Draft version June 12, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

AGNBoost: A Machine Learning Approach to AGN Identification with JWST/NIRCam+MIRI Colors and Photometry

Kurt Hamblin [1], Allison Kirkpatrick [1], Bren E. Backhaus [1], Gregory Troiani [1], Fabio Pacucci [2,3], Jonathan R. Trump [4], Alexander de la Vega [5], L. Y. Aaron Yung [6], Jeyhan S. Kartaltepe [7], Dale D. Kocevski [8], Anton M. Koekemoer [6], Erini Lambrides [9,*], Casey Papovich [10,11], Kaila Ronayne [10,11], Guang Yang [12,13], Pablo Arrabal Haro [14,†], Micaela B. Bagley [14,15], Mark Dickinson [16], Steven L. Finkelstein [15,17] and Nor Pirzkal [18]

[1] *Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA*
[2] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden St, Cambridge, MA 02138, USA*
[3] *Black Hole Initiative, Harvard University, 20 Garden St, Cambridge, MA 02138, USA*
[4] *Department of Physics, 196 Auditorium Road, Unit 3046, University of Connecticut, Storrs, CT 06269, USA*
[5] *Department of Physics and Astronomy, University of California, 900 University Ave, Riverside, CA 92521, USA*
[6] *Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*
[7] *Laboratory for Multiwavelength Astrophysics, School of Physics and Astronomy, Rochester Institute of Technology, 84 Lomb Memorial Drive, Rochester, NY 14623, USA*
[8] *Department of Physics and Astronomy, Colby College, Waterville, ME 04901, USA*
[9] *NASA-Goddard Space Flight Center, Code 662, Greenbelt, MD, 20771, USA*
[10] *Department of Physics and Astronomy, Texas A&M University, College Station, TX, 77843-4242 USA*
[11] *George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX, 77843-4242 USA*
[12] *Nanjing Institute of Astronomical Optics & Technology, Chinese Academy of Sciences, Nanjing 210042, China*
[13] *CAS Key Laboratory of Astronomical Optics & Technology, Nanjing Institute of Astronomical Optics & Technology, Nanjing 210042, China*
[14] *Astrophysics Science Division, NASA Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, MD 20771, USA*
[15] *Department of Astronomy, The University of Texas at Austin, Austin, TX, USA*
[16] *NSF's National Optical-Infrared Astronomy Research Laboratory, 950 N. Cherry Ave., Tucson, AZ 85719, USA*
[17] *Cosmic Frontier Center, The University of Texas at Austin, Austin, TX, USA*
[18] *ESA/AURA Space Telescope Science Institute*

ABSTRACT

We present `AGNBoost`, a machine learning framework utilizing `XGBoostLSS` to identify AGN and estimate redshifts from JWST NIRCam and MIRI photometry. `AGNBoost` constructs 121 input features from 7 NIRCam and 4 MIRI bands—including magnitudes, colors, and squared color terms—to simultaneously predict the fraction of mid-IR $3-30\,\mu$m emission attributable to an AGN power law ($\mathrm{frac_{AGN}}$) and photometric redshift. Each model is trained on a sample of 10^6 simulated galaxies from `CIGALE` providing ground truth values of both $\mathrm{frac_{AGN}}$ and redshift. Models are tested against both mock `CIGALE` galaxies set aside for testing and 698 observations from the JWST MIRI EGS Galaxy and AGN (MEGA) survey. On mock galaxies, `AGNBoost` achieves 15% outlier fractions of 0.19% ($\mathrm{frac_{AGN}}$) and 0.63% (redshift), with a root mean square error (σ_{RMSE}) of 0.027 for $\mathrm{frac_{AGN}}$ and a normalized mean absolute deviation (σ_{NMAD}) of 0.011 for redshift. On MEGA galaxies with spectroscopic redshifts, `AGNBoost` achieves $\sigma_{\mathrm{NMAD}} = 0.074$ and 17.05% outliers, with most outliers at $z_{\mathrm{spec}} > 2$. `AGNBoost` $\mathrm{frac_{AGN}}$ estimates broadly agree with `CIGALE` fitting ($\sigma_{\mathrm{RMSE}} = 0.183$, 20.41% outliers), and `AGNBoost` finds a similar number of AGNs as `CIGALE` SED fitting. The flexible framework of `AGNBoost` allows straightforward incorporation of additional photometric bands and derived qualities, and simple re-training for other variables of interest. `AGNBoost`'s computational efficiency makes it well-suited for wide-sky surveys requiring rapid AGN identification and redshift estimation.

* NPP Fellow

† NASA Postdoctoral Fellow

1. INTRODUCTION

The James Webb Space Telescope (JWST) has revolutionized our ability to study galaxies in the mid-infrared, achieving sensitivity levels an order of magnitude deeper than Spitzer/MIPS in significantly shorter integration times (Rigby et al. 2023; Gardner et al. 2006, 2023). The Mid-Infrared Instrument (MIRI; Ni et al. 2021) provides unprecedented wavelength coverage from $5 - 28.5\,\mu$m across nine photometric filters, enabling detailed characterization of dusty galaxies out to the so-called "cosmic noon" epoch ($z \sim 1{-}3$), during which the bulk of cosmic star formation and black hole growth occurred (Heckman et al. 2004; Hopkins 2004; Fontana et al. 2006; Pérez-González et al. 2008; Shankar et al. 2009; Aird et al. 2010; Kormendy & Ho 2013; Madau & Dickinson 2014; Förster Schreiber & Wuyts 2020). This enhanced sensitivity and spectral coverage is particularly essential for studying the growth of galaxies and active galactic nuclei (AGN) over cosmic time and, notably, at cosmic noon. JWST/MIRI is able to identify heavily obscured AGNs at cosmic noon that are entirely missed in traditional optical surveys, allowing for the study of supermassive black hole-host coevolution (e.g. Yang et al. 2023; Kirkpatrick et al. 2023). This requires disentangling the emission from AGN and highly star-forming galaxies (SFGs), a process that requires accurate models and often significant computational resources, especially in the case of large surveys.

The mid-IR emission from AGNs and SFGs originates from fundamentally different physical processes. In AGN, the central supermassive black hole is surrounded by a dusty torus that absorbs and reprocesses radiation from the accretion disk. This dust, heated to temperatures over 1000 K, produces a characteristic power-law continuum ($f_\nu \propto \nu^{-\alpha}$) in the mid-IR (Stern et al. 2005). The slope and intensity of this continuum reflect both the intrinsic AGN luminosity and the geometry of the obscuring torus (Toba et al. 2014; Laurent et al. 2000; Stalevski et al. 2016, e.g.). In contrast, SFGs exhibit strong emission features from polycyclic aromatic hydrocarbons (PAHs), with the most prominent features at 6.2, 7.7, 11.2, and 12.7 μm, excited by UV photons from young stellar populations (Draine & Li 2001; Peeters et al. 2004; Magdis et al. 2012). The relative strengths of these PAH features to the underlying continuum provide key diagnostics to distinguish active from non-active galaxies.

With the improved mid-IR coverage and sensitivity of JWST, MIRI surveys are also finding populations of mid-IR weak galaxies at cosmic noon. These populations are most likely mid-IR weak due to intrinsically low luminosities, and these sources dominate the MIRI population at $L_{IR} < 10^{10}\,L_\odot$. The mid-IR emission of these mid-IR weak galaxies can look very similar to the mid-IR power law of an AGN torus, making it difficult to distinguish between the two. AGN and mid-IR weak galaxies can only be distinguished from comparison of mid-IR emission and rest-frame near-IR emission. In particular, the ability to identify the stellar bump at $\lambda \sim 1.6\,\mu$m and the stellar emission minimum at $\lambda \sim 5\,\mu$m has proven crucial to discriminate between AGN and mid-IR weak galaxies (Kirkpatrick et al. 2023).

The interplay between spectral features and redshift makes the identification of AGN with mid-IR photometry difficult unless the redshift is known a priori. Observing the hot dust grains of the AGNs torus at rest-frame wavelengths of $\lambda \approx 3 - 10\,\mu$m is crucial for identifying AGN and constraining the IR AGN luminosity (Durodola et al. 2024), since the resulting power-law emission is distinct from mid-IR SFG emission (Padovani et al. 2017; Franceschini et al. 1991). Importantly, the optical depth is low at mid-IR wavelengths so the AGN emission is not strongly diminished by obscuring dust (Hickox & Alexander 2018). AGN identification cannot reliably be done in the near-IR due to the higher optical depth caused by obscuring dust and contamination from the peak of galaxy stellar emission at near-IR wavelengths. Additionally, systematic redshifting of PAH features in and out of JWST/MIRI bands can cause SFGs to mimic the rising power-law signature characteristic of AGN (Kirkpatrick et al. 2023). Figure 1, which is derived from a viewable animation, illustrates this degeneracy in the mid-IR emission of SFGs, AGNs, and mid-IR weak galaxies.

Historically, AGNs have been photometrically identified in the mid-IR with various color selections (Kirkpatrick et al. 2013; Assef et al. 2018; Lacy et al. 2004, 2007; Sajina et al. 2005; Stern et al. 2005; Donley et al. 2012; Alonso-Herrero et al. 2016; Messias et al. 2012; Assef et al. 2013; Wu et al. 2012). These color selections make use of the intrinsic spectral differences between AGNs and other sources to divide them into distinct regions of color-space. While mid-IR color selections are capable, their effectiveness is notably limited to galaxies with AGNs that dominate the SED, where $\gtrsim 50\%$ of the mid-IR emission is attributable to an AGN (Assef et al. 2013). AGNs that do not dominate the SED can mix with prominent PAH features to produce a similar spectral shape as a SFG with intrinsically weak PAH features (Pope et al. 2008; Sajina et al. 2012; Kirkpatrick et al. 2012). This degeneracy in spectral shape means that weak PAH features alone are not a robust indicator of the presence of an AGN.

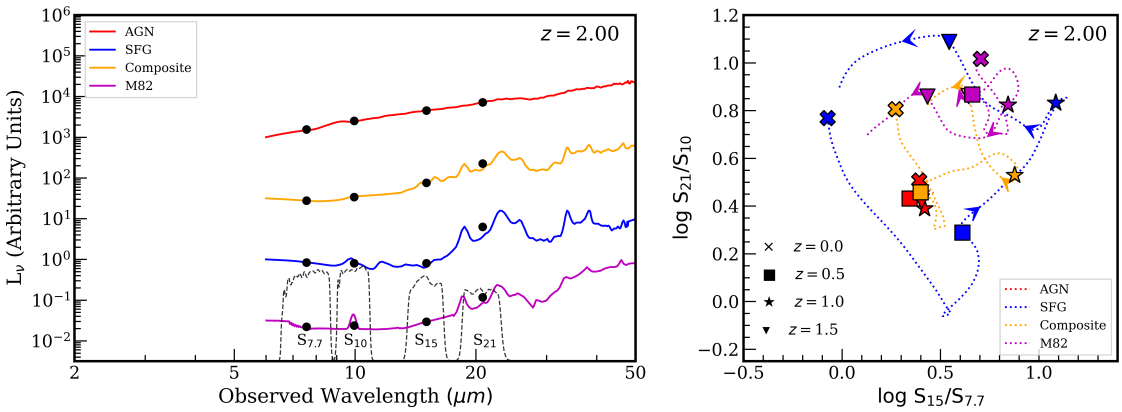

Figure 1. A single frame of an animation illustrating the mid-IR color properties of SF+AGN composites, SFGs, and mid-IR weak galaxies. The AGN, Composite, and SFG templates are from Kirkpatrick et al. (2015). Here, M82 (Förster Schreiber et al. 2003) serves as a proxy for a SFG with weak PAH features. Without prior redshift information, it is not possible to draw boundaries in this color space that robustly separate these classes of sources. The full animation can be viewed here: https://hamblin-ku.github.io/ColorAnimation/sed_color_animation.mp4.

In recent years, spectral energy distribution (SED) fitting methods (see Iyer et al. 2025; Conroy 2013 for a review) have emerged as an alternative to color selection for AGN identification. These methods simultaneously model multiple physical components from the UV to the far-IR by generating libraries of theoretical templates and statistically comparing them to observed photometry. While more comprehensive than simple color cuts, traditional parametric SED fitting suffers from two notable limitations that impede its application to large galaxy surveys. First, the computational demands can be prohibitive, with runtimes ranging from days to weeks for catalogs of a few thousand sources. Second, the results are highly sensitive to the chosen parameter grid or parameter priors, meaning that the SED fitting process is rarely a one-time activity and can require multiple iterations, further increasing the needed computation time (Pacifici et al. 2023). Non-parametric approaches, such as the dense basis method of Iyer & Gawiser (2017), have emerged to address these limitations, offering substantial improvements in computational efficiency ($\sim$ 30 minutes for 1000 galaxies) and reduced sensitivity to parameter priors. While these advances represent significant progress, there remains value in developing alternative efficient methods, namely for AGN-related tasks, that can provide immediate feedback to identify small sub-populations of interest in large catalogs for follow-up study.

