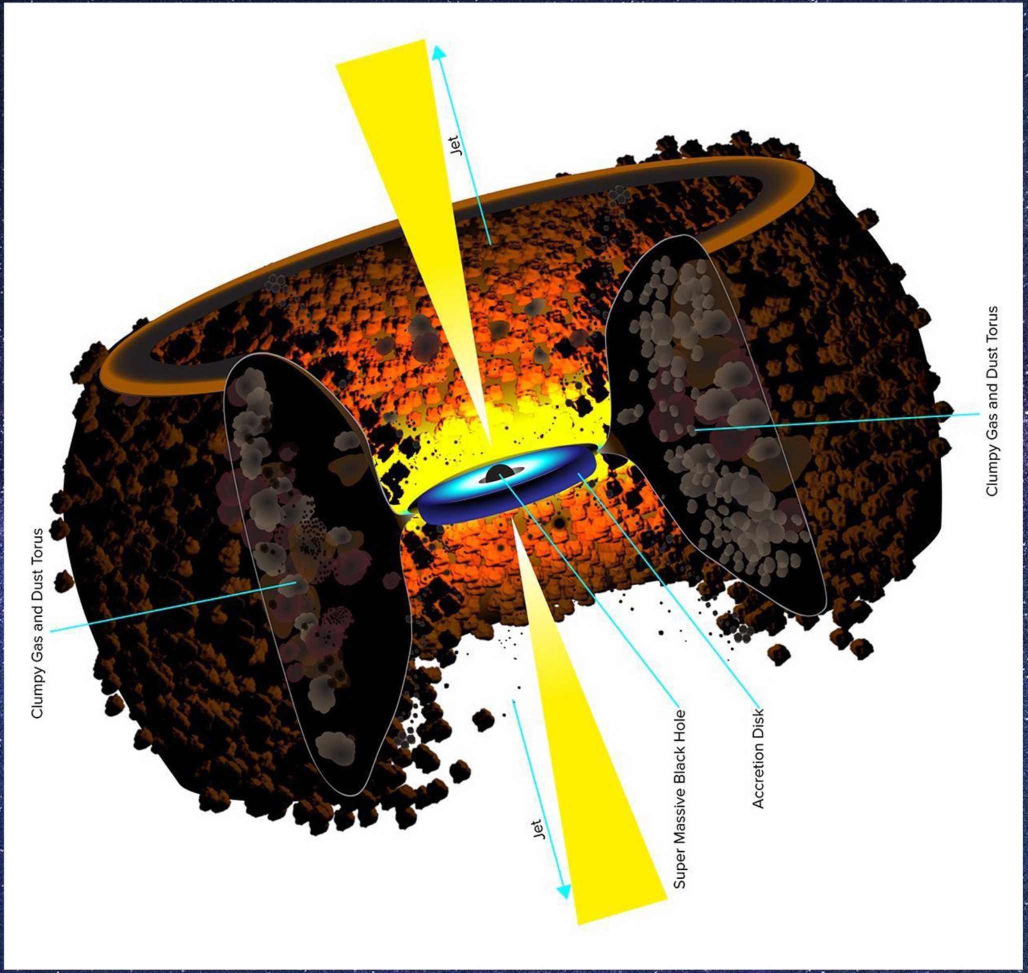


Future Work

- We applied for simultaneous *NuSTAR* - XMM-*Newton* observations of 6 AGN with predicted Log(NH) > 23.80.
- Use it on large source catalogs (XMM & *Chandra*) to get an NH estimate for many AGN at once.



