

UC Berkeley

UC Berkeley Previously Published Works

Title

Beyond traditional emission-line diagnostics: Using autoencoders to uncover active galactic nuclei in DESI spectra

Permalink

<https://escholarship.org/uc/item/7z13523r>

Journal

Astronomy & Astrophysics, 713

ISSN

0004-6361

Authors

Alcolea, JA

Siudek, M

Eriksen, M

et al.

Publication Date

2026-09-01

DOI

10.1051/0004-6361/202660335

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Beyond traditional emission-line diagnostics: Using autoencoders to uncover active galactic nuclei in DESI spectra

J. A. Alcolea^{1,*}, M. Siudek^{2,*}, M. Eriksen^{3,4}, M. Mezcua^{1,5,*}, R. Pucha^{6,7}, S. Juneau⁸, S. Gontcho A Gontcho⁹, S. Panda¹⁰, J. Aguilar¹¹, S. Ahlen¹², D. Bianchi^{13,14}, A. Brodzeller¹¹, D. Brooks¹⁵, F. J. Castander^{1,5}, T. Claybaugh¹¹, A. Cuceu¹¹, A. de la Macorra¹⁶, B. Dey^{17,18}, P. Doel¹⁵, S. Ferraro^{11,19}, A. Font-Ribera^{3,20}, J. E. Forero-Romero^{21,22}, E. Gaztañaga^{1,5,23}, G. Gutierrez²⁴, C. Hahn²⁵, H. K. Herrera-Alcántar^{26,27}, R. Joyce⁸, R. Kehoe²⁸, D. Kirkby²⁹, T. Kisner¹¹, A. Kremin¹¹, O. Lahav¹⁵, C. Lamman³⁰, M. Landriau¹¹, L. Le Guillou³¹, A. Meisner⁸, R. Miquel^{3,20}, J. Moustakas³², S. Nadathur²³, W. Percival^{33,34,35}, F. Prada³⁶, I. Pérez-Ràfols³⁷, G. Rossi³⁸, E. Sanchez³⁹, E. Schlafly⁴⁰, D. Schlegel¹¹, M. Schubnell⁴¹, J. Silber¹¹, D. Sprayberry⁸, G. Tarlé⁴¹, B. A. Weaver⁸, and H. Zou⁴²

(Affiliations can be found after the references)

Received 9 April 2026 / Accepted 9 July 2026

ABSTRACT

Context. The growing volume of spectroscopic data in modern surveys motivates the development of data-driven approaches that complement traditional emission-line diagnostics for the identification of active galactic nuclei (AGNs).

Aims. We develop and validate a machine-learning (ML) framework for galaxy and AGN classification that exploits the full optical spectrum, using unsupervised learning to capture spectral features within a semi-supervised classification framework that complements classical approaches.

Methods. We adapted the SPENDER autoencoder architecture to compress galaxy spectra obtained with the Dark Energy Spectroscopic Instrument (DESI) into a low-dimensional latent space that encodes physically meaningful label-agnostic information. We introduced a quantitative classification method via a k-d tree nearest-neighbor search in latent space. The model was trained on 50 222 spectra from the DESI main survey, corresponding to the internal Guadalupe dataset, released as part of data release 1 (DR1), and restricted to $z \leq 0.5$. We validated the performance using labels derived from the FastSpecFit emission line measurements defining seven galaxy classes: AGN, broad-line (BL), composite, star-forming, passive, retired, and other.

Results. The model achieves a high performance in identifying AGN and BL sources, with accuracies of 0.952 and 0.965, respectively. The model reliably recovers AGN and BL sources even in spectra with a low signal-to-noise ratio and identifies AGNs missed by single-diagnostic standard methods. Our classification metrics are benchmarked against traditional diagnostics, and we show that they represent lower limits of the true model performance. The latent space learned by the autoencoder not only separates AGN and non-AGN galaxies, but also correlates with key physical properties such as stellar mass and star formation rate, demonstrating that the method captures astrophysical meaningful features.

Conclusions. This work demonstrates that unsupervised spectral representation learning implemented within a semi-supervised classification framework offers a powerful and scalable pathway toward more complete and reliable AGN catalogs for current and upcoming large spectroscopic surveys.

Key words. techniques: spectroscopic – galaxies: active – galaxies: general – galaxies: nuclei – galaxies: Seyfert – galaxies: statistics

1. Introduction

Active galactic nuclei (AGNs) are among the most powerful and energetic sources in the Universe (Padovani et al. 2017). They are powered by accretion of material onto supermassive black holes (SMBHs; $M_{\text{BH}} \geq 10^6 M_{\odot}$) that are located at the center of galaxies (Salpeter 1964; Zel'dovich 1964). AGNs are found in ~ 1 –10% of all galaxies (Beckmann & Shrader 2012) and emit across the entire electromagnetic spectrum, from radio to gamma-rays. It is crucial to study AGNs for understanding galaxy evolution and black hole growth (Heckman & Best 2014; Harrison & Ramos Almeida 2024). For instance, AGN feedback plays a crucial role in regulating, and possibly altogether halting, star

formation processes within massive galaxies (Ramasawmy et al. 2019). This phenomenon is commonly incorporated into cosmological simulations to ensure that the simulated galaxy properties agree with observations (Rosito et al. 2021; Vogelsberger et al. 2014). Moreover, studying AGNs hosted in dwarf galaxies can provide insights into understanding the formation of seed black holes in the early Universe (Mezcua 2017). The study of AGNs can even constrain cosmological parameters and test whether the dark energy density varies with time (Risaliti & Lusso 2019).

Despite the large variety of AGN classes defined over the years (the so-called AGN zoo), the unified model of AGNs proposes a common physical structure (Antonucci 1993; Urry & Padovani 1995; Beckmann & Shrader 2012), with observational differences arising mainly from the orientation, accretion rate, presence of jets, and properties of the host galaxy (Padovani et al. 2017). A particularly interesting type of AGNs are

* Corresponding authors: joseantalcolea@gmail.com; msiudek@iac.es; mezcua@ice.csic.es

broad-line (BL) AGNs, whose spectra exhibit broad emission lines (e.g., $H\alpha$ and $H\beta$). These lines originate in the broad-line region (BLR), a compact and fast-moving gas region located close to the SMBH (Padovani et al. 2017). The BLR not only enables SMBH mass estimates through virial techniques (Greene & Ho 2005; Moran et al. 2014; Suh et al. 2020), but also offers valuable information on accretion processes, the structure of the inner AGN, and the co-evolution of SMBHs with their host galaxies (Netzer 2015; Shen et al. 2019). It is worth noting that transient broad emission lines can also be produced by stellar processes such as Type II supernovae, luminous blue variables, or Wolf–Rayet stars, particularly in low-mass galaxies, and that follow-up spectroscopy is often required to distinguish between AGN activity and these transient stellar phenomena (Baldassare et al. 2016).