Machine learning based techniques are well-suited for such tasks when speed is a priority. Unlike traditional color-selection methods, machine learning algorithms can simultaneously analyze multiple colors across all available bands, enabling the identification of subtle spectral features that distinguish AGNs from SFGs. This multi-dimensional approach is potentially powerful for breaking degeneracies that arise when PAH features are redshifted between bands, as the models can learn to recognize patterns in how multiple colors change simultaneously. Recently, machine learning algorithms have been successfully applied to color selection tasks and redshift estimation in a variety of wide-field optical and infrared surveys (Holwerda et al. 2021; Bai et al. 2018; Krakowski et al. 2016; Morello et al. 2018; Chao et al. 2019; Mechbal et al. 2024). The abundance of accurately labeled data and spectroscopic coverage makes wide field surveys natural candidates for the application of machine learning models, but there have been few attempts to apply these models to surveys with sparse spectroscopic coverage and smaller sample sizes (N$\sim$ 100 $-$ 1000). This lack of spectroscopic coverage and overall limited amounts of data poses a challenge to the application of machine learning algorithms, which are generally prone to over-fitting and poor generalization on small datasets (Hastie et al. 2001).

In this paper, we explore the potential of machine learning models to supplant traditional color selection when applied to small datasets (N $\sim$ 100 $-$ 1000) with our machine learning model `AGNBoost`, based on the `XGBoostLSS` algorithm (März 2019). `AGNBoost` is trained on the JWST/ NIRCam+MIRI photometric observations and derived quantities of a sample of mock galaxies from `CIGALE` (Boquien et al. 2019), and

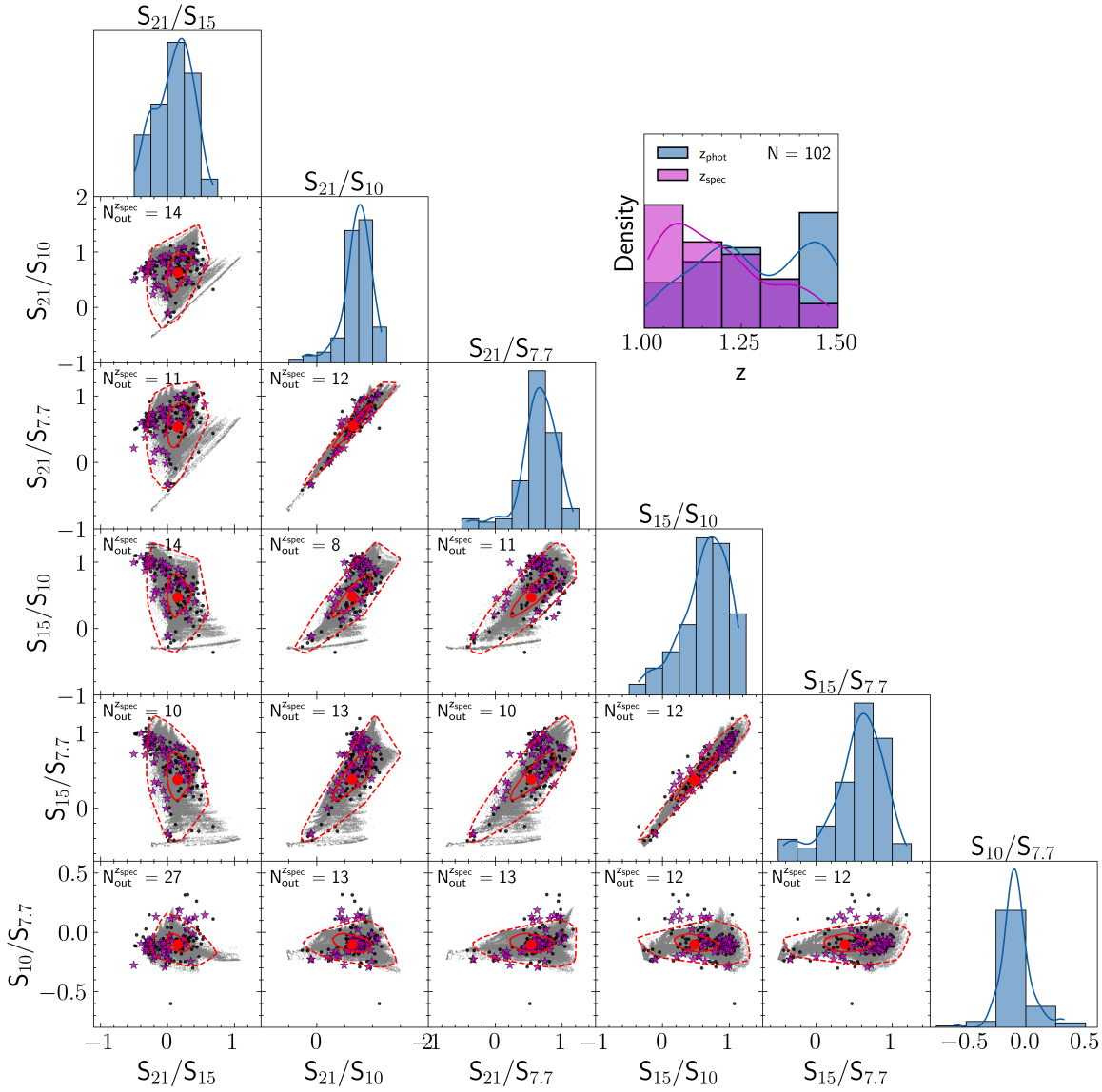

Figure 2. JWST/MIRI color comparison between mock `CIGALE` galaxies and MEGA observations in the $1.0 \leq z < 1.5$ redshift bin. MEGA sources with spectroscopic redshifts are shown as purple stars, those with photometric redshifts as black points, and individual `CIGALE` galaxies as gray points. Red regions show bagplots (the bivariate analog of boxplots; Rousseeuw et al. 1999), where the filled area contains 50% of the `CIGALE` data, the central point marks the depth median, and the dashed line indicates the outlier boundary. The MEGA galaxy colors show good agreement with the `CIGALE` mock galaxy color distribution, validating the representativeness of our training data. AGNBoost performance for MEGA galaxies lying outside the `CIGALE` color space is examined in Figure A1 in Appendix A.

we test it against both the mock galaxies and a set of real JWST/NIRCam+MIRI observations from the MIRI EGS Galaxy and AGN (MEGA) survey (Backhaus et al. 2025). In Section 2 we describe our training sample construction and the MEGA observations. Section 3 details the `AGNBoost` methodology and implementation. Section 4 presents our model performance and assessment. Section 5 summarizes our conclusions and outlines future applications.

2. DATA AND SAMPLE SELECTION

2.1. `CIGALE` *Training Sample*

The foundation of `AGNBoost`'s training relies on a comprehensive set of mock galaxies generated using the Code Investigating GALaxy Emission (`CIGALE`; Boquien et al. 2019). We use `CIGALE` v2022.1 in the 'save_fluxes' mode to generate $\approx 10^9$ unique galaxy templates using `CIGALE`'s physical models, which include both stellar and AGN emission components. The total adopted parameters are summarized in Table 1.

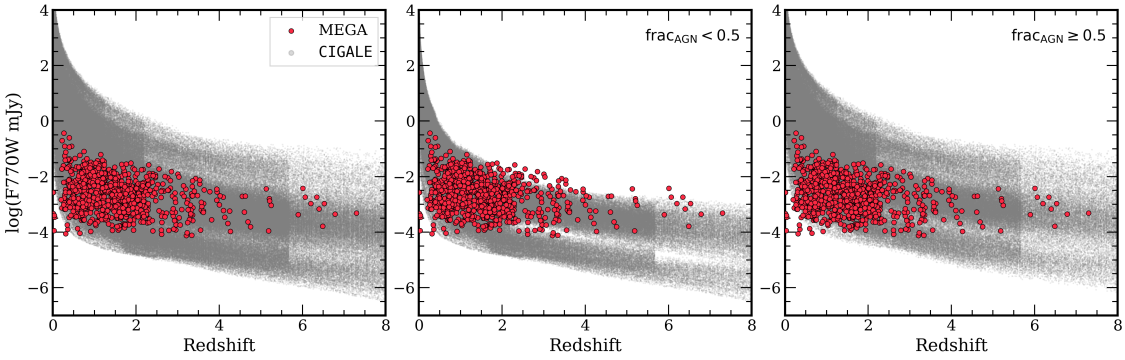

Figure 3. JWST/MIRI F770W photometry distributions comparing mock `CIGALE` galaxies (gray) and MEGA observations (red). The left panel shows the full `CIGALE` mock set, the middle panel shows `CIGALE` non-AGN ($\mathrm{frac_{AGN}} < 0.5$), and the right panel shows `CIGALE` AGN ($\mathrm{frac_{AGN}} \geq 0.5$). The same MEGA sample is plotted in all three panels for comparison. The `CIGALE` simulations provide excellent coverage of the MEGA F770W photometry, with only 6 ($\sim 0.9\%$) very low-redshift sources ($z \sim 0.01$) falling outside the simulated range.

For star formation, we adopt a standard delayed star formation history (SFH; `sfhdelayed` in `CIGALE`), with e-folding times ranging from $0.1-5$ Gyr and stellar ages ranging from $0.5-5$ Gyr. We also allow for starbursts with an age of 20 Myr and a mass fraction of $f_{\mathrm{burst}} = 0.0, 0.02, 0.1$. For the stellar component, we adopt the Bruzual & Charlot (2003) stellar population models (module `bc03` in `CIGALE`) with a Chabrier initial mass function (Chabrier 2003). Dust attenuation is modeled using the `dustatt_modified_starburst` module which is based on the attenuation curve of Calzetti et al. (2000) but extended to short wavelengths ($91.2-150$ nm) with Leitherer et al. (2002). The allowed range of color excess is $E(B-V) = 0-1$. Galactic dust emission is modeled with the `dl2014` module (Draine et al. 2014), which models dust emission with a diffused emission component and a photodissociating component associated with star formation. We use a PAH mass fraction (q_{PAH}) varying from $q_{PAH} = 1.77-5.26$, a minimum radiation parameter of diffuse dust (U_{min}) ranging from $U_{min} = 1-10$, and relative strengths of diffuse and star forming components (γ) of $\gamma = 0.01-0.2$. These parameter ranges were chosen to envelope ranges of reasonable values expected for MEGA galaxies, and the gridding was chosen to permit `CIGALE` fitting of the MEGA catalog in a reasonable timeframe (∼days on a personal computer).

AGN emission is modeled using the `SKIRTOR` clumpy torus model, based on the clumpy torus models from Stalevski et al. (2012, 2016). In the `SKIRTOR` module, the relative strength of AGN in the IR is set by the $\mathrm{frac_{AGN}}$ parameter, which corresponds to the fraction of $3-30\,\mu$m mid-IR light attributable to an AGN power law. We let $\mathrm{frac_{AGN}}$ range from $\mathrm{frac_{AGN}} = 0.0-1$ in steps of 0.1. We set the viewing angle to 70°, a typical value for type 2 AGN, since we expect very few type 1 AGNs in small-area JWST surveys (Yang et al. 2023; Kirkpatrick et al. 2023). We also vary the AGN polar dust temperature, allowing $\mathrm{T_{dust}} = 100, 500, 900, 1300$ K. The remaining `SKIRTOR` parameters generally have minor effects on SED shapes (Yang et al. 2019), so we adopt default values to reduce the necessary computation time. Finally, we choose a redshift grid evenly spaced in $\log(1+z)$ with 100 steps from $z = 0.01-8.0$. We adopt this redshift range to fully cover the spread of MEGA redshifts. In total, this `CIGALE` parameter space created $\approx 10^9$ galaxies, which we uniformly downsample to N$= 10^6$ due to the computational limitations of training machine learning models on large datasets[1].

To validate the mock set of `CIGALE` galaxies generated from the above `CIGALE` parameter space, we compare the colors and photometric observations of the mock galaxies to those of galaxies observed in MEGA. Figure 2 demonstrates that the mock galaxies reproduce the observed color distributions of the MEGA observations across all redshifts. In Appendix A, we investigate the `AGNBoost` performance of MEGA galaxies that lie outside the `CIGALE` color space. We also directly compare the photometric observations of the `CIGALE` simulations to MEGA in Figure 3. The `CIGALE` simulations fully cover the range of MEGA photometric observations, but with different distribution shapes. In our testing, we found that sampling the distributions of mock photometry to match that of the real MEGA galaxies had no major effect on `AGNBoost` predictive performance.

[1] We have made this mock `CIGALE` catalog publicly available on the Harvard Dataverse (Hamblin 2025)

The complete mock sample is divided following standard machine learning practices (Hastie et al. 2001): (60%) for training, (20%) for model validation during model tuning, (20%) for final testing. This division ensures sufficient statistics in each subset while maintaining independence between training and evaluation. Each subset is carefully constructed to maintain the same distribution of physical parameters ($\mathrm{frac_{AGN}}$, redshift) as the full sample.

2.2. *MEGA Observations*

We test the performance of our models on data from the MIRI EGS Galaxy and AGN (MEGA) survey (PI A. Kirkpatrick; Backhaus et al. 2025), a four-band MIRI survey with 25 pointings of the Extended Groth Strip (EGS) field. Three of the pointings used only the three reddest filters (F1000W, F1500W, F2100W), while the remainder also added a blue filter (F770W). MEGA builds upon existing coverage of the EGS field (Davis et al. 2007), providing MIRI observations for 68.9% of the NIRCam area from the Cosmic Evolution Early Release Science Survey (CEERS; Finkelstein et al. 2025). For full details of the observations, reductions, and catalogs see Backhaus et al. (2025).