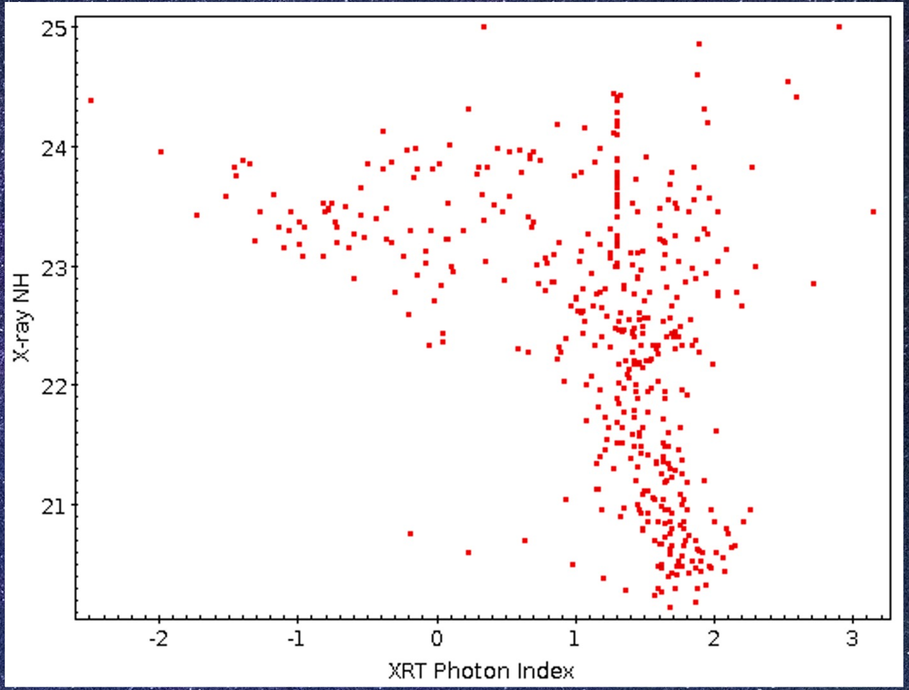
AGN Structure



Credit: B. Saxton NRAO/AUI/NSF

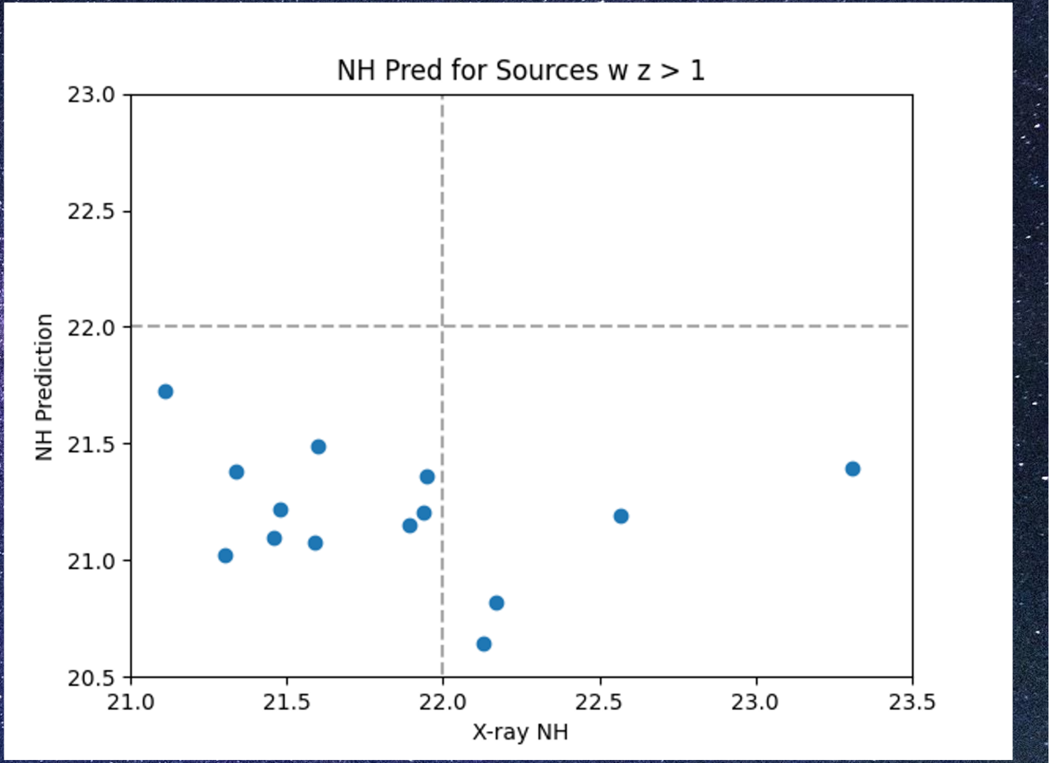
Ways to Improve the Algorithm

- NuLANDS sample has ~100 AGN with *NuSTAR* confirmed NH values. Add these sources to my training sample.
- Look into adding other parameters, such as XRT Photon Index and redder MIR (IRAS) data.
- Find high-z heavily obscured sources.