Methods for identifying AGNs exploit their multiwavelength emission. While X-ray, radio, infrared (IR), or gamma-ray observations can effectively reveal AGN activity (Padovani et al. 2017), optical spectroscopy remains the primary tool for AGN selection in large extragalactic surveys and has been shown to provide a more complete AGN census (Juneau et al. 2013). The standard approach to spectroscopic AGN classification relies on emission-line diagnostics such as the Baldwin–Phillips–Terlevich (BPT) diagram (Baldwin et al. 1981), which uses ratios of prominent emission lines (e.g., $[\text{O III}]/H\beta$ vs. $[\text{N II}]/H\alpha$) to distinguish between ionization mechanisms driven by AGNs, star formation, or a combination of the two. However, these methods typically focus on a limited number of emission lines, requiring prior spectral fitting and often missing the broader spectral diversity of galaxies. Moreover, it remains challenging to establish a uniform selection method because different techniques for identifying AGNs often yield distinct and only partially overlapping samples (Juneau et al. 2013; Padovani et al. 2017; Hickox & Alexander 2018).

Large spectroscopic surveys have transformed our view of galaxy populations over the past two decades. Projects such as the Sloan Digital Sky Survey (SDSS; Wu et al. 2023), and more recently, the Dark Energy Spectroscopic Instrument¹ (DESI; DESI Collaboration 2026a), are delivering optical spectra for millions of galaxies, enabling statistical studies of unprecedented scope.

Even with the Early Data Release (EDR; DESI Collaboration 2024) alone, DESI has already led to major scientific advances using classical analysis techniques based on a sample of approximately 1.8 million unique targets. For instance, Siudek et al. (2024) produced a value-added catalog with physical properties for more than 1.3 million galaxies using CIGALE spectral energy distribution fitting (Boquien et al. 2019), and explored its use as a proxy for AGN identification (Siudek et al. 2025). Similarly, Pucha et al. (2025) tripled the census of dwarf AGN candidates through emission-line fitting, while Guo et al. (2024) increased the known population of changing-look AGNs by approximately 30%. Additional studies revealed new connections between dust reddening and radio emission in quasars (Fawcett et al. 2023) and provided improved single-epoch black hole mass estimates for over 55 000 DESI quasars (Pan et al. 2025). These results highlight the scientific power of DESI even when exploited with traditional model-based approaches.

The DESI data release 1 (DR1; DESI Collaboration 2026a) has already dramatically increased the volume and diversity of available spectra, including more than 18.7 million unique targets. Forthcoming releases, such as data release 2 (DR2;

Abdul Karim et al. 2025), are expected to expand the dataset to more than 40 million sources in the coming years. This growth will be further amplified by future massive spectroscopic surveys such as the 4-metre Multi-Object Spectroscopic Telescope (4MOST; De Jong et al. 2019), the Prime Focus Spectrograph (PFS; Tamura et al. 2016), and the William Herschel Telescope Enhanced Area Velocity Explorer (WEAVE; Jin et al. 2024), which will collectively provide spectroscopic observations for millions of additional sources.

It is inherently challenging to extract physical information from galaxy spectra at this scale. Classical model-fitting and line-based analysis techniques are computationally intensive and can become major bottlenecks when applied to datasets of this size and complexity (Pattnaik et al. 2025). These limitations naturally motivate the adoption of machine-learning (ML) approaches for spectroscopic analysis in astronomy (Angeloudi et al. 2024; Siudek et al. 2025), which offer the scalability and flexibility required to exploit the full information content of galaxy spectra beyond a limited set of emission-line measurements (Huertas-Company & Lanusse 2023).

Briefly, ML algorithms fall into two main categories: supervised methods, which require labeled data, and unsupervised methods, which do not. There are categories in between, such as semi-supervised learning, which combines elements of supervised and unsupervised learning, typically by learning representations from unlabeled data and subsequently using labeled information to guide classification. Several recent studies have demonstrated the potential of these methods (e.g., Siudek et al. 2018b,a; Hviding et al. 2024, Pérez-Díaz et al. 2024), highlighting in particular the power of unsupervised approaches, which allow for the analysis of the full spectrum without prior assumptions or labels. The application of unsupervised ML to galaxy spectra offers several advantages: (i) it enables the exploitation of multiple AGN signatures simultaneously across the full spectral range, (ii) it uncovers natural groupings in a high-dimensional data-driven space, and (iii) it avoids biases introduced by relying on predefined labels or selection criteria (Siudek et al. 2018b). By using the entire spectral information without imposing strict thresholds, unsupervised methods not only enhance AGN identification, but also open the door to discovering diverse AGN types based on their intrinsic spectral features.

Several recent studies have demonstrated the potential of ML methods applied to galaxy spectra. For instance, Siudek et al. (2018b,a) applied a Fisher expectation-maximization algorithm to the VIMOS Public Extragalactic Redshift Survey, categorizing over 50 000 VIPERS galaxies into 12 classes based on shared physical and spectral properties. This approach led to AGN identification that exceeded the mid-IR classification (Siudek et al. 2022a) and allowed them to discover rare sources that were missed by standard methods (Lisiecki et al. 2023; Siudek et al. 2022b; Mezcua et al. 2023).

Focusing on dimensionality reduction, Portillo et al. (2020) showed that an autoencoder (a neural network that compresses data into a low-dimensional representation (encoding) and then reconstructs it (decoding)) can condense SDSS spectra into just six latent parameters while retaining enough information to accurately reconstruct the original spectra. This nonlinear approach outperforms traditional methods such as a principal component analysis. Teimoorinia et al. (2022) expanded on this by adding convolutional layers to extract correlated features from the MaNGA survey galaxy spectra, while Pat et al. (2022) employed a probabilistic autoencoder (Böhm et al. 2023) on SDSS DR16 to improve spectral reconstruction and

¹ <https://www.desi.lbl.gov/>

detect outliers in the latent space. More recently, Nicolaou et al. (2026) applied a variational autoencoder to approximately 200 000 DESI spectra, demonstrating that it can compress spectral dimensionality by a factor of 100 while retaining enough information to accurately reconstruct features, and successfully identifying anomalous spectra with artifacts and with unusual physical characteristics.