We cross-match our MEGA sources with the CEERS NIRCam catalog of Finkelstein et al. (2023), matching to the nearest object within 0.2", which includes HST+NIRCam photometry and photometric redshifts derived from `EAZY` (Brammer et al. 2008) fitting of HST+NIRCam photometry. These redshifts are the peaks of the redshift posterior probability distributions given by `EAZY`. We also cross-match to the CANDELS (Koekemoer et al. 2011; Grogin et al. 2011) EGS redshift catalog of Kodra et al. (2023) to get spectroscopic redshifts where available. In total, there are 2803 sources with NIRCam counterparts, but we limit our sample to the 701 sources with photometric redshift estimates and detections of signal-to-noise above 3 in F2100W+F1500W+F770W, 217 of which have spectroscopic redshifts. Of these 701 sources, only 3 are missing a single photometric observation. We exclude these 3 from our work, but in Sections 3.4 and 4.5, we detail and test the ability to reliably characterize sources with missing photometry by using statistical imputation. Since the MEGA sources do not have existing estimates of $\mathrm{frac_{AGN}}$, in Section 4.2 we use `CIGALE` to fit these sources with the parameters from Table 1.

3. METHODOLOGY

3.1. `XGBoostLSS` *Framework*

The framework of `AGNBoost` is based on the machine learning algorithm `XGBoostLSS` (März 2019), which extends `XGBoost` (eXtreme Gradient Boosting; Chen & Guestrin 2016), an ensemble boosted decision tree algorithm, to probabilistic forecasting.

Decision trees are a non-parametric method with a hierarchical tree-like structure that learn simple decision rules from data. Broadly, ensemble models are machine learning methods that combine the predictions of multiple base learners, such as decision trees, in some way (e.g., averaging, voting, stacking, etc.) to achieve a more effective overall model (Zhou 2012; Ganaie et al. 2021). The generalization ability (i.e., the model's ability to perform on unseen data) of the overall ensemble is generally better than that of any of the base learners, and ensemble methods are less prone to falling into local optima (Dietterich 2000). Boosting methods, such as `XGBoost`, are a class of ensemble techniques that aim to reduce model bias by combining sequentially trained base learners into one powerful learner.

While `XGBoost` is a powerful algorithm, it is statistically limited to modeling only the conditional mean $\mathbb{E}(Y|X = x)$ and treats higher moments of the conditional distribution as fixed nuisance parameters (März 2019). These assumptions are only valid when data is symmetrically Gaussian with a constant variance, but real observations are not generally so well behaved and exhibit characteristics of variable higher moments (e.g., heterosekdasticity, skewness, or kurtosis). `XGBoostLSS` bridges this statistical modeling gap by connecting `XGBoost` to Generalized Additive Models for Location Scale and Shape (GAMLSS; Stasinopoulos & Rigby 2007) to predict the entire conditional distribution $F_Y(y|x)$.

`XGBoostLSS` has a few built-in features that make it appealing for application to astronomical data. There is built-in support for both L1 (Lasso regression) and L2 (Ridge regression) regularization to avoid model overfitting to spurious features or noise. It also has the ability to handle missing data through a technique known as sparsity-aware split finding, in which a default direction is assigned at every branch in a tree that results in the lowest corresponding loss.

We train separate `XGBoostLSS` models for estimation of $\mathrm{frac_{AGN}}$ and redshift. Since $\mathrm{frac_{AGN}}$ is intrinsically bounded $[0, 1)$, we choose to estimate a zero-inflated beta distribution for $\mathrm{frac_{AGN}}$. The beta distribution, commonly used in Bayesian inference, is particularly flexible at modeling parameters bounded on the interval (0,1), but is not defined at zero. Per the `CIGALE`

Table 1. `CIGALE` Simulation Parameters

Module	Parameter	Symbol	Values
Star formation history	Stellar e-folding time	$\tau_{\rm star}$	0.1, 0.5, 1, 5 Gyr
`sfhdelayed`	Stellar age	$t_{\rm star}$	0.5, 1, 3, 5, 7 Gyr
	Burst mass fraction	$f_{\rm burst}$	0.0, 0.02, 0.1
Simple stellar population	Initial mass function	–	Chabrier (2003)
`bc03`	Metallicity	Z	0.02
Nebular emission	Ionization parameter	$\log U$	-2.0
`nebular`	Gas metallicity	$Z_{\rm gas}$	0.02
Dust attenuation	Color excess of nebular lines	$E(B-V)_{\rm line}$	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
`dustatt_modified_starburst`			0.6, 0.7, 0.8, 0.9, 1.0
	ratio between line and continuum $E(B-V)$	$\frac{E(B-V)_{\rm line}}{E(B-V)_{\rm cont}}$	1
	PAH mass fraction	$q_{\rm PAH}$	1.77, 3.19, 5.26
Galactic dust emission	Minimum radiation field	$U_{\rm min}$	0.1, 1.0, 10
`dl2014`	Fraction of PDR emission	γ	0.01, 0.1, 0.2
AGN (UV-to-IR) emission	Viewing angle	$\theta_{\rm AGN}$	70°
`skirtor2016`	AGN contribution to IR luminosity	$\rm frac_{AGN}$	0.1–0.9 (step 0.1), 0.99
	Wavelength range where $\rm frac_{AGN}$ is defined	$\lambda_{\rm AGN}$	3–30 μm
	Extinction in polar direction	$E(B-V)_{\rm polar}$	0.03, 0.1
	Polar dust temperature	$\rm T_{dust}$	100, 500, 900, 1300 K
Redshift+IGM	Source redshift	z	0.01–8.0 (100 steps)
`redshifting`			

NOTE—For parameters not listed here, we adopt default values.

modeling performed in this work, galaxies are allowed $\rm frac_{AGN} = 0$ (i.e., entirely SFG), which the zero-inflated beta distribution accounts for by adding a Bernoulli distribution at $\rm frac_{AGN} = 0$. Conversely, redshift is theoretically bounded $[0, \inf)$, but since this work is focused on a large redshift range ($0.01 < z < 8$), we choose to transform redshift to the range (0,1) with a modified sigmoid transformation ($\Phi(z) = 2/(1 + e^{-az}) - 1$; $a = 0.4$) so that we can model the transformed redshift with a beta distribution. This transformation provides variable compression of redshift onto the (0,1) interval, controlled by the parameter a which was chosen to create better resolution in the intermediate redshift range ($2 < z < 5$), where PAH features redshift out of the NIRCam+MIRI band coverage, making redshift estimation more challenging due to the loss of key spectral diagnostics. Compared to the scale factor transformation ($a(z) = 1/(1+z)$), this modified sigmoid transformation better spreads redshifts over the unit interval.

Each model shares the same inputs: 7 NIRCam bands (F115W, F150W, F200W, F277W, F356W, F410M, F44W) + 4 MIRI bands (F770W, F1000W, F1500W, F2100W), 55 derived colors from all unique filter combinations to avoid using unnecessary reciprocal colors, and the square of those 55 colors. We include the square of the 55 colors in order to emphasize the effects of continuum features in color space, and to enhance the models' ability to learn non-linear color relationships. We package our trained`XGBoostLSS` models into a repository of publicly available Python code, `AGNBoost`[2].

3.2. *Model Training and Optimization*

Due to the large size of our training set, a methodical approach was necessary in order to tune `AGNBoost`'s models in a reasonable amount of time. Broadly, to tune the configurable variables of our models, known as hyperparameters, we employ a multi-stage search procedure, where in the first stage we separately tune tree parameters (the last three rows of Table 2) and boosting parameters (the remaining rows of Table 2), and in the final stage we find the optimal number of boosting iterations. The first stage of tuning is performed with 2-fold cross-validation and a high learning rate (0.8) on the combination of the training and validation sets, and the second tuning stage is performed with a low training

[2] https://github.com/hamblin-ku/AGNBoost

Table 2. `AGNBoost` Hyperparameters

Hyperparameter	Description	Range Explored	Model	Optimal Value
max_depth	Max depth of trees	[3,10]	frac_{AGN}	5
			z	10
gamma	Min loss reduction to further partition leaf node	$[10^{-8}, 40]$	frac_{AGN}	6.87e-02
			z	5.83e-6
min_child_weight	Min sum of hessian needed to partition	[1,250]	frac_{AGN}	26
			z	21
lambda	L2 regularization on weights	[1,150]	frac_{AGN}	99.17
			z	6.14
alpha	L1 regularization on weights	$[10^{-3}$,100]	frac_{AGN}	2.96e-6
			z	5.23e-6
subsample	Subsample ratio of training instances per boosting iteration	[0.7, 1.0]	frac_{AGN}	1.00
			z	0.76
colsample_bytree	Subsample ratio of columns when constructing each tree	[0.5, 1.0]	frac_{AGN}	0.85
			z	0.81
tree_method	Tree construction algorithm in XGBoost	['hist','approx']	frac_{AGN}	'hist'
			z	'approx'
stabilization	Stabilization of gradients and Hessians to improve model convergence	['none','L2', 'MAD']	frac_{AGN}	'none'
			z	'none'
Response_fn	Transformation function of distribution parameters	['softplus', 'exp']	frac_{AGN}	'softplus'
			z	'softplus'

NOTE—For `XGBoostLSS`/`XGBoost` hyperparameters not listed here, we use default values.

rate (0.01) on just the training set with the validation set used to stop the training procedure when the loss on the validation set increases (known as early stopping). The primary benefit of this approach is that the high learning rate of the first two stages allows fast model tuning, while the final stage can use these now confidently identified parameters to find the optimal boosting duration. This process is outlined Figure 4.

For each stage of hyperparameter tuning, we adopt a Bayesian approach with `Optuna` (Akiba et al. 2019), an open source hyperparameter optimization framework. Unlike traditional grid search methods, which try all possible combinations within a given hyperparameter space, Bayesian hyperparameter tuning efficiently explores the hyperparameter space by continuously updating the parameter probability density functions based on the results of all trials.

3.3. *Uncertainty Quantification*

`XGBoostLSS` provides two primary means of quantifying uncertainties of estimates. The conditional distributions modeled by `XGBoostLSS` capture the uncertainty that arises due to the inherent and unpredictable randomness of the data, known as the aleatoric uncertainty (Hüllermeier & Waegeman 2021). The epistemic uncertainty, or the uncertainty due to a lack of model knowledge, is estimated with the virtual ensemble method of Ustimenko et al. (2020). In brief, a virtual ensemble is created from M number of sub-models within each trained `XGBoostLSS` model and each sub-model is used to make predictions. The variance of these M observations is used as a robust estimate of the epistemic uncertainty. Notably, virtual ensembles are obtained from *one* already trained model, while creating a true ensemble of N `XGBoostLSS` models would require N times the computational cost, so this approach can be considered computationally "free."

We also derive prediction uncertainties due to source photometric uncertainty. For each source, we create 100 mock sources with Monte Carlo, by considering the measured fluxes and associated flux errors as the mean and standard deviation of a normal flux distribution. With these 100 mock sources we create 100 point estimates from each model, corresponding to the peak of their beta

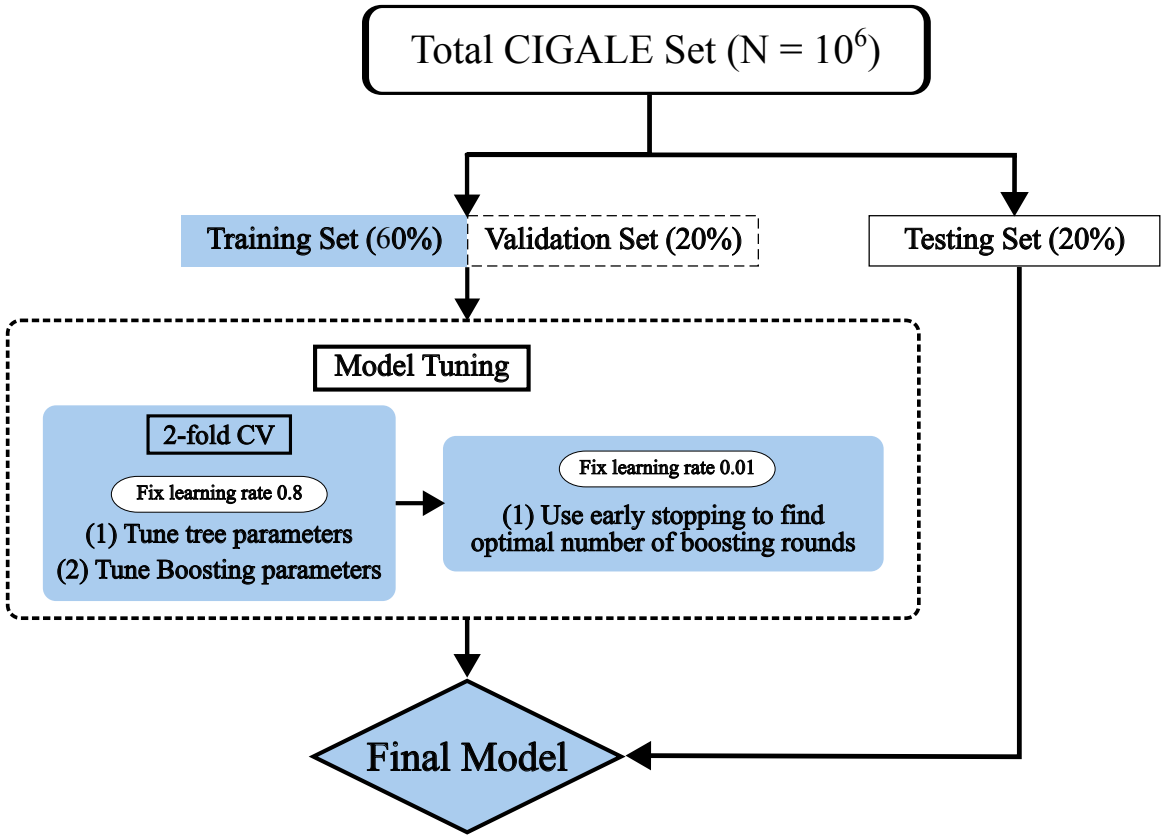

Figure 4. Flowchart illustrating the `AGNBoost` training procedure. The multi-stage approach optimizes computational efficiency by first performing coarse hyperparameter tuning, then fine-tuning the number of boosting rounds. In the first stage of tuning, the training and validation sets are combined for 2-fold cross-validation (CV) to identify optimal hyperparameters. In the second stage of tuning, these hyperparameters are used to train models on the training set alone, with the validation set used for early stopping to determine the optimal number of boosting rounds (i.e., when validation loss begins to increase). The final model is trained using these optimized parameters on the combined training and validation datasets to maximize the use of available training data.

distributions, and use the standard deviation of these point estimates to quantify an uncertainty due to photometric uncertainty for every source in each `AGNBoost` model. The total uncertainty for each source is derived by summing these three measures in quadrature.