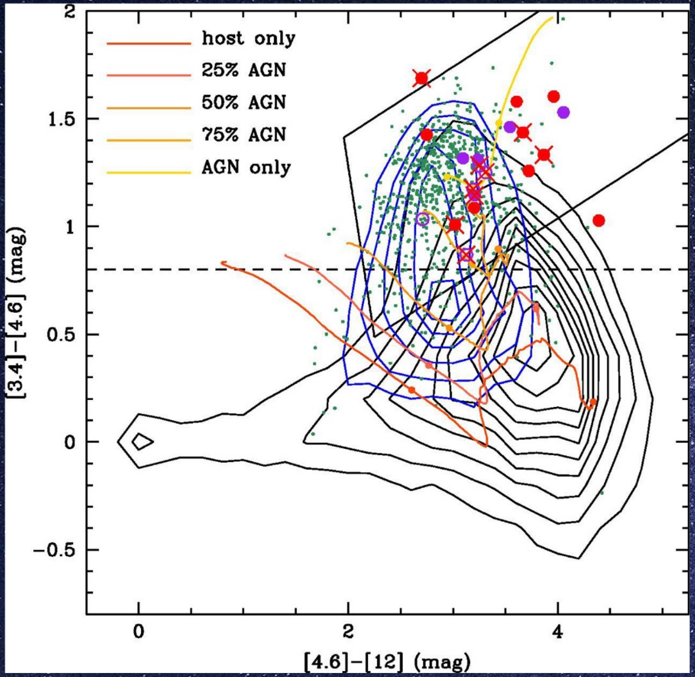


Testing Against High-z Sources

- There are 14 sources in my training sample with z > 1, and 4 with z > 3.
- Our algorithm correctly classifies 10/14 (71%) as unobscured.