SPENDER (Melchior et al. 2023; Liang et al. 2023b) is the basis for our work. This autoencoder is designed to analyze and reconstruct galaxy spectra. By encoding observed-frame spectra and applying redshift transformations, it produces a latent representation that enables consistent comparison of spectral features across galaxies at different redshifts. This approach reduces redshift-driven biases and is robust to common artifacts, allowing spectra from large surveys to be ingested directly without extensive preprocessing.

SPENDER has been successfully applied to a variety of spectroscopic analyses, including accurate reconstructions and noise suppression in SDSS-II data (Melchior et al. 2023), outlier detection (Liang et al. 2023b), and the identification of anomalies in the DESI Bright Galaxy Survey, such as a redshift bias in the Redrock² algorithm (Bailey et al., in prep.) caused by M-dwarf contamination (Liang et al. 2023a). It has also been adapted to reconstruct intrinsic quasar (QSO) spectra and to measure the Ly α forest (Hahn et al. 2025). By encoding spectra into a few latent variables, this method facilitates data exploration and interpretation, as demonstrated in studies linking optical and IR emission (Jespersen et al. 2026). These applications illustrate the growing impact of ML in spectroscopic research, enabling analyses that are difficult to achieve with traditional techniques.

In this work, we take a step toward automated AGN identification and galaxy classification by adapting the SPENDER architecture within a semi-supervised workflow. The autoencoder learns a low-dimensional spectral representation without using labels, while class assignments are derived from reference classifications based on emission-line diagnostics and propagated according to the proximity of similar spectra in the learned latent space. We apply this method to a representative sample of 50 241 galaxies from DESI DR1 (DESI Collaboration 2026a), demonstrating its ability to recover AGN candidates and highlighting its potential for scalable AGN searches in upcoming large spectroscopic datasets.

This paper is organized as follows. Section 2 describes the DESI data, sample selection, and labeling strategy. Section 3 presents the SPENDER-based method and the k-d tree classification framework. The results are presented in Section 4, including AGN identification and analysis of discrepant classifications. We interpret these results in Section 5 and summarize our main conclusions in Section 6.

2. Data

2.1. The DESI Survey

The DESI spectrograph, which is robotically fiber-positioned and reconfigurable in 3 minutes, is situated at the Mayall 4-meter telescope at the Kitt Peak National Observatory (DESI Collaboration 2022; Poppett et al. 2024). The advanced DESI optics and 5000 robotic positioners enable swift, high-resolution spectral measurements of galaxies across 14 000 square degrees, with coverage extending from 360 nm to 980 nm. This allows us to explore redshifts up to $z \sim 1.7$ for emission-line galaxies, while QSOs are observed to significantly higher redshifts,

with Lyman- α forest science enabled up to $z \sim 3.5$ (DESI Collaboration 2016; Silber et al. 2023; Miller et al. 2024).

2.2. Guadalupe dataset

Data from the first two months of the DESI main survey (referred to internally as Guadalupe) were employed in this work. This dataset comprises 43 of the 261 observation nights of the main survey (Schlafly et al. 2024) and is part of DR1 (DESI Collaboration 2026a). Guadalupe was designed to follow the standard DESI survey strategy and target selection and is therefore considered broadly representative of the main galaxy sample (DESI Collaboration 2026a). The dataset contains more than three million galaxies, 1 178 007 of which are galaxies and QSOs in our redshift range of interest: $z \in [0.001, 0.5]$. The lower threshold avoids contamination from sources within the Milky Way (Pucha et al. 2025), while the upper limit ensured that key emission lines such as H α remained within the DESI spectral range. Future work will extend the analysis to higher redshifts. Each spectrum consists of 7781 spectral measurements spanning the observed wavelength range $\lambda_{\text{obs}} = [3600, 9824]$ Å.

2.3. Preprocessing and quality cuts

The DESI spectra are automatically processed through the DESI spectroscopic pipeline (Guy et al. 2023), followed by spectral classification and redshift determination using the Redrock fitting pipeline (Anand et al. 2024). Redrock assigns a spectral type (GALAXY, QSO, or STAR), redshift (z), and a warning bitmask (ZWARN), among other parameters. We only retained entries with a galaxy or QSO spectral type, with ZWARN = 0 and no fiber issues (COADD_FIBERSTATUS = 0), following DESI recommendations (DESI Collaboration 2024, 2026a).

2.4. Spectral classification and labeling

Since SPENDER does not produce direct class predictions, external labels are required to assign spectral classes and evaluate the classification performance. We used reference labels derived from emission-line measurements obtained with FastSpecFit v3.1 (Moustakas 2023), a spectrophotometric fitting algorithm that provides emission-line measurements and broadband photometry for DESI spectra. FastSpecFit decomposes permitted Balmer lines into narrow and broad components only when the inclusion of a broad component leads to a statistically significant improvement of the fit. Based on these measurements, galaxy classes are assigned according to the following criteria, summarized in Table 1:

- BL sources³ are labeled when the broad component of the H α emission is significantly detected, with an amplitude-over-noise ratio (AON) ≥ 3 and a signal-to-noise ratio (S/N) ≥ 5 , while the narrow component of the H α line has S/N ≥ 3 .
- The remaining galaxies, those with S/N ≥ 3 in H α , H β , [OIII] λ 5007, and [NII] λ 6584, are classified using the BPT diagram with the criteria from Kewley (Kewley et al. 2001) and Kauffmann (Kauffmann et al. 2003) (hereafter the [NII]–BPT diagram) into narrow-line (NL) AGNs, star-forming, or composite.
- Galaxies that fail to meet these thresholds are further categorized by the H α equivalent width (EW) into passive ($0 < \text{EW}_{\text{H}\alpha} < 0.5$ Å) or retired ($0.5 \leq \text{EW}_{\text{H}\alpha} \leq 3$ Å).

³ BL emission is characteristic of AGNs, so these sources are expected to be AGN-dominated, although their BPT classification was not examined.