3.4. *Missing Photometry Imputation*

It is common for observed MIRI sources to be missing detections in one or more bands, due to the way that bright PAH and absorption features shift through the MIRI bands. While `AGNBoost` is able to make predictions for sources with missing photometric observations, we also test whether replacing these missing values with statistical imputation can improve the model performance. We implement and test a Generative Adversarial Network (GAN) for photometry imputation, using the `SGAIN` algorithm of Neves et al. (2021)[3]. `SGAIN` benefits from increased stability and performance over comparative GAN-based methods. `SGAIN` consists of a generator and a discriminator, where the generator's goal is to accurately impute missing data and the discriminator's goal is to distinguish between observed and imputed data. This adversarial training process ensures that the generator's imputation performance increases over training.

We integrate `SGAIN` into a separate module of `AGNBoost`, allowing users to impute missing photometry from their catalogs before performing data analysis. We use cross-validation to optimize the mini-batch size, the α parameter, and the optimal number of iterations of `SGAIN`. Following standard practices, photometric observations are standardized to the range (0,1) to avoid problems due to scale. The `SGAIN` imputation process is performed 100 times. To account for imputation uncertainty due to photometric uncertainty, at each iteration the existing flux measurements are randomly sampled from a normal distribution of mean equal to the flux

[3] https://github.com/dtneves/ICCS_2021

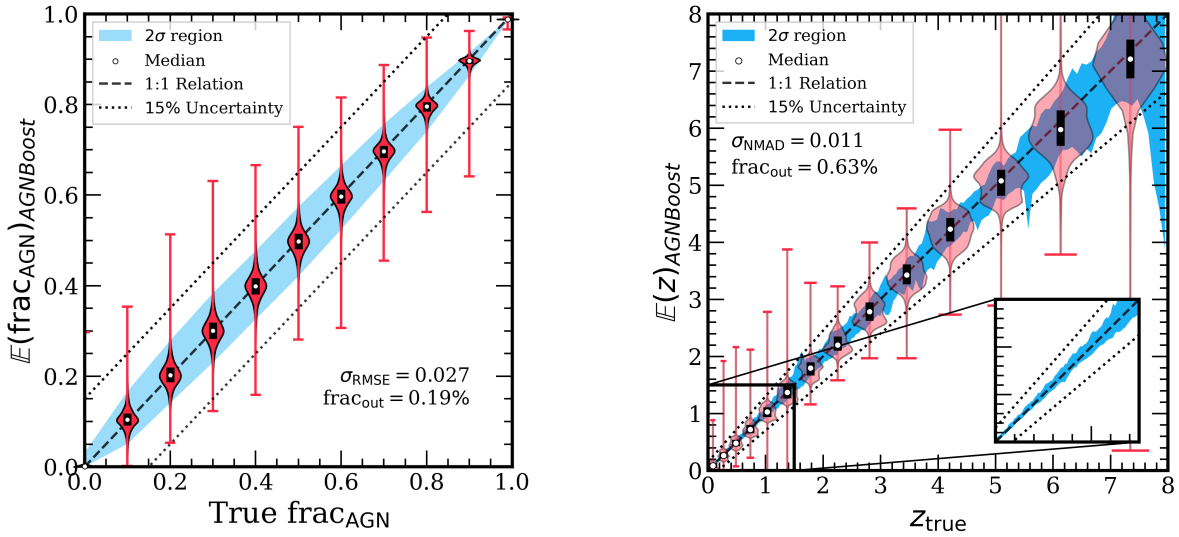

Figure 5. Expectation values of `AGNBoost` predictions on the test set of mock CIGALE data, for $\mathrm{frac_{AGN}}$ (left) and redshift (right) models. The violin plots show the distribution of predicted expectation values at each bin of true value, with white dots marking medians, solid black lines indicating 25th-75th percentiles, and whiskers extending to the extrema. The light blue shading represents the 2σ (95%) interval, and the black dashed and dotted lines correspond to the 1:1 relation and 15% uncertainty regions, respectively. Both models demonstrate excellent performance with predictions tightly centered on the 1:1 relation and sub-1% outlier fractions.

value and standard deviation equal to the flux error. The final imputed value is calculated as the average of these imputation trials, and its error is given by the standard deviation of these trials.

4. RESULTS AND DISCUSSION

4.1. *Performance on Mock Data*

The results in Figure 5 show the model performance on the `CIGALE` test set (20% of the total sample). To evaluate the performance of each model, we calculate the root mean square error (RMSE; $\sigma_{\mathrm{RMSE}} = \sqrt{\mathrm{mean}\{(x_{\mathrm{pred}} - x_{\mathrm{true}})^2\}}$), and the outlier fraction ($\mathrm{frac_{out}}$). For the $\mathrm{frac_{AGN}}$ model, we define $\mathrm{frac_{out}}$ as $|\Delta \mathrm{frac_{AGN}}| > 0.15$, and for the redshift model we define the $\mathrm{frac_{out}}$ as $|\Delta z| \,/\, (1 + z_{\mathrm{true}}) > 0.15$. Across all values of $\mathrm{frac_{AGN}}$, only $\sim 0.19\%$ of the test sample lie outside the 15% uncertainty region, and we find the RMSE to be $\sigma_{\mathrm{RMSE}} = 0.027$. The redshift model shows similar performance, with a tight spread around the 1:1 relation out to $z = 8$ and a normalized median absolute deviation (NMAD; $\sigma_{\mathrm{NMAD}} = 1.48 \times \mathrm{median}\{|\Delta z - \mathrm{median}(\Delta z)|/(1 + z_{\mathrm{spec}})\}$) of $\sigma_{\mathrm{NMAD}} = 0.011$. Only $\sim 0.63\%$ of mock sources lie outside the drawn 15% uncertainty region.

To determine how sensitive `AGNBoost` is to photometric uncertainty, we add realistic photometric uncertainties to the mock data and use these photometric uncertainties to generate new mock sources perturbed by photometric uncertainty. We use the final MEGA sample to get realistic measures of photometric uncertainty by creating linear fits in log-uncertainty vs log-photometry space. We then use these fits to get estimates of the photometric uncertainty for each source in the mock catalog, and then use these uncertainties to perturb the mock fluxes according to a normal flux distribution. Figure 6 shows the resulting `AGNBoost` predictive performance with the added photometric uncertainties. For $\mathrm{frac_{AGN}}$ estimation, `AGNBoost` maintains median predictions on the 1:1 relation with $\sigma_{\mathrm{RMSE}} = 0.120$, but the 15% outlier fraction has increased considerably to $\mathrm{frac_{out}} = 11.06\%$. For photometric redshift estimation, the introduced photometric uncertainty causes `AGNBoost` to systematically underestimate redshift for $z > 4$, but `AGNBoost` maintains median predictions following the 1:1 relation for $z < 3.5$. With the added photometric uncertainty, we find the photometric redshift 15% outlier fraction to be $\mathrm{frac_{out}} = 17.66\%$ and $\sigma_{\mathrm{NMAD}} = 0.023$. We note that the NMAD has only increased by a factor of two since the majority of the mock sources have $z < 2$, where the predicted photometric redshift spread is low.

4.2. *Performance on MEGA Sources with Complete Observations*

Here, we evaluate how `AGNBoost` performs on the final set of 698 MEGA galaxies with no missing photometric observations. Since the MEGA catalog does not have existing estimates of $\mathrm{frac_{AGN}}$, we use `CIGALE` to

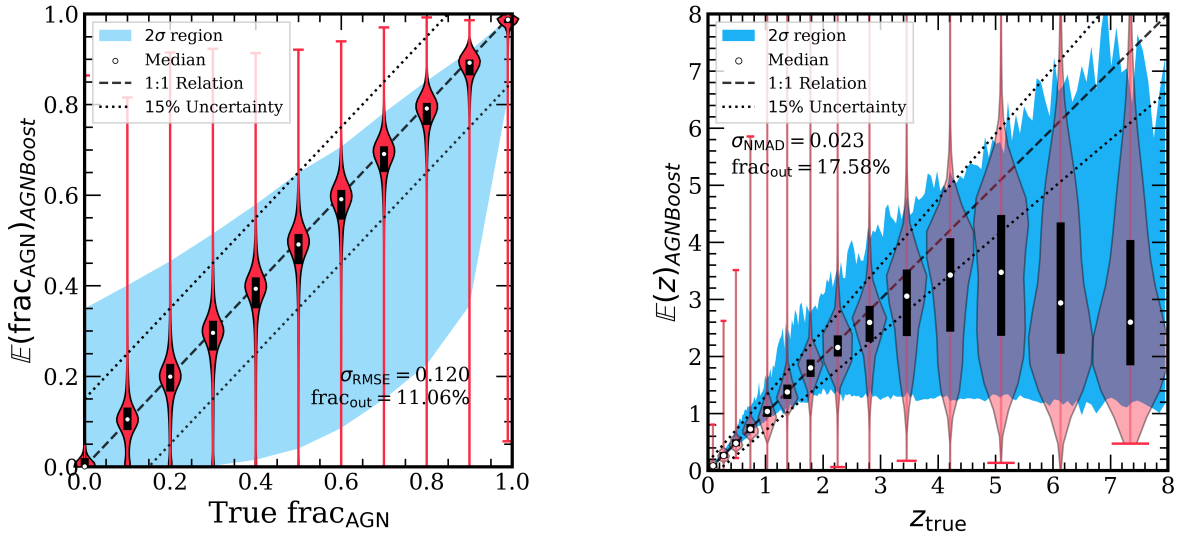

Figure 6. Expectation values of `AGNBoost` predictions on the test set of mock CIGALE data with realistic photometric uncertainty added, for $\mathrm{frac_{AGN}}$ (left) and redshift (right) models. The violin plots show the distribution of predicted expectation values at each bin of true value, with white dots marking medians, solid black lines indicating 25th-75th percentiles, and whiskers extending to the extrema. The light blue shading represents the 2σ (95%) interval, and the black dashed and dotted lines correspond to the 1:1 relation and 15% uncertainty regions, respectively. For $\mathrm{frac_{AGN}}$ estimation, introducing realistic photometric uncertainty opens up the outlier ranges considerably, but the median 1:1 relation performance is maintained. For photometric redshift estimation, the photometric uncertainties cause `AGNBoost` to underestimate the redshift for $z > 4$, but the median 1:1 relation is maintained for $z < 3.5$.

fit the MEGA catalog. We adopt the same `CIGALE` parameter space presented in Table 1, but use the existing redshifts for each source so that `CIGALE` does not fit for redshift. In total, 246 ($\sim$ 40%) sources have fits with an acceptable reduced chi-squared (χ^2_{red}) of $0.5 < \chi^2_{\mathrm{red}} < 2$. We adopt the Bayesian output values of $\mathrm{frac_{AGN}}$ instead of the best-fit output. These Bayesian outputs are the probability-weighted values across all the `CIGALE` models, and are generally more robust than the best-fit values from the minimum-χ^2 model alone (Yang et al. 2023). Figure 7 shows the distribution of Bayesian $\mathrm{frac_{AGN}}$ estimates. The results of this SED-fitting should not be interpreted as ground-truth, but instead serve as a point of comparison of what performance can be expected of `CIGALE` SED-fitting without spending the time to tune the fitting parameter space.