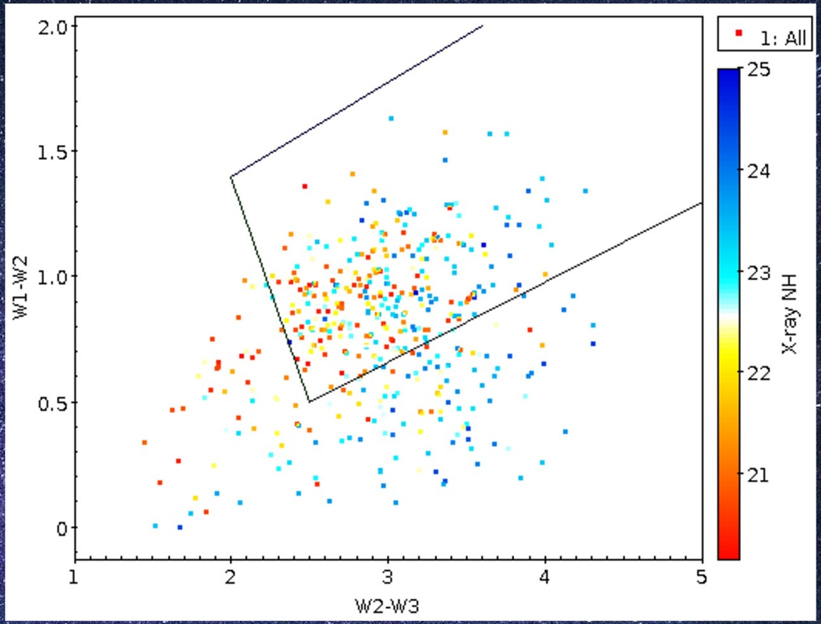


WISE Colors



Rovilos et al. 2014

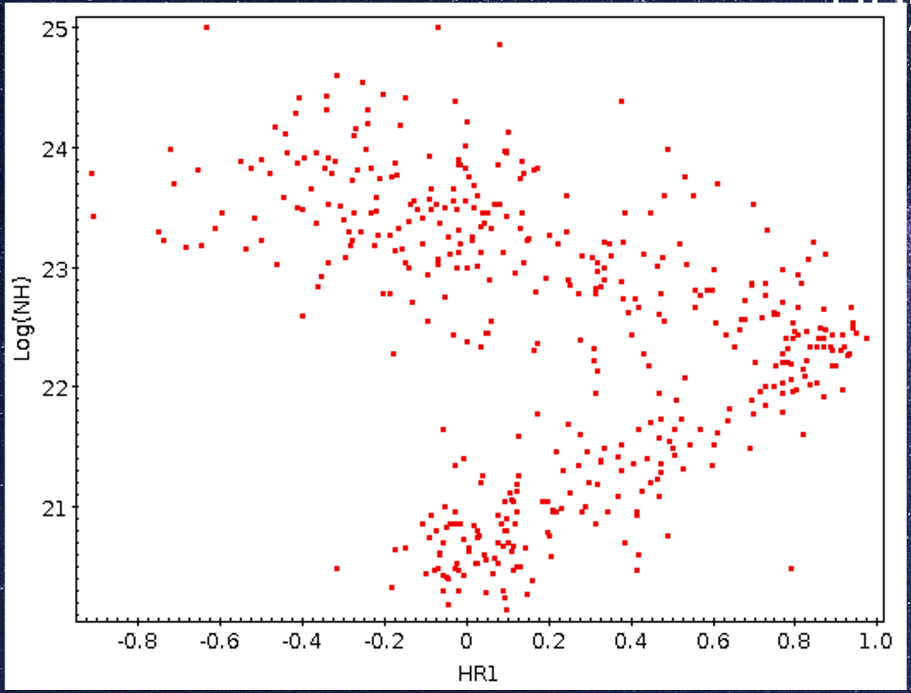
W1-W2 VS W2-W3

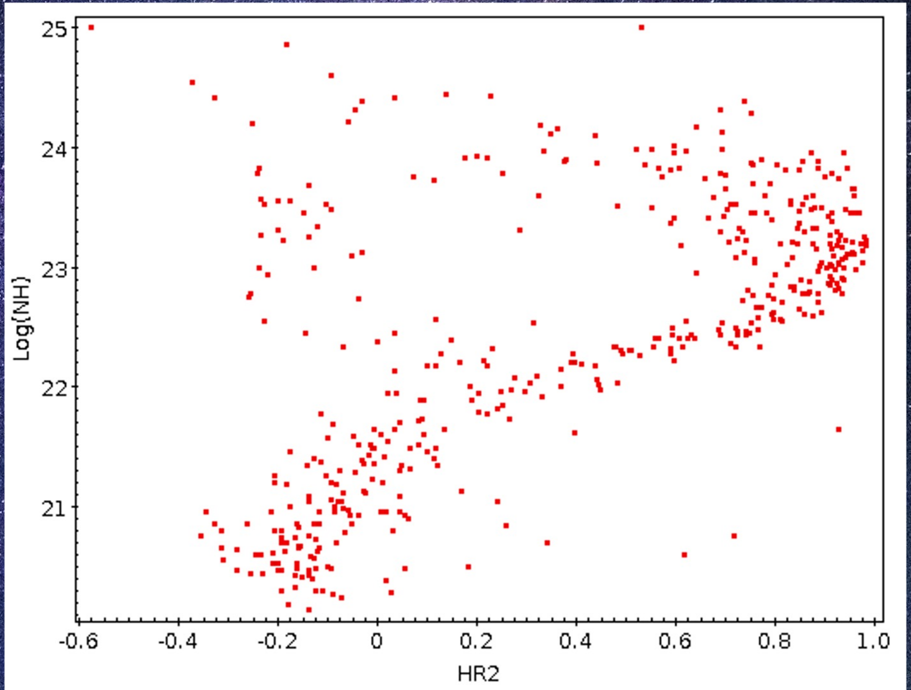


NH vs HRs

HR1

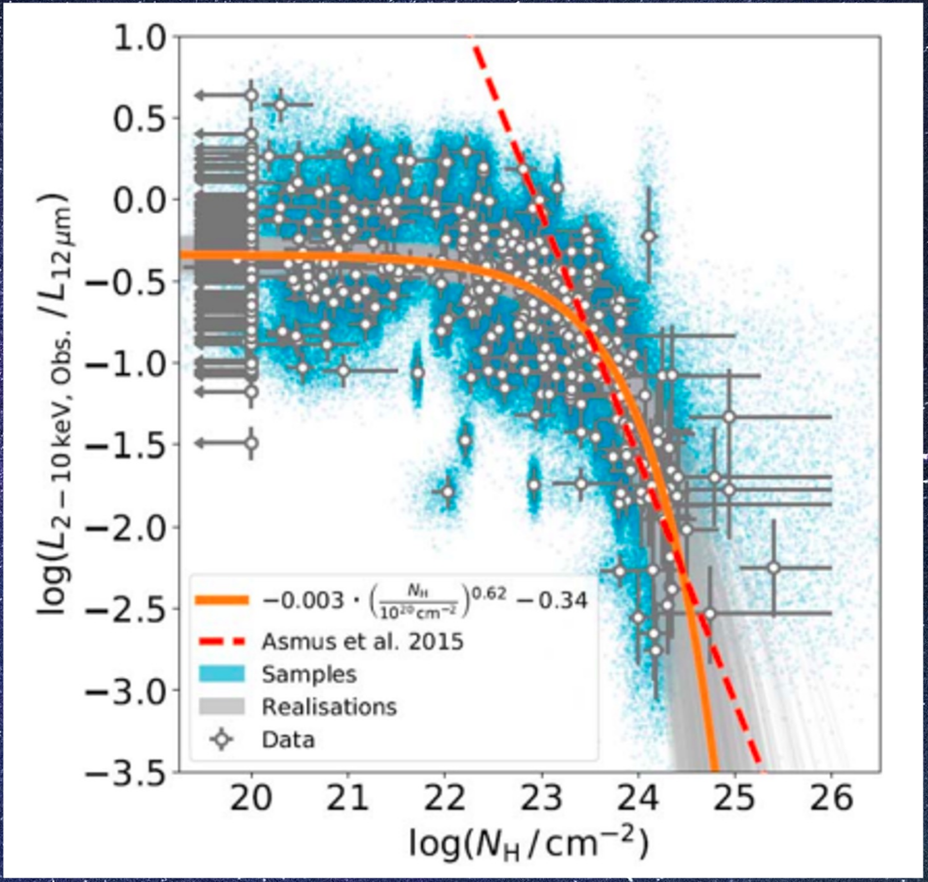
HR2





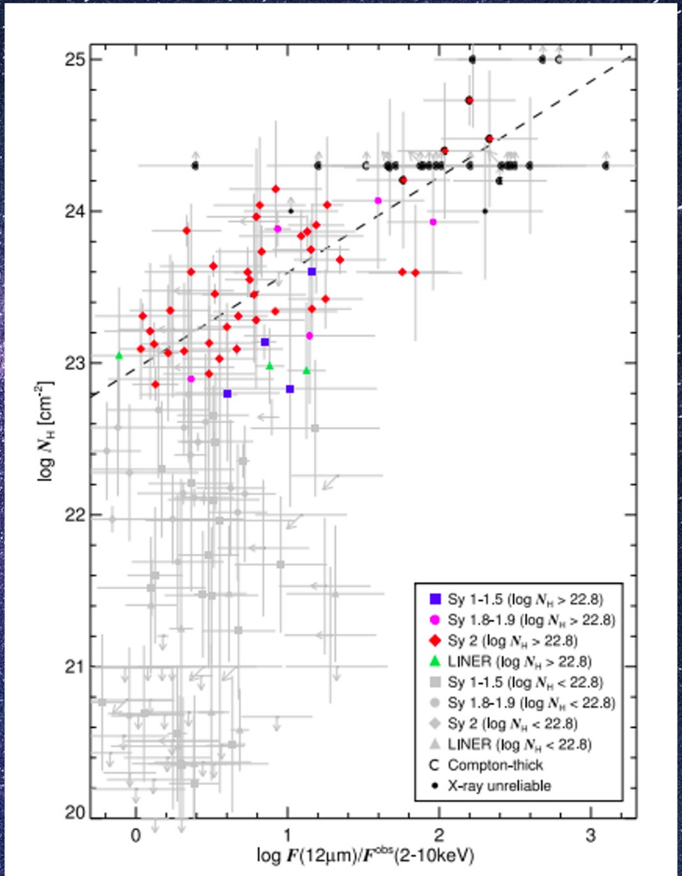
Pfeifle et al. 2022

- Very large scatter in the data
- Final sample used 456 AGN, compared to 152 from Asmus et al. 2015



Asmus et al. 2015

- Smaller data sample than Pfeifle.
- Only used 152 AGN, all with Log(NH) > 22.8 (colored points in the plot)



Standard Error for Correlation Coefficient

- r = 0.86
- n = 91 (number of sources in test sample)

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- **SE = 0.05**