² <https://redrock.readthedocs.io>

Table 1. Galaxy classes, subpopulations, and selection criteria for the representative Guadalupe dataset.

Galaxy class	N (%)	Galaxy subpopulation	N (%)	Criterion
AGN	12 217 (24.3%)	BL	5592 (11.1%)	$S/N_{H\alpha,n} \geq 3$, $S/N_{H\alpha,b} \geq 5$, $AON_{H\alpha,b} \geq 3$ [N II]-BPT diagram & $S/N_{lines} \geq 3$
		NL	6625 (13.2%)	
Star-forming	14 791 (29.5%)	Composite	7271 (14.5%)	[N II]-BPT diagram & $S/N_{lines} \geq 3$ [N II]-BPT diagram & $S/N_{lines} \geq 3$
		Star-forming	7520 (15.0%)	
Other/passive	23 214 (46.2%)	Passive	7473 (14.9%)	$0 < EW_{H\alpha} < 0.5 \text{ \AA}$
		Retired	7908 (15.7%)	$0.5 \leq EW_{H\alpha} \leq 3 \text{ \AA}$
		Other	7833 (15.6%)	None of the above criteria are met
Total	50 222			

Notes. Counts correspond to the representative Guadalupe dataset after masking. Here, “n” refers to the narrow component and “b” to the broad component of $H\alpha$.

- Galaxies that do not satisfy any of the above criteria are classified as other. This heterogeneous category includes sources whose nature cannot be robustly determined based on the available emission-line measurements, such as peculiar galaxies, spectral outliers, or objects with ambiguous or low-S/N features. In many cases, assignment to the other class arises as a consequence of the S/N thresholds applied to emission-line measurements when defining the remaining galaxy classes.

In addition to these seven galaxy subclasses, we categorized the dataset into three main galaxy classes: AGNs (including NL AGNs and BL), star-forming (including star-forming and composite), and other/passive, the latter encompassing galaxies classified as retired, passive, and other. The performance of the proposed method is evaluated using both the three-class and seven-class classification schemes.

2.5. Sample selection

The sample of 50 241 galaxies we used was constructed by retaining all NL and BL AGNs, which are much less numerous than the other classes. For the remaining classes, galaxies were randomly selected in bins of absolute magnitude and redshift, following the procedure of Siudek et al. (2024). This selection ensured a balanced class representation and uniform coverage across the magnitude–redshift plane.

2.6. Flux normalization and masking

To normalize the flux, we computed the median flux over a flat region in the rest-frame range $\lambda_{rest} = [5300, 5850] \text{ \AA}$, and the spectrum was divided by this value, centering it at unity. This suppressed the overall amplitude as a dominant feature and mitigated redshift-dependent scaling (Melchior et al. 2023). Sources containing infinity or NaN values in the normalized flux, as well as those with a negative normalization factor, were excluded. After these cuts, the final dataset contained 50 222 sources, as summarized in Table 1.

2.7. Value-added catalogs

Additionally, the value-added catalog⁴ from Siudek et al. (2024) was used to obtain additional variables, including the star formation rate (SFR) and stellar mass. These parameters were derived

by applying the CIGALE spectral energy distribution fitting code (Boquien et al. 2019) to the multiwavelength photometry of DESI galaxies (Dey et al. 2019).

To further assess the reliability of the AGN predictions by our model, the AGN/galaxy classification value-added catalog (hereafter AGN/galaxy VAC) from DESI DR1⁵ was used as well. This comprehensive catalog incorporates AGN diagnostics based on a variety of optical, ultraviolet, and IR criteria, using emission-line measurements from FastSpecFit and mid-IR photometry from WISE. A detailed description of the catalog will be provided in Juneau et al. (in preparation).

3. Methods

3.1. Model

The SPENDER architecture (Melchior et al. 2023; Liang et al. 2023a,b) was adapted to compress observed galaxy spectra into a low-dimensional latent representation.

The encoder operates on the observed-frame spectra and produces an initial representation. Then, a multilayer perceptron (Goodfellow et al. 2016) explicitly incorporates the redshift to transform this representation into a latent space in the rest-frame. Finally, the decoder reconstructs the spectra in the rest frame. However, the flexibility of the autoencoder introduces degeneracy in the latent space, allowing similar spectra to be represented in multiple ways. Consequently, while the latent space itself does not need to explicitly encode redshift information because reconstruction occurs in the rest frame, spectra at different redshifts can still be mapped to distinct latent representations (Liang et al. 2023b). To address this, a total loss function is employed, combining three components: fidelity loss, which quantifies reconstruction accuracy using inverse-variance weighted mean squared error; similarity loss, which ensures proximity of similar spectra in latent space; and consistency loss, which enforces closeness between original and augmented spectra. These losses are defined following Liang et al. (2023b).

Following Melchior et al. (2023), we adopted a latent space dimensionality of ten. It was shown that increasing the latent space from six to ten dimensions leads to small but consistent improvements in the reconstruction performance, as quantified by the total loss function, without introducing signs of overfitting. This choice therefore provided a good compromise

⁴ <https://data.desi.lbl.gov/doc/releases/edr/vac/cigale/>

⁵ <https://data.desi.lbl.gov/doc/releases/dr1/vac/agnqso/>

between compact representation and accurate spectral reconstruction.

For every spectrum, a mask was implemented on top of the 100 strongest telluric lines, where zero weights were assigned in the 5 Å region around the line centers (Melchior et al. 2023). When a mask was implemented, the spectra remained unchanged; only the weights were set to zero. This ensured that these regions did not contribute to the final loss during training.

3.2. Training

Several preliminary training runs were performed to optimize the training pipeline, focusing on learning-rate adjustments. The final training scheme ensured loss function convergence over 60 epochs by employing a cyclic learning-rate scheduler, specifically, PyTorch’s OneCycleLR (PyTorch 2025). In this setup, the learning rate follows a cyclical pattern within each epoch, with a peak value that is manually decreased by an order of magnitude every 20 epochs, from 10^{-3} to 10^{-5} . To mitigate stochastic effects, the training procedure was initially repeated in five independent runs, but we observed negligible differences in the loss function and evaluation metrics between runs. Consequently, only a single representative training run is reported in this work.