The left side of Figure 8 compares the redshift estimates of `AGNBoost` to the photometric redshifts from Finkelstein et al. (2023). There is agreement between the two redshift estimates at $z < 2$, but disagreement at $z > 3$. It is important to note that the Finkelstein et al. (2023) photometric redshifts were estimated using HST+NIRCam observations only ($0.6 - 4.4\,\mu$m), meaning that they are able to trace the Balmer break as early as $z \gtrsim 0.5$, but the NIRCam+MIRI coverage ($1.2 - 21\,\mu$m) of `AGNBoost` means that `AGNBoost` can only trace the Balmer break for $z \gtrsim 1.9$. Additionally, `AGNBoost` is able to trace the mid-IR PAH features up to $z \sim 2.4$

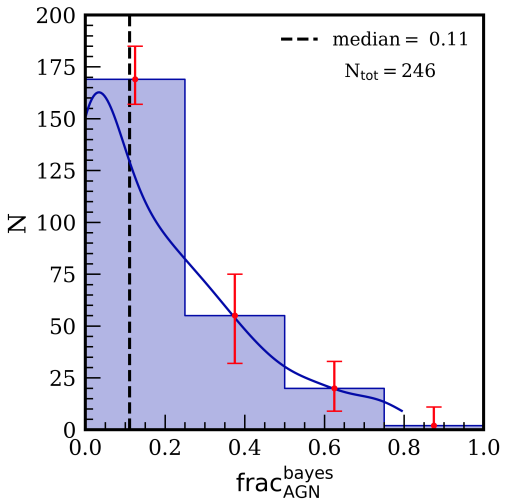

Figure 7. Histogram of `CIGALE`-derived $\mathrm{frac_{AGN}}$ values for the 246 MEGA sources with well-constrained `CIGALE` fits ($0.5 < \chi^2_{\mathrm{red}} \leq 2$). Error bars represent uncertainties propagated from the CIGALE Bayesian $\mathrm{frac_{AGN}}$ uncertainties. The vertical dashed line indicates the median $\mathrm{frac_{AGN}}$ value, and the blue curve shows a kernel density estimate (KDE) fit to the distribution.

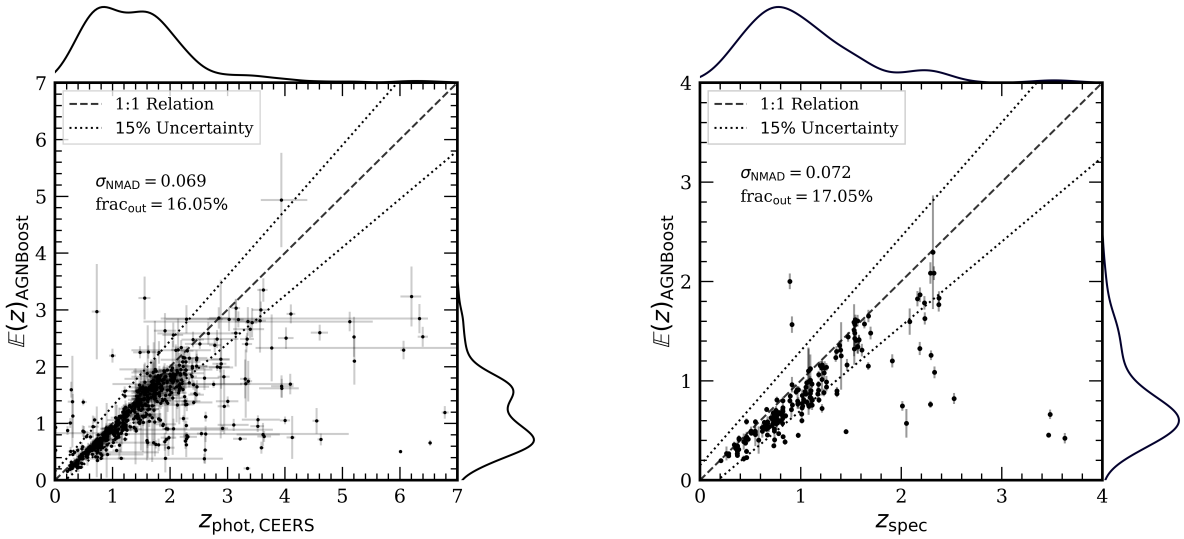

Figure 8. Expectation values of `AGNBoost` redshift predictions compared to existing photometric redshifts (left) and spectroscopic redshifts (right). The left panel shows `AGNBoost` predictions against `EAZY` photometric redshifts from Finkelstein et al. (2023) for the full MEGA sample, with photo-z uncertainties indicating 68% confidence intervals. The right panel compares `AGNBoost` predictions to spectroscopic redshifts for 217 MEGA sources. The black dashed lines show the 1:1 relation and dotted lines indicate 15% uncertainty regions. Marginal distributions along each axis show the underlying redshift distributions, with redshift quality metrics ($\sigma_{\rm NMAD}$ and $\rm frac_{out}$) displayed. `AGNBoost` redshifts show good agreement with spectroscopic values, particularly for $z < 2$, but tend to underestimate redshifts at $z > 3.5$ compared to existing photometric redshift estimates.

The right side of Figure 8 shows the photometric redshift estimation performance of `AGNBoost` compared to the available spectroscopic redshifts. For the sample with spectroscopic redshifts, we obtain an acceptable NMAD of $\sigma_{\rm NMAD} = 0.074$ and a outlier fraction of $\rm frac_{out} = 16.67\%$. The majority of these outliers are above $z_{\rm spec} > 2$ where all but the 6.2 μm PAH complex have redshifted out of the MIRI bands. We note that the 3 catastrophic outliers with $z_{\rm spec} > 3$ have Finkelstein et al. (2023) photometric redshifts in agreement with those from `AGNBoost`. The SEDs of these sources exhibit a weak Balmer break, which is likely causing the photometric redshifts to be underestimated. This is supported by the redshift probability distribution functions of these sources from Finkelstein et al. (2023), which show no high-redshift peaks.

Figure 9 shows the $\rm frac_{AGN}$ performance of `AGNBoost` on the subset of 246 MEGA galaxies with $0.5 < \chi^2_{red} < 2$ that are missing no photometric observations. The majority of sources that are identified as SFGs from `CIGALE` fitting are also identified as SFGs with `AGNBoost`. That is, 173 ($\sim$ 70%) galaxies have $\rm frac_{AGN} < 0.3$ from `AGNboost` and $\rm frac^{bayes}_{AGN} < 0.3$ from `CIGALE`. We adopt the selection criteria of Kirkpatrick et al. (2017) and classify sources with $\rm frac_{AGN} > 0.3$ as a candidate AGN or AGN+SF composite (hereafter simply referred to as AGN candidates). There are 41 ($\sim$ 17%) sources in Figure 9 that `AGNBoost` identifies as AGN candidates. In

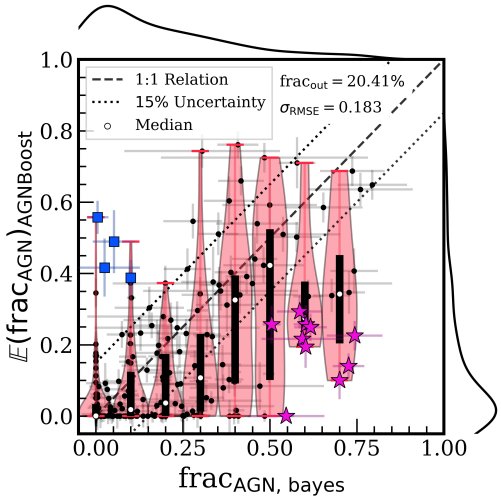

Figure 9. Expectation values of `AGNBoost` $\rm frac_{AGN}$ predictions compared to `CIGALE` Bayesian $\rm frac_{AGN}$ estimates for the 246 MEGA sources with complete photometry and well-constrained `CIGALE` fits. The violin plots show the distribution of `AGNBoost` predictions at each binned `CIGALE` value, with white dots marking medians, solid black lines indicating 25th-75th percentiles, and whiskers extending to extrema. Blue squares highlight sources classified as AGNs by `AGNBoost` but as SFGs by `CIGALE`, while purple stars show the reverse (i.e., AGN by `CIGALE`, SFG by `AGNBoost`). `AGNBoost` $\rm frac_{AGN}$ estimates show broad agreement with `CIGALE`.

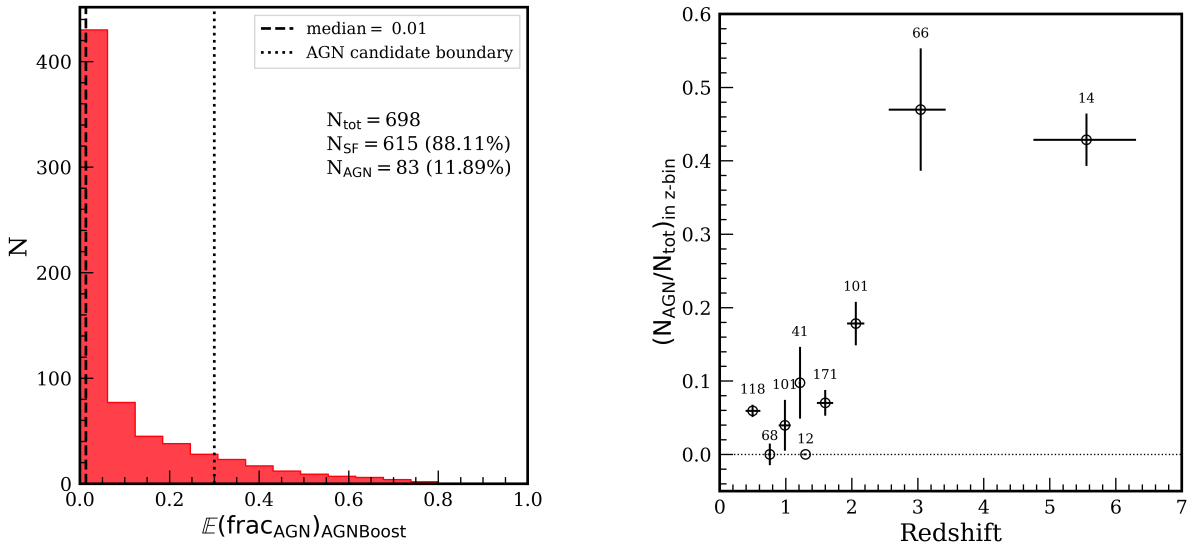

Figure 10. Left: Histogram of $\mathrm{frac_{AGN}}$ values for all 698 MEGA galaxies. Using $\mathrm{frac_{AGN}} > 0.3$ as the AGN threshold, `AGNBoost` identifies 83 ($\sim 12\%$) AGN candidates and 615 ($\sim 88\%$) SFGs. Right: The AGN population fraction ($\mathrm{frac_{AGN}} > 0.3$) in redshift bins of $\Delta z = 0.5$, requiring a minimum of 10 galaxies per bin. Points are plotted at the median redshift of each bin, with redshift error bars showing 25th-75th percentiles and the AGN population fraction uncertainties are showing the average shifting between across the $\mathrm{frac_{AGN}} = 0.3$ threshold due to `AGNBoost` $\mathrm{frac_{AGN}}$ uncertainty. Redshift bins are determined using the Bayesian blocks algorithm (Scargle 1998), with the number of galaxies in each bin labeled above each point. `AGNBoost` identifies a significant high-redshift AGN population, with $\sim 45\%$ of galaxies at $3 < z < 6$ classified as AGN candidates. Validation of these AGN candidates will require X-ray confirmation or spectroscopic confirmation through broad lines and high ionization lines.

total, `AGNBoost` estimates a similar number of AGN as `CIGALE`.

We find 14 sources with significant disagreements in estimated $\mathrm{frac_{AGN}}$ values. There are 10 sources (plotted as purple stars in Figure 9) that `AGNBoost` identifies as SFGs ($\mathrm{frac_{AGN}} < 0.3$) but `CIGALE` identifies as AGNs ($\mathrm{frac_{AGN}^{bayes}} > 0.5$), and 4 sources (plotted as blue squares Figure 9) that `AGBoost` identifies as AGNs ($\mathrm{frac_{AGN}} > 0.35$) but `CIGALE` identifies as SFGs ($\mathrm{frac_{AGN}^{bayes}} < 0.2$). To assess these discrepancies, we re-fit the 10 sources identified as AGNs by `CIGALE` without an AGN component entirely, and re-fit the 4 identified as SFGs by `CIGALE` with a forced AGN component. All 10 sources identified as AGNs by `CIGALE` have $0.5 < \chi^2_{red} < 2$ for the SED fits without an AGN component. Upon visual inspection, we find that 5 of these sources have better SED fits without an AGN component, and the remaining 5 are fit equally well with or without an AGN component. Only 1 of the sources identified as a SFG by `CIGALE` has a worse fit with a forced AGN component, and the remaining 3 are fit equally well with or without an AGN component. We conclude that the `CIGALE` fits for these sources are unreliable.

Thus, it is evident that `AGNBoost` is, at the very least, successfully identifying AGN candidates. The left panel of Figure 10 shows that, across the entire MEGA sample of 698 sources, `AGNBoost` finds 83 ($\sim 12\%$) sources that are AGN candidates. Notably, `AGNBoost` finds a large fraction of AGNs at $3 < z < 6$ of $\sim 45\%$, significantly higher than the of $\sim 10\%$ fraction of broad-line AGNs found by Maiolino et al. (2024) at $4 < z < 6$. Validation of these AGN candidates found by `AGNBoost` will require X-ray confirmation or spectroscopic confirmation through broad lines and high ionization lines. The reduced size of this candidate sample will make this necessary follow-up more achievable.

4.3. *Photometry Imputation Performance*

Following the procedure of Luo et al. (2024), we test the photometry imputation performance by creating new datasets with different missing rates (defined as the number of total missing photometric observations per source) from the final set of 698 MEGA sources with no missing photometric observations. These subsamples are created by randomly deleting photometric observations for each source according to the missing rate of the sub-sample. With this method, we create three new datasets: a one-band missing sample (OMS), a two-band missing sample (TwMS), and a three-band missing sample (TrMS). To assess the accuracy of imputation, we calculate the normalized root mean square error (NRMSE; $\sigma_{\mathrm{NRMSE}} = \sigma_{\mathrm{RMSE}}/\mathrm{mean}\{x_{\mathrm{pred}}\}$) of each band

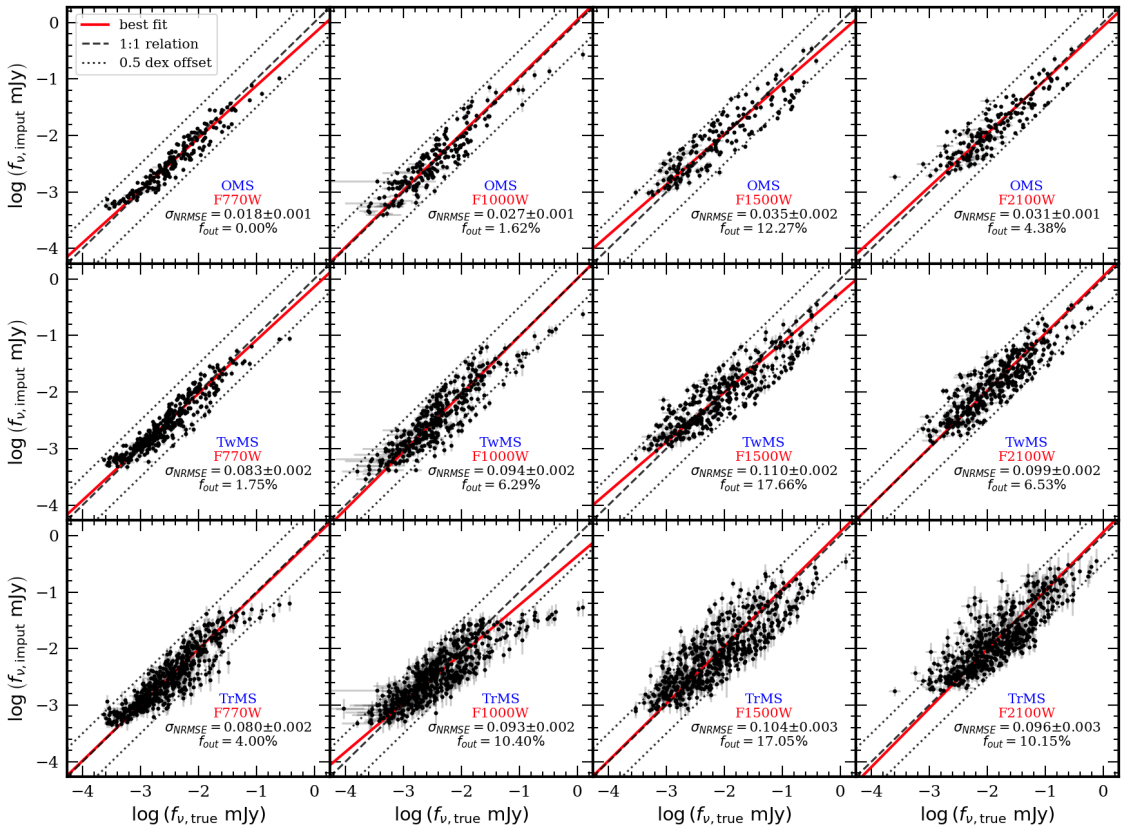

Figure 11. `SGAIN` photometry imputation performance for samples missing one band (OMS), two bands (TwMS), and three bands (TrMS). The black dashed line shows the 1:1 relation, the dotted line indicates the 0.5 dex uncertainty region, and the red line shows the line of best fit. The normalized root mean square error (σ_{NRMSE}) and outlier fractions are displayed for each band. `SGAIN` demonstrates notably strong performance when only one MIRI band is missing. While imputation performance degrades with the number of missing MIRI bands, `SGAIN` still performs relatively well even when three of the four MIRI bands are missing.

for each new dataset. Figure 11 shows the imputation performance for the OMS, TwMS, and TrMS samples. The imputation process becomes more uncertain as the number of missing bands increases, but the outlier fraction (within 0.5 dex) is still below 20% in all bands when there are three missing MIRI bands in Figure 11.

4.4. *Effects of Imputed Photometric Observations on* `AGNBoost` *Predictions*

Figure 12 and Figure 13 show how replacing missing MIRI photometric observations with `SGAIN` imputation improves `AGNBoost`'s ability to estimate $\mathrm{frac_{AGN}}$ and photometric redshift, respectively. Notably, when there are one or two missing MIRI bands, photometry imputation cuts the outlier fraction of $\mathrm{frac_{AGN}}$ estimation in half. `AGNBoost` systematically underestimates the $\mathrm{frac_{AGN}}$ when running with missing values. Appendix B investigates the wavelength-dependent effects of missing bands on $\mathrm{frac_{AGN}}$ estimation. For photometric redshift estimation, imputing one missing photometric band reduces the outlier fraction by a factor of 3, and by a factor of 2 when two or three bands are imputed. The photometric imputation recovers `AGNBoost`'s good redshift performance for $z < 2$, but using `AGNBoost` with missing MIRI bands introduces a positive z-offset for $z < 2$. `AGNBoost`'s photometric redshift estimation is much more sensitive to missing and imputed values than its $\mathrm{frac_{AGN}}$ estimation.

4.5. *Feature Importance and Model Interpretation*

XGBoostLSS provides native integration with SHapley Additive exPlanations (SHAP), a framework for interpreting black box models (Lundberg & Lee 2017). SHAP assigns each feature in the model an importance value for individual predictions, allowing interpretation of the model predictions. Of particular use are SHAP's 'beeswarm' plots, which rank features by importance and depict how high/low individual feature values affect predictions. In Figure 14 we show beeswarm plots for both the $\mathrm{frac_{AGN}}$ and redshift models of the top 10

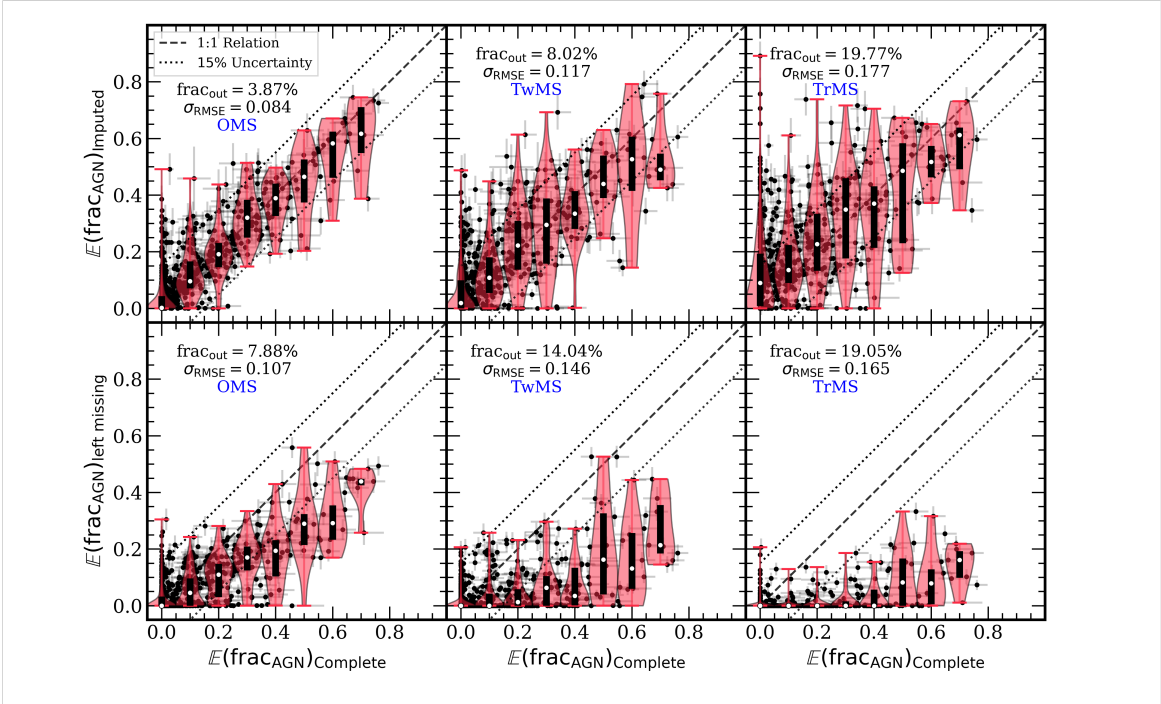

Figure 12. Comparison of `AGNBoost` frac$_{AGN}$ estimation performance with and without `SGAIN` imputation for missing MIRI photometry. The top row shows AGNBoost results when the randomly removed bands in the one missing band (OMS), two missing bands (TwMS), and three missing bands (TrMS) sets are replaced with using `SGAIN`-imputed values. The bottom row shows `AGNBoost` performance when missing bands are left blank (i.e., no imputation). `SGAIN` imputation significantly improves frac$_{AGN}$ estimation compared to running `AGNBoost` with missing photometric data, with the benefit most pronounced when three MIRI bands are absent.

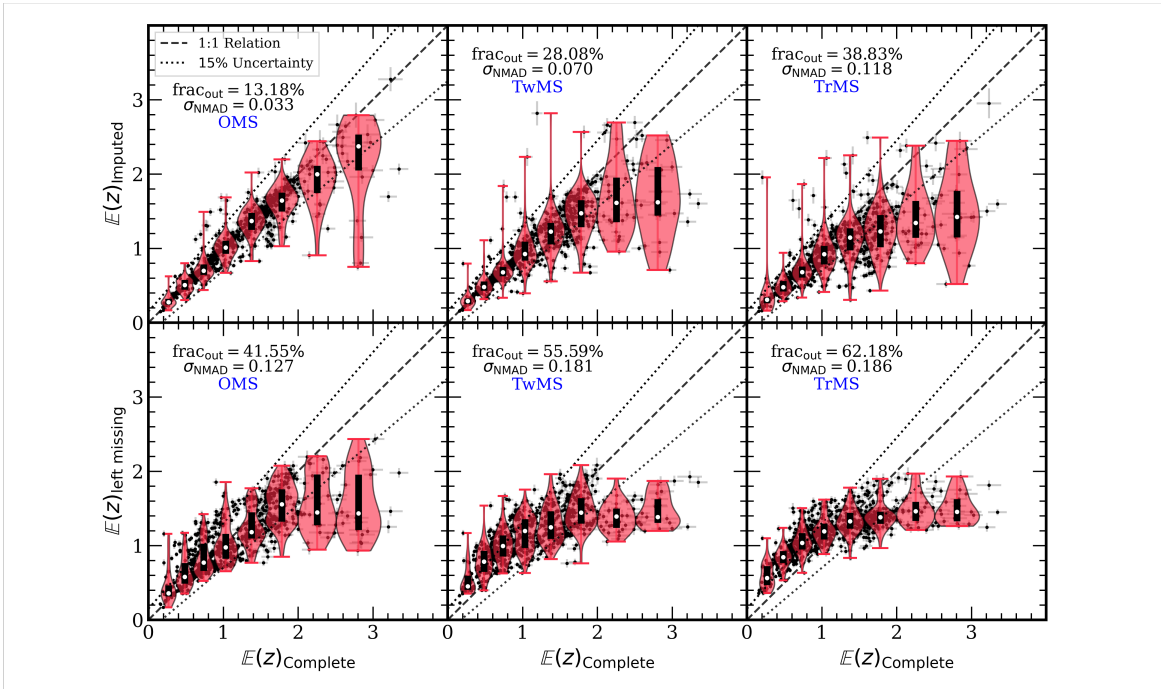

Figure 13. Comparison of `AGNBoost` photometric redshift estimation performance with and without `SGAIN` imputation for missing MIRI photometry. The top row shows `AGNBoost` results using `SGAIN`-imputed values for randomly removed bands: one missing band (OMS), two missing bands (TwMS), and three missing bands (TrMS). The bottom row shows `AGNBoost` performance when missing bands are left blank (i.e., no imputation). Running `AGNBoost` with missing photometric data introduces notable positive z-bias for $z < 1.5$, particularly when two or three bands are absent. `SGAIN` imputation successfully recovers the median 1:1 relation out to $z \sim 3$ when one MIRI band is missing, but underestimates redshift at $z \geq 2$ when two or three MIRI bands are missing.