3.3. UMAP

For visualization purposes, we implemented the uniform manifold approximation and projection (UMAP) method (McInnes et al. 2020). This dimensionality-reduction technique takes the ten-dimensional latent space as input and projects it into a two-dimensional space, where the two parameters are subsequently plotted. The different galaxy classes are then color-coded a posteriori to examine their distribution and clustering in the latent space.

The umap package (McInnes et al. 2018) was used for this implementation with Python 3.10, adopting the default hyperparameters ($n_neighbors=15$, $min_dist=0.1$, and $metric=euclidean$). Since UMAP projections are non-unique and depend on the adopted hyperparameters and stochastic initialization, we assessed the robustness of the resulting visual structures by repeating the projection with different random seeds and alternative parameter configurations. Although the exact geometry of the two-dimensional embedding varied between realizations, as expected from the stochastic nature of the algorithm, the main structures and class separations discussed in this work remained stable.

Plotting the two-dimensional UMAP projection of the latent space enabled us to straightforwardly visually inspect the data while approximately preserving the topological structure. However, information is inevitably lost during the dimensionality reduction process, and distances or apparent separations in the projected space should not be interpreted quantitatively. Therefore, all quantitative analyses and classification metrics presented in this work were computed directly within the original ten-dimensional latent space rather than from the two-dimensional UMAP projection, which was only used for visualization purposes.

3.4. Classification: k-d tree

The SPENDER model does not produce direct class predictions, and class assignments must therefore be derived from the learned spectral representation. To this end, we implemented a k-d tree model (Bentley 1975) within the trained ten-dimensional latent

space. Galaxy classes are predicted via a majority vote among the ten nearest neighbors, exploiting the proximity of spectra with similar spectral features in the learned representation. The neighbor labels were derived from FastSpecFit emission-line measurements (Section 2) and were used as reference classifications for the class assignment and metric evaluation. This nearest-neighbor approach accommodates complex data distributions without imposing predefined decision boundaries. This step completes the semi-supervised workflow: the spectral representation is learned without labels, while class assignments are derived from reference classifications propagated through proximity in latent space.

To determine the hyperparameters of the nearest-neighbor classifier using the k-d tree algorithm, we performed iterative runs exploring several distance metrics, including Minkowski, Manhattan, and Chebyshev distances, and leaf sizes between 10 and 60 in steps of 10. The model performance was highly stable across the explored hyperparameter space. Variations in leaf size produced no measurable impact on the classification metrics, while changing the distance metric led to variations below $\sim 0.4\%$ in global accuracy and macro F1 score (see Section 3.5 and Appendix A for the metric definitions). The Minkowski metric yielded the best overall AGN performance, although the improvement relative to other metrics was marginal ($<0.15\%$ in AGN accuracy). We therefore adopted the default scikit-learn parameters (Pedregosa et al. 2011), using a leaf size of 40 and the Minkowski metric with $p = 2$, implemented in Python 3.10.

3.5. Performance metrics

The class predictions from the k-d tree, along with the reference labels derived from FastSpecFit line measurements, enabled us to evaluate the model performance through various metrics. In multiclass classification, several strategies exist for computing the evaluation metrics. The one-versus-one strategy considers all pairwise combinations of classes, while one-versus-rest binarizes the problem by comparing each class against all others (Pedregosa et al. 2011). In our analysis, we adopted the latter strategy.

The definitions of all performance metrics we used are provided in Appendix A. All metrics were computed using the scikit-learn package (Pedregosa et al. 2011), with Python 3.10. The `multi_class` parameter was set to one-versus-rest.

4. Results

4.1. Identification of AGNs

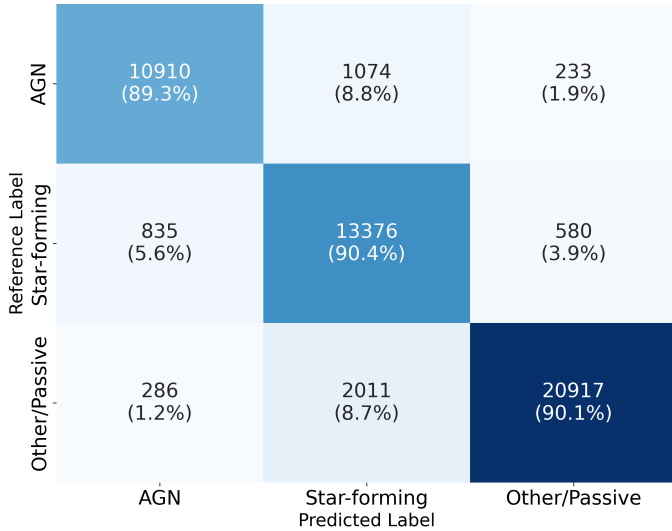
To assess the performance of the model in identifying AGNs, we first used the three main galaxy classes: AGNs (including NL AGNs and BL), star-forming (including star-forming and composite), and other/passive, the latter encompassing galaxies classified as other, retired, and passive. Table 2 summarizes the classification performance of the SPENDER model based on these categories.

The model achieved the highest accuracy and area under the curve (AUC) for the AGN classification (0.952 and 0.976, respectively), as also reflected in the receiver-operating characteristic (ROC) curves in Fig. B.1 panel (a), where the AGN curve lies slightly above those of the other classes, indicating a marginally better trade-off between sensitivity and specificity across thresholds. The performance for star-forming and other/passive galaxies is slightly lower but remains robust, with

Table 2. Classification performance for the three galaxy classes.

Galaxy class	Accuracy	Precision	Recall	FPR	F1-score	AUC
AGN (BL + NL AGN)	0.952	0.907	0.893	0.030	0.900	0.976
Star-forming (star-forming + composite)	0.910	0.813	0.904	0.087	0.856	0.956
Other/passive (passive + retired + other)	0.938	0.963	0.901	0.030	0.930	0.971

Notes. The metric definitions are provided in Appendix A.