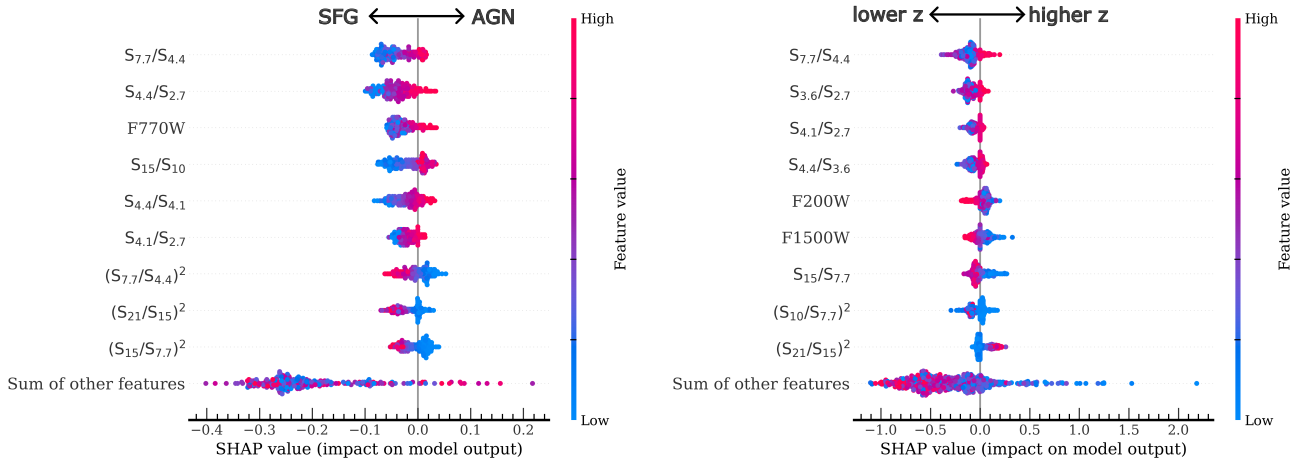

Figure 14. SHAP beeswarm plots showing feature importance for the $\mathrm{frac_{AGN}}$ model (left) and redshift model (right). Features are ranked by importance (top to bottom), with each dot representing an individual source. Colors indicate feature values from low (blue) to high (red), which for color-based features corresponds to bluer and redder colors, respectively. Negative SHAP values push predictions toward low-$\mathrm{frac_{AGN}}$ (i.e., SFGs) or low redshifts, while positive values push predictions toward high-$\mathrm{frac_{AGN}}$ (i.e., AGNs) or high redshifts. MIRI-based features rank among the most important predictors for both models, highlighting the critical role of mid-infrared photometry in AGN identification and redshift estimation.

most important features, where each point is an individual MEGA galaxy colored by the value of the listed features.

For the $\mathrm{frac_{AGN}}$ model, we see that $S_{7.7}/S_{4.4}$, $S_{4.4}/S_{2.7}$, F770W, and S_{15}/S_{10} are the four most important features. For all four of these colors, 'blue' colors are pushing the model towards smaller values of $\mathrm{frac_{AGN}}$ (i.e., non-AGN), whereas 'red' colors push the model towards larger values of $\mathrm{frac_{AGN}}$ (i.e., AGN). The fact that bluer values of S_{15}/S_{10} are pushing the model towards non-AGN suggests that the model is successfully identifying PAH features, and the importance of red $S_{7.7}/S_{4.4}$, $S_{4.4}/S_{2.7}$ for AGN suggests that `AGNBoost` is identifying the characteristic red slope of AGN at these wavelengths. Furthermore we can see that larger flux values in F770W and F1000W also push the model towards larger values of $\mathrm{frac_{AGN}}$. For the redshift model, the top four features are $S_{7.7}/S_{4.4}$, $S_{3.6}/S_{2.7}$, $S_{4.1}/S_{2.7}$, and $(S_{4.4}/S_{3.6})^2$. Notably, all 4 MIRI bands in MEGA appear in the top 10 features at least once for both the $\mathrm{frac_{AGN}}$ and redshift models, confirming that mid-IR information is important for finding AGN and estimating their redshifts (Yang et al. 2023; Pérez-González et al. 2024; Leung et al. 2024; Durodola et al. 2024; Kocevski et al. 2024; Kirkpatrick et al. 2023).

5. SUMMARY

In this work we presented `AGNBoost`, a publicly-available machine learning model utilizing the XGBoostLSS algorithm to estimate $\mathrm{frac_{AGN}}$ and photometric redshift for NIRCam+MIRI galaxies. `AGNBoost` is trained on a large set of mock galaxies from `CIGALE`, which are checked to be representative of real JWST/NIRCam+MIRI photometric observations and colors. We test `AGNBoost` on both a subset of mock galaxies and on a sample of MEGA galaxies. `AGNBoost` provides a computationally fast method of identifying subsamples of candidate AGNs, without needing to devote substantial time to SED fitting of a large catalog. On a catalog of N $\sim$ 1000 sources, `AGNBoost` provides parameter estimates and uncertainties in minutes on a standard laptop. Importantly, `AGNBoost` is able to identify AGNs without knowledge of redshift, which are often not available.

`AGNBoost` provides robust estimates of prediction uncertainty, capturing both the aleatoric uncertainty due to the intrinsic randomness in observations, the epistemic uncertainty due to a lack of model knowledge, and the prediction uncertainty due to photometric uncertainty. `AGNBoost` can also provide feature importance estimates with SHAP plots to understand the model predictions. From these SHAP plots we found evidence of PAH features in the S_{15}/S_{10} color, where 'blue' S_{15}/S_{10} colors push the model predictions to lower values of $\mathrm{frac_{AGN}}$. Importantly, `AGNBoost` provides simple and efficient handling of missing photometric observations, thus providing $\mathrm{frac_{AGN}}$ and redshift estimates for all sources.

The adaptable framework of `AGNBoost` allows straightforward incorporation of additional photometric bands and derived quantities as desired. Training new models is simple with the built-in hyperparameter optimization functions of `AGNBoost`, and can be performed on per-

sonal computers. It is also possible to extend AGNBoost to estimate other parameters of interest such as star formation rates or stellar masses. The computational efficiency and scalability of AGNBoost make it ideally suited for the new era of wide-sky surveys, where rapid identification of AGN candidates will be essential for effective target selection and follow-up observations.

The code for AGNBoost is publicly available on GitHub at https://github.com/hamblin-ku/AGNBoost.

K.H. gratefully acknowledges support from a NASA/FINESST award, 80NSSC22K1594. The MEGA JWST data used in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST) at the Space Telescope Science Institute. The specific observations analyzed can be accessed via DOI: 10.17909/cqes-eh14X.

Software: AstroPy (Astropy Collaboration et al. 2013, 2018, 2022), NumPy (Harris et al. 2020), Matplotlib (Hunter 2007), scikit-learn (Pedregosa et al. 2011), Scipy (Virtanen et al. 2020), TOPCAT (Taylor 2005), STILTS (Taylor 2006), pandas (pandas development team 2020), XGBoostLSS (März 2019), Optuna (Akiba et al. 2019), SGAIN (Neves et al. 2021), SHAP (Lundberg et al. 2020)

APPENDIX

A. PERFORMANCE ON GALAXIES OUTSIDE OF THE CIGALE COLOR SPACE

We test to see if there are any noticeable performance impacts with AGNBoost for sources that are not covered by the CIGALE MIRI color space. Figure A1 looks at the performance of both of AGNBoost's models for sources that lie outside the CIGALE MIRI color space. For the $\mathrm{frac_{AGN}}$ model, there appear to be no trends in performance. However, for the redshift model, estimation performance degrades for high-redshift sources ($z > 2$) that fall outside the CIGALE training distribution in multiple color combinations. All extreme redshift outliers at $z > 3$ are found outside the CIGALE color space at least 7 times.

B. WAVELENGTH DEPENDENT EFFECTS OF MISSING PHOTOMETRIC BANDS

Figure 12 demonstrates that AGNBoost systematically underestimates $\mathrm{frac_{AGN}}$ when MIRI bands are missing and not replaced with statistical imputation. Here, we investigate whether there is a physical reason for this effect (i.e., a wavelength dependence) or if this is entirely due to artifacts of AGNBoost's sparsity aware split finding algorithm, used when features are missing. Figure B1 shows how missing or imputed values in the first 4 NIRCam bands (F115W, F150W, F200W, F277W) effect AGNBoost's $\mathrm{frac_{AGN}}$ estimates. Missing NIRCam bands cause AGNBoost to systematically overestimate $\mathrm{frac_{AGN}}$. This indicates that AGNBoost's $\mathrm{frac_{AGN}}$ estimation with missing values does depend on the wavelengths of the missing values.

REFERENCES

Aird, J., Nandra, K., Laird, E. S., et al. 2010, Monthly Notices of the Royal Astronomical Society, 401, 2531, doi: 10.1111/j.1365-2966.2009.15829.x

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19 (New York, NY, USA: Association for Computing Machinery), 2623–2631, doi: 10.1145/3292500.3330701

Alonso-Herrero, A., Esquej, P., Roche, P. F., et al. 2016, Monthly Notices of the Royal Astronomical Society, 455, 563, doi: 10.1093/mnras/stv2342

Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, The Astrophysical Journal, 772, 26, doi: 10.1088/0004-637X/772/1/26

Assef, R. J., Prieto, J. L., Stern, D., et al. 2018, ApJ, 866, 26, doi: 10.3847/1538-4357/aaddf7

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, doi: 10.1051/0004-6361/201322068

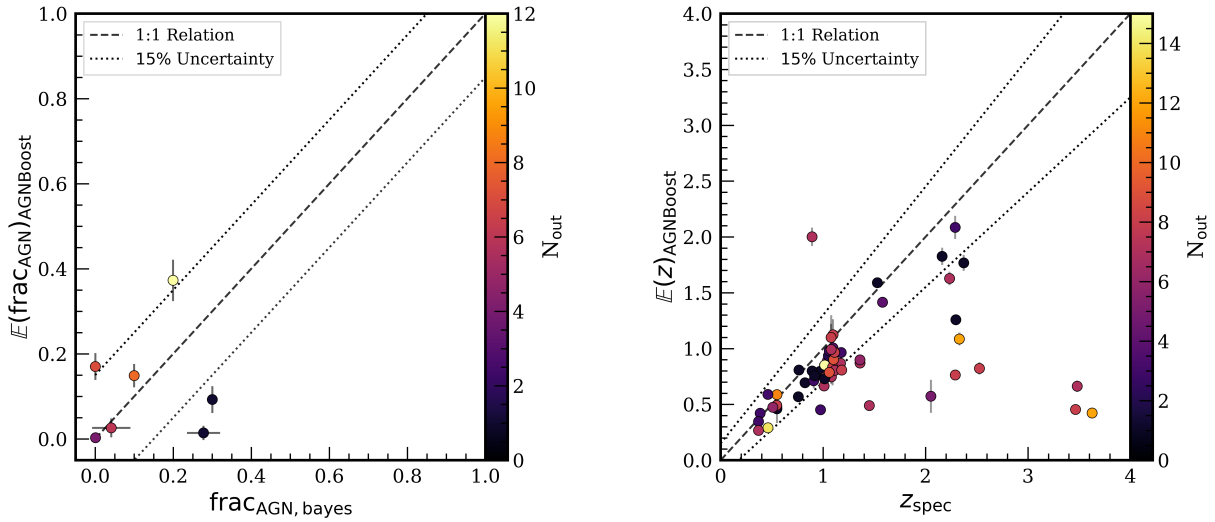

Figure A1. Expectation values of `AGNBoost` predictions for MEGA sources found outside the `CIGALE` MIRI color space (as shown in Figure 2) at least once. Sources are colored by the number of MIRI color combinations for which they fall outside the `CIGALE` distribution. The left panel shows sources from the 246 MEGA galaxies with reliable `CIGALE` fits ($0.5 < \chi^2_{\mathrm{red}} < 2$), and the right panel shows MEGA galaxies with available spectroscopic redshifts. While $\mathrm{frac_{AGN}}$ predictions show no systematic trends related to lying outside the `CIGALE` color space (left), redshift estimation performance degrades for high-redshift sources ($z > 2$).

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: 10.3847/1538-3881/aabc4f

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167, doi: 10.3847/1538-4357/ac7c74

Backhaus, B. E., Kirkpatrick, A., Yang, G., et al. 2025, MEGA Mass Assembly with JWST: The MIRI EGS Galaxy and AGN Survey. https://arxiv.org/abs/2503.19078

Bai, Y., Liu, J., Wang, S., & Yang, F. 2018, AJ, 157, 9, doi: 10.3847/1538-3881/aaf009

Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103, doi: 10.1051/0004-6361/201834156

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503, doi: 10.1086/591786

Bruzual, G., & Charlot, S. 2003, Monthly Notices of the Royal Astronomical Society, 344, 1000, doi: 10.1046/j.1365-8711.2003.06897.x

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, The Astrophysical Journal, 533, 682, doi: 10.1086/308692

Chabrier, G. 2003, PASP, 115, 763, doi: 10.1086/376392

Chao, L., Wen-hui, Z., & Ji-ming, L. 2019, Chinese Astronomy and Astrophysics, 43, 539, doi: 10.1016/j.chinastron.2019.11.005

Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: Association for Computing Machinery), 785–794, doi: 10.1145/2939672.2939785

Conroy, C. 2013, Annual Review of Astronomy and Astrophysics, 51, 393, doi: https://doi.org/10.1146/annurev-astro-082812-141017

Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJL, 660, L1, doi: 10.1086/517931

Dietterich, T. G. 2000, in Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00 (Berlin, Heidelberg: Springer-Verlag), 1–15

Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, The Astrophysical Journal, 748, 142, doi: 10.1088/0004-637X/748/2/142

Draine, B. T., & Li, A. 2001, ApJ, 551, 807, doi: 10.1086/320227

Draine, B. T., Aniano, G., Krause, O., et al. 2014, The Astrophysical Journal, 780, 172, doi: 10.1088/0004-637X/780/2/172

Durodola, E., Pacucci, F., & Hickox, R. C. 2024, arXiv e-prints, arXiv:2406.10329, doi: 10.48550/arXiv.2406.10329

Finkelstein, S. L., Bagley, M. B., Ferguson, H. C., et al. 2023, The Astrophysical Journal Letters, 946, L13, doi: 10.3847/2041-8213/acade4

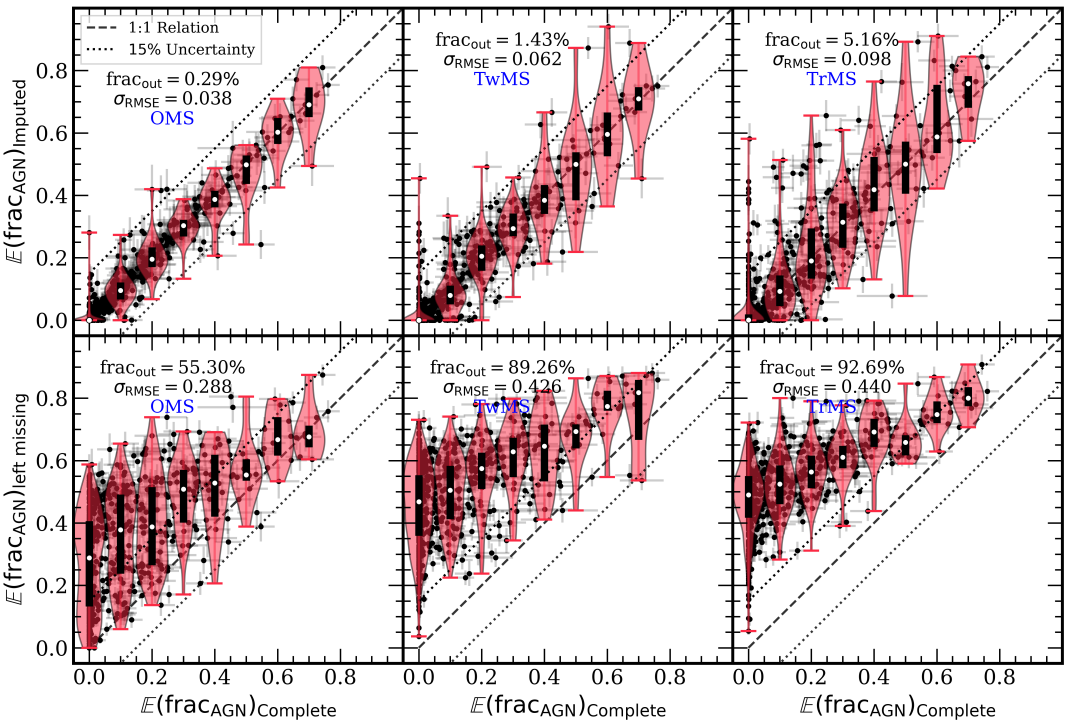

Figure B1. Comparison of `AGNBoost` frac$_{AGN}$ estimation performance with and without `SGAIN` imputation for missing NIRCam photometry (F115W, F150W, F200W, F277W). The top row shows `AGNBoost` results using `SGAIN`-imputed values for randomly removed bands: one missing band (OMS), two missing bands (TwMS), and three missing bands (TrMS). The bottom row shows `AGNBoost` performance when missing bands are left blank (i.e., no imputation). `SGAIN` imputation successfully recovers accurate frac$_{AGN}$ estimates, while running `AGNBoost` with missing NIRCam photometry leads to systematic overestimation of frac$_{AGN}$ values.

Finkelstein, S. L., Bagley, M. B., Haro, P. A., et al. 2025, The Cosmic Evolution Early Release Science Survey (CEERS), arXiv, doi: 10.48550/arXiv.2501.04085

Fontana, A., Salimbeni, S., Grazian, A., et al. 2006, Astronomy and Astrophysics, 459, 745, doi: 10.1051/0004-6361:20065475

Förster Schreiber, N. M., Sauvage, M., Charmandaris, V., et al. 2003, A&A, 399, 833, doi: 10.1051/0004-6361:20021719

Franceschini, A., Toffolatti, L., Mazzei, P., Danese, L., & de Zotti, G. 1991, A&AS, 89, 285

Förster Schreiber, N. M., & Wuyts, S. 2020, Annual Review of Astronomy and Astrophysics, 58, 661, doi: 10.1146/annurev-astro-032620-021910

Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. 2021, CoRR, abs/2104.02395

Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, SSRv, 123, 485, doi: 10.1007/s11214-006-8315-7

Gardner, J. P., Mather, J. C., Abbott, R., et al. 2023, PASP, 135, 068001, doi: 10.1088/1538-3873/acd1b5

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35, doi: 10.1088/0067-0049/197/2/35

Hamblin, K. 2025, AGNBoost CIGALE Mock Dataset, V1, Harvard Dataverse, doi: 10.7910/DVN/YYGZ3P

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357, doi: 10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning, Springer Series in Statistics (New York, NY, USA: Springer New York Inc.)

Heckman, T. M., Kauffmann, G., Brinchmann, J., et al. 2004, The Astrophysical Journal, 613, 109, doi: 10.1086/422872

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

Holwerda, B. W., Wu, J. F., Keel, W. C., et al. 2021, ApJ, 914, 142, doi: 10.3847/1538-4357/abffcc

Hopkins, A. M. 2004, The Astrophysical Journal, 615, 209, doi: 10.1086/424032

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Hüllermeier, E., & Waegeman, W. 2021, Mach Learn, 110, 457, doi: 10.1007/s10994-021-05946-3

Iyer, K., & Gawiser, E. 2017, in Galaxy Evolution Across Time, 2, doi: 10.5281/zenodo.805863

Iyer, K. G., Pacifici, C., Calistro-Rivera, G., & Lovell, C. C. 2025, arXiv e-prints, arXiv:2502.17680, doi: 10.48550/arXiv.2502.17680

Kirkpatrick, A., Pope, A., Sajina, A., et al. 2015, ApJ, 814, 9, doi: 10.1088/0004-637X/814/1/9

Kirkpatrick, A., Pope, A., Alexander, D. M., et al. 2012, The Astrophysical Journal, 759, 139, doi: 10.1088/0004-637X/759/2/139

Kirkpatrick, A., Pope, A., Charmandaris, V., et al. 2013, ApJ, 763, 123, doi: 10.1088/0004-637X/763/2/123

Kirkpatrick, A., Alberts, S., Pope, A., et al. 2017, ApJ, 849, 111, doi: 10.3847/1538-4357/aa911d

Kirkpatrick, A., Yang, G., Le Bail, A., et al. 2023, The Astrophysical Journal, 959, L7, doi: 10.3847/2041-8213/ad0b14

Kocevski, D. D., Finkelstein, S. L., Barro, G., et al. 2024, arXiv e-prints, arXiv:2404.03576, doi: 10.48550/arXiv.2404.03576

Kodra, D., Andrews, B. H., Newman, J. A., et al. 2023, The Astrophysical Journal, 942, 36, doi: 10.3847/1538-4357/ac9f12

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36, doi: 10.1088/0067-0049/197/2/36

Kormendy, J., & Ho, L. C. 2013, Annual Review of Astronomy and Astrophysics, 51, 511, doi: 10.1146/annurev-astro-082708-101811

Krakowski, T., Małek, K., Bilicki, M., et al. 2016, A&A, 596, A39, doi: 10.1051/0004-6361/201629165

Lacy, M., Petric, A. O., Sajina, A., et al. 2007, The Astronomical Journal, 133, 186, doi: 10.1086/509617

Lacy, M., Storrie-Lombardi, L. J., Sajina, A., et al. 2004, The Astrophysical Journal Supplement Series, 154, 166, doi: 10.1086/422816

Laurent, O., Mirabel, I. F., Charmandaris, V., et al. 2000, Astronomy and Astrophysics, 359, 887, doi: 10.48550/arXiv.astro-ph/0005376

Leitherer, C., Li, I. H., Calzetti, D., & Heckman, T. M. 2002, The Astrophysical Journal Supplement Series, 140, 303, doi: 10.1086/342486

Leung, G. C. K., Finkelstein, S. L., Pérez-González, P. G., et al. 2024, arXiv e-prints, arXiv:2411.12005, doi: 10.48550/arXiv.2411.12005

Lundberg, S. M., & Lee, S.-I. 2017, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 4768–4777

Lundberg, S. M., Erion, G., Chen, H., et al. 2020, Nature Machine Intelligence, 2, 2522

Luo, Z., Tang, Z., Chen, Z., et al. 2024, Monthly Notices of the Royal Astronomical Society, 531, 3539, doi: 10.1093/mnras/stae1397

Madau, P., & Dickinson, M. 2014, Annual Review of Astronomy and Astrophysics, 52, 415, doi: 10.1146/annurev-astro-081811-125615

Magdis, G. E., Daddi, E., Béthermin, M., et al. 2012, The Astrophysical Journal, 760, 6, doi: 10.1088/0004-637X/760/1/6

Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024, A&A, 691, A145, doi: 10.1051/0004-6361/202347640

Mechbal, S., Ackermann, M., & Kowalski, M. 2024, A&A, 685, A107, doi: 10.1051/0004-6361/202346557

Messias, H., Afonso, J., Salvato, M., Mobasher, B., & Hopkins, A. M. 2012, The Astrophysical Journal, 754, 120, doi: 10.1088/0004-637X/754/2/120

Morello, G., Morris, P. W., Van Dyk, S. D., Marston, A. P., & Mauerhan, J. C. 2018, Monthly Notices of the Royal Astronomical Society, 473, 2565, doi: 10.1093/mnras/stx2474

März, A. 2019, XGBoostLSS – An extension of XGBoost to probabilistic forecasting. https://arxiv.org/abs/1907.03178

Neves, D. T., Naik, M. G., & Proença, A. 2021, in Computational Science – ICCS 2021, ed. M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Cham: Springer International Publishing), 98–113

Ni, Q., Brandt, W. N., Yang, G., et al. 2021, MNRAS, 500, 4989, doi: 10.1093/mnras/staa3514

Pacifici, C., Iyer, K. G., Mobasher, B., et al. 2023, ApJ, 944, 141, doi: 10.3847/1538-4357/acacff

Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, A&A Rv, 25, 2, doi: 10.1007/s00159-017-0102-9

pandas development team, T. 2020, pandas-dev/pandas: Pandas, latest, Zenodo, doi: 10.5281/zenodo.3509134

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Peeters, E., Spoon, H. W. W., & Tielens, A. G. G. M. 2004, The Astrophysical Journal, 613, 986, doi: 10.1086/423237

Pérez-González, P. G., Barro, G., Rieke, G. H., et al. 2024, ApJ, 968, 4, doi: 10.3847/1538-4357/ad38bb

Pope, A., Chary, R.-R., Alexander, D. M., et al. 2008, The Astrophysical Journal, 675, 1171, doi: 10.1086/527030

Pérez-González, P. G., Rieke, G. H., Villar, V., et al. 2008, The Astrophysical Journal, 675, 234, doi: 10.1086/523690

Rigby, J., Perrin, M., McElwain, M., et al. 2023, Publications of the Astronomical Society of the Pacific, 135, 048001, doi: 10.1088/1538-3873/acb293

Rousseeuw, P. J., Ruts, I., & and, J. W. T. 1999, The American Statistician, 53, 382, doi: 10.1080/00031305.1999.10474494

Sajina, A., Lacy, M., & Scott, D. 2005, The Astrophysical Journal, 621, 256, doi: 10.1086/426536

Sajina, A., Yan, L., Fadda, D., Dasyra, K., & Huynh, M. 2012, The Astrophysical Journal, 757, 13, doi: 10.1088/0004-637X/757/1/13

Scargle, J. D. 1998, ApJ, 504, 405, doi: 10.1086/306064

Shankar, F., Weinberg, D. H., & Miralda-Escudé, J. 2009, The Astrophysical Journal, 690, 20, doi: 10.1088/0004-637X/690/1/20

Stalevski, M., Fritz, J., Baes, M., Nakos, T., & Popović, L. C. 2012, Monthly Notices of the Royal Astronomical Society, 420, 2756, doi: 10.1111/j.1365-2966.2011.19775.x

Stalevski, M., Ricci, C., Ueda, Y., et al. 2016, Monthly Notices of the Royal Astronomical Society, 458, 2288, doi: 10.1093/mnras/stw444

Stasinopoulos, D. M., & Rigby, R. A. 2007, Journal of Statistical Software, 23, 1, doi: 10.18637/jss.v023.i07

Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163, doi: 10.1086/432523

Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Taylor, M. B. 2006, in Astronomical Society of the Pacific Conference Series, Vol. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, 666

Toba, Y., Oyabu, S., Matsuhara, H., et al. 2014, ApJ, 788, 45, doi: 10.1088/0004-637X/788/1/45

Ustimenko, A., Prokhorenkova, L., & Malinin, A. 2020, CoRR, abs/2006.10562

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: 10.1038/s41592-019-0686-2

Wu, X.-B., Hao, G., Jia, Z., Zhang, Y., & Peng, N. 2012, The Astronomical Journal, 144, 49, doi: 10.1088/0004-6256/144/2/49

Yang, G., Boquien, M., Buat, V., et al. 2019, Monthly Notices of the Royal Astronomical Society, 491, 740, doi: 10.1093/mnras/stz3001

Yang, G., Caputi, K. I., Papovich, C., et al. 2023, The Astrophysical Journal Letters, 950, L5, doi: 10.3847/2041-8213/acd639

Zhou, Z.-H. 2012, Ensemble Methods: Foundations and Algorithms, 1st edn. (Chapman & Hall/CRC)