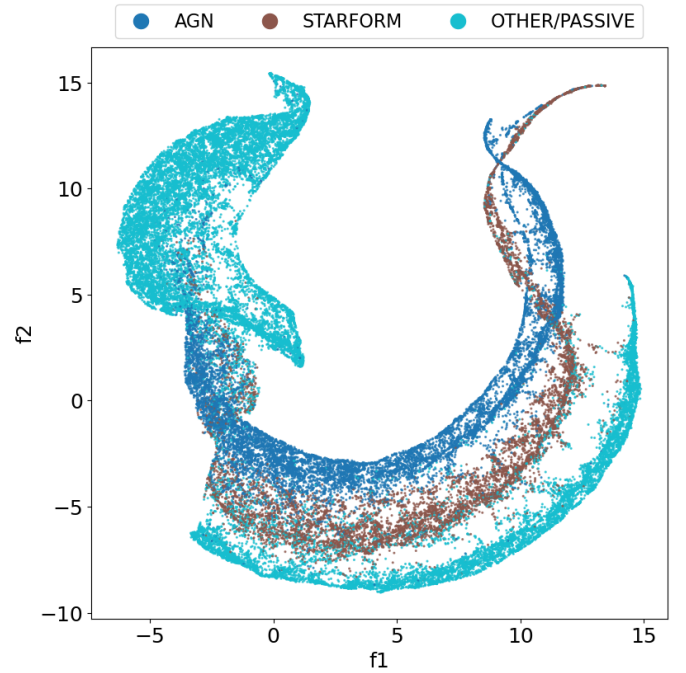
**Fig. 1.** Confusion matrix for the three-class galaxy classification. The percentages are computed with respect to the reference labels.

accuracy values above 90% and AUC scores exceeding 0.97. The star-forming class shows the highest recall (0.904), suggesting the model is particularly effective at recovering most of these sources. However, the recall values are very similar across the three classes, indicating a balanced sensitivity. The other/passive class attains the highest precision (0.963) and a strong F1-score (0.930), reflecting a high degree of purity and consistency in the predicted labels. Overall, the results highlight that the model can classify galaxies reliably in all categories.

Figure 1 presents the confusion matrix for the three-class classification. The model performs well in the AGN identification; only 835 and 286 sources predicted as AGNs are labeled as star-forming and other/passive, respectively. These cases account for 9.3% of all AGN predictions, while 10 910 (90.7%) sources are consistently identified as AGNs.

It is important to note that the metrics and confusion matrix values we presented rely on classifications derived from the emission-line measurements provided by FastSpecFit. These classifications were used here as reference labels to evaluate the model. However, as discussed in Section 4.4 and Appendix C, some sources exhibit weak emission lines (e.g., H α , H β , [OIII] λ 5007, or [NII] λ 6584) with an S/N below 3, leading to their classification as other even though they might be genuine AGNs or star-forming galaxies. Moreover, as discussed in Section 4.4, some of these sources might still be AGNs according to alternative diagnostics beyond the [NII]–BPT diagram. Therefore, the reported accuracy values should be treated as lower bounds on the true model performance.

To further explore how the SPENDER model encodes spectral information, we projected the ten-dimensional latent space

**Fig. 2.** UMAP projection of the ten-dimensional latent space into two dimensions. Galaxies are color-coded a posteriori according to their FastSpecFit classification: AGNs (blue), star-forming (brown), and other/passive (cyan).

onto a two-dimensional plane using UMAP (McInnes et al. 2020) (Fig. 2). The three galaxy classes, which were not used in the training, were color-coded a posteriori. The clustering of these categories suggests that the latent space retains meaningful physical information by grouping galaxies with similar spectral properties together.

The UMAP embedding exhibits a curved, bifurcated morphology, with three distinct branches that become narrower toward the left end. The AGN population (blue) follows a well-defined arc that remains relatively separated from the other/passive class (cyan). The other/passive galaxies are distributed across two regions: a cluster in the upper left corner, and a more compact branch in the lower right corner. The star-forming galaxies (brown) are interspersed between the AGN and other/passive populations, forming a diffuse region on the left side with additional mixing with the AGN class and a sharper, more compact branch on the right.

4.2. Identification of galaxy subpopulations

In addition to distinguishing between AGN, star-forming, and other/passive galaxies, this section evaluates the model performance in classifying an extended set of seven galaxy classes:

Table 3. Performance metrics for classification of the extended galaxy population into seven classes.

Galaxy class	Accuracy	Precision	Recall	FPR	F1-score	AUC
BL	0.965	0.839	0.845	0.020	0.842	0.973
NL AGN	0.937	0.768	0.746	0.034	0.757	0.946
Composite	0.875	0.561	0.630	0.083	0.594	0.897
Star-forming	0.904	0.669	0.716	0.062	0.692	0.937
Passive	0.848	0.491	0.597	0.108	0.539	0.890
Retired	0.833	0.468	0.468	0.099	0.468	0.875
Other	0.847	0.516	0.327	0.057	0.400	0.787

Notes. The metric definitions are provided in Appendix A.

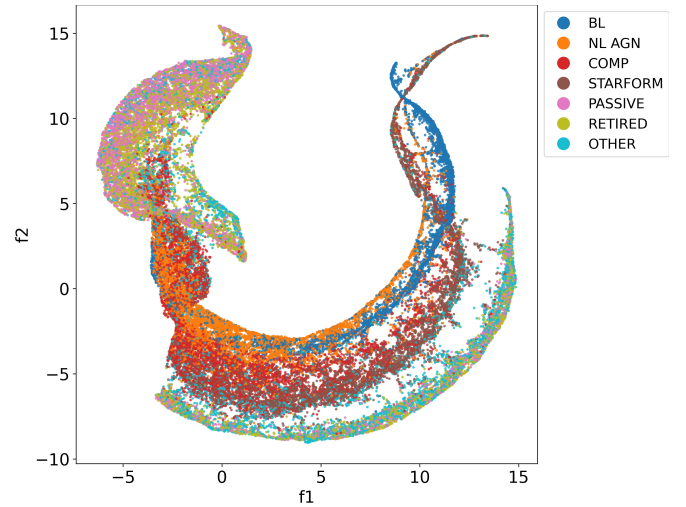
Reference Label	BL	4725 (84.5%)	643 (11.5%)	172 (3.1%)	30 (0.5%)	0 (0.0%)	1 (0.0%)	21 (0.4%)
	NL AGN	585 (8.8%)	4940 (74.6%)	712 (10.7%)	178 (2.7%)	12 (0.2%)	91 (1.4%)	107 (1.6%)
	COMP	202 (2.8%)	602 (8.3%)	4583 (63.0%)	1454 (20.0%)	23 (0.3%)	99 (1.4%)	308 (4.2%)
	STARFORM	19 (0.3%)	46 (0.6%)	1756 (23.4%)	5385 (71.6%)	12 (0.2%)	30 (0.4%)	272 (3.6%)
	PASSIVE	3 (0.0%)	9 (0.1%)	9 (0.1%)	18 (0.2%)	4459 (59.7%)	2293 (30.7%)	682 (9.1%)
	RETIRED	10 (0.1%)	20 (0.3%)	39 (0.5%)	14 (0.2%)	3112 (39.4%)	3697 (46.8%)	1016 (12.8%)
	OTHER	87 (1.1%)	175 (2.2%)	896 (11.4%)	970 (12.4%)	1456 (18.6%)	1686 (21.5%)	2563 (32.7%)
		BL	NL AGN	COMP	STARFORM	PASSIVE	RETIRED	OTHER
Predicted Label								

Fig. 3. Confusion matrix for the extended galaxy classification. The percentages are computed with respect to the reference labels.

NL AGNs, BL, star-forming, composite, passive, retired, and other.

Table 3 summarizes the classification performance for this extended galaxy taxonomy. The model achieves its best performance for BL, with the highest accuracy (0.965) and AUC (0.973), also evident in Fig. B.1 panel (b), where the BL ROC curve remains well above the curves of the other classes across most thresholds. NL AGNs are also well identified, with an accuracy of 0.937 and an AUC of 0.946. The performance for star-forming and composite galaxies is solid, with accuracy values of 0.904 and 0.875, and AUCs of 0.936 and 0.897, respectively. It is more challenging to identify passive and retired galaxies. The accuracy (0.848 and 0.833) and AUC scores are lower (0.890 and 0.875). Galaxies classified as other show the lowest recall (0.327) and F1-score (0.400), indicating that this heterogeneous group is the most difficult to distinguish reliably. This difficulty is also reflected in Fig. B.1 panel (b), where its ROC curve lies closest to the diagonal.

The confusion matrix (Figure 3) shows that disagreements in the NL AGN class predominantly involve BL and composite galaxies. In the UMAP embedding of the ten latent features (Figure 4), colored by these extended galaxy classes, NL AGNs, BL sources, and a mixture of star-forming and composite galaxies occupy distinct regions of the latent space. Notably, the BL sources and NL AGN regions are well separated, reinforcing the claim that the model can distinguish these two classes. However, the relation between star-forming and composite galaxies

**Fig. 4.** UMAP embedding of the ten latent features of the SPENDERmodel. The points are colored a posteriori according to seven galaxy classes: NL AGNs, BL, composite, star-forming, passive, retired, and other.

is more complex, with significant overlap between them, especially on the left side of the latent space, where the branches are closer together. This reflects the inherent ambiguity in their spectral features, as recent studies using dimensionality reduction or latent space clustering showed that the distinction between star formation-dominated and composite sources is gradual rather than sharp (Pat et al. 2022; Tous et al. 2025).

The classification challenges are most evident for passive and retired galaxies, which frequently misclassify into one another, as observed in the confusion matrix (Figure 3). Figure 4 shows that these two classes tend to cluster together in the latent space. Retired, passive, and other galaxies form a distinct structure in the UMAP embedding, appearing to split into two separate regions: a bulge in the upper left corner, and a second cluster located at lower f2 UMAP values. As discussed in Section 4.3, this separation reflects underlying physical differences between these galaxy populations and is not solely a projection artifact.

4.3. Latent distribution of physical properties

The UMAP embedding of the latent space shows clear visual trends when color-coded by galaxy properties from the value-added catalog presented in Siudek et al. (2024) (Figure 5). Although the SPENDER architecture incorporates additional loss terms to discourage the encoding of redshift information, redshift trends can still appear in the latent representation. As discussed in Liang et al. (2023b), this effect primarily arises

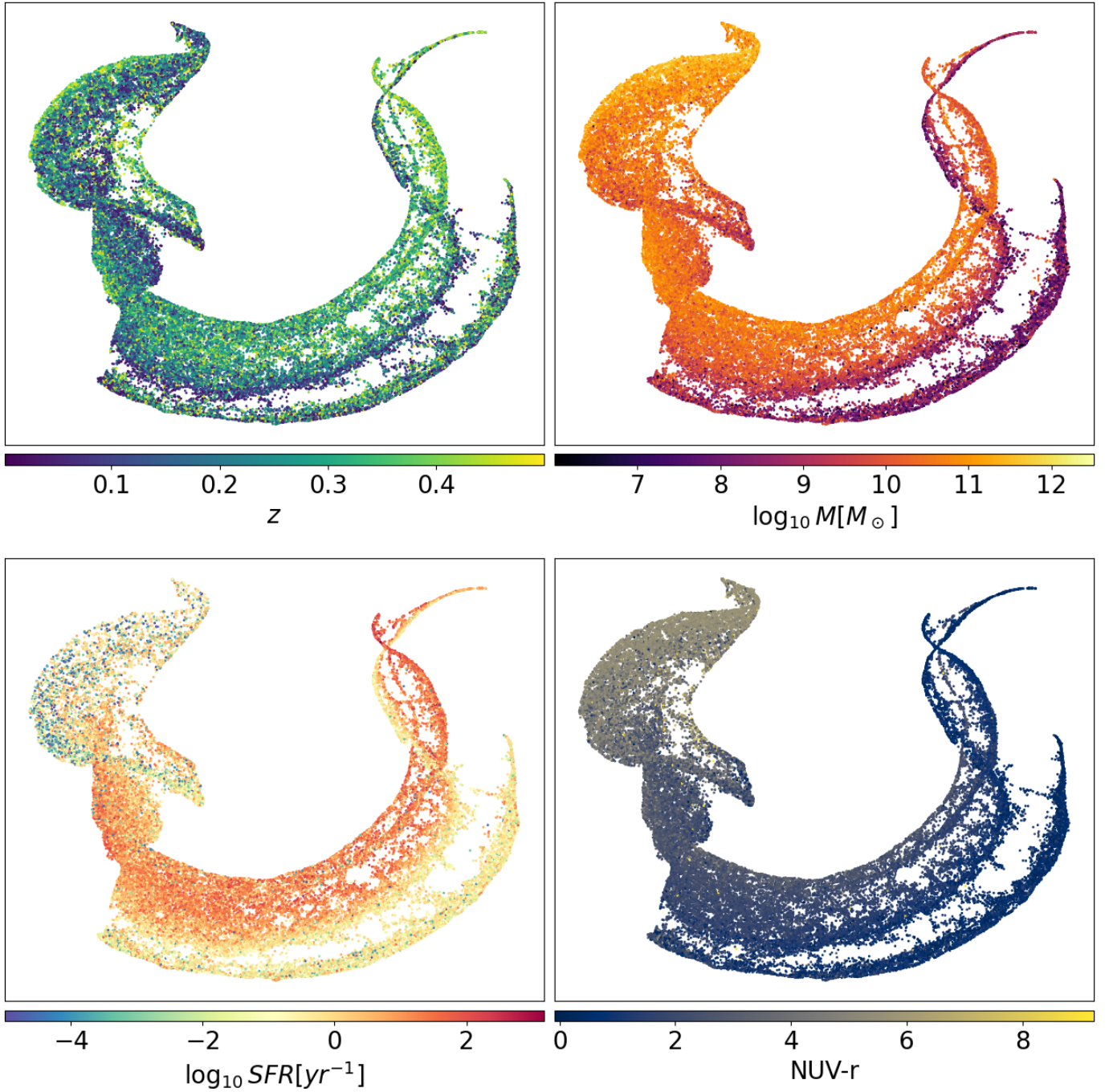


Fig. 5. UMAP embedding of the ten-dimensional latent space, color-coded by redshifts (top left), stellar masses (top right), specific SFRs (lower left), and NUV- r color (lower right).

from selection effects in magnitude-limited samples and not from an explicit encoding of redshift. In particular, high-redshift galaxies are preferentially massive, while low-mass galaxies are only observed at low redshift, which naturally induces correlated trends between redshift and stellar mass in the latent space. The NL AGN and BL branches show distinct redshift gradients, particularly in BL, where the far right segment is dominated by $z \in [0.4, 0.5]$. By contrast, stellar masses show comparatively weak variation along both branches. Star-forming galaxies display a vertical gradient in redshift and mass, with the most massive objects preferentially located toward the left end of the sequence.

Objects with ongoing or recent star formation within the last billion years typically have near-UV (NUV) $-r$ colors < 5.5 , as reported in Johansson et al. (2016); Chilingarian & Zolotukhin (2011). The observed NUV- r gradient therefore suggests that galaxies in the top right cluster are consistent with little or no star formation within this period, whereas those in the bottom right branch show signatures of more recent activity. This interpretation is supported by the strong correlation between NUV- r color and SFR shown in the lower left panel of Figure 5.

The structures dominated by passive, retired, and other galaxies (see Figure 4) show that the upper left cluster tends to contain more massive systems with lower SFRs than the

second cluster located at lower UMAP f2 values. This suggests that the observed separation reflects underlying physical differences between these populations, and that stellar population properties may capture distinctions not fully described by the current $EW_{H\alpha}$ -based separation between passive and retired galaxies. In particular, the passive population appears to be substantially more complex and heterogeneous than implied by the standard spectroscopic classification. This interpretation is consistent with previous studies highlighting the diversity of passive systems (Siudek et al. 2018b,a; Portillo et al. 2020; Siudek et al. 2022a).

4.4. Sources with discrepant classifications

In this section, we analyze sources for which the model prediction and the reference classification disagree, with a particular focus on the identification of AGNs. This analysis combines stacked spectra of off-diagonal populations in the confusion matrix with independent validation using the DESI DR1 AGN/galaxy VAC.

4.4.1. Assessing AGN predictions with stacked spectra

To further assess the reliability of our model, we constructed stacked spectra for each off-diagonal cell of the confusion matrix (Fig. 1), corresponding to sources for which the model prediction and the reference classification disagreed. The stack spectra were computed using the *desigal* package (DESI Collaboration 2026b), applying an inverse-variance weighted mean with bootstrap resampling to ensure robustness.

We first focused on the 286 sources labeled as other/passive by classical diagnostics but predicted as AGNs by our model. The emission-line measurements derived from the stacked spectrum using *FastSpecFit* are shown in the [NII]–BPT diagram (Fig. 6), including the Seyfert/LINER division from Schawinski et al. (2007), where Seyfert galaxies exhibit high-ionization emission lines and LINERs lower-ionization spectra (Padovani et al. 2017). The stack lies in the Seyfert region, indicating a high-ionization nucleus consistent with AGN activity.

Although many individual spectra lack a sufficient S/N to be classified using standard diagnostic diagrams, the stacked spectrum demonstrates that the line ratios of this population are consistent with AGN excitation on average. This is consistent with the fact that 144 out of 286 sources (50.3%) fail the $S/N > 3$ requirement in at least one of the BPT lines, with 121 of these (84.0%) failing in $H\beta$ and/or [OIII], highlighting the dominant role of the vertical BPT axis. A detailed breakdown of the limiting emission lines is provided in Appendix C.

Moreover, 98 of these 286 sources (34.3%) are classified as BL sources within the extended classification scheme, although they are not initially labeled as such due to a low S/N in the $H\alpha$ broad component. The stacked spectrum reveals a clear broad $H\alpha$ component with an FWHM exceeding $\sim 1200 \text{ km s}^{-1}$, providing additional evidence of their BL nature (see Appendix E). A comprehensive analysis of all discrepant prediction–label combinations, including additional diagnostic diagrams, is presented in Appendix D.

4.4.2. Using the AGN/galaxy VAC for DESI DR1

To further validate the AGN predictions made by our model, we cross-matched the 1121 sources predicted as AGNs but labeled as star-forming (835, 74.5%) or other/passive (286, 25.5%) (see Fig. 1) with the AGN/galaxy VAC of DESI DR1, which compiles AGN identifications based on multiple diagnostics beyond the

Table 4. AGNs identified by the DESI DR1 AGN/galaxy VAC among sources predicted as AGNs but labeled differently.

AGN diagnostic method	Number of sources (%)
Total unique AGNs detected	1048 (93.5%)
Detected by ≥ 2 methods	885 (78.9%)
Detected by ≥ 3 methods	523 (46.6%)
Optical AGN (non-BPT)	974 (86.9%)
BPT: Seyfert (robust)	176 (15.7%)
Mass-excitation AGN	822 (73.3%)
WHAN: strong AGN	792 (70.6%)
WISE color diagnostic	317 (28.3%)
Kinematics-excitation AGN	294 (26.2%)
BPT (Seyfert/LINER)	239 (21.3%)
Broad emission lines	216 (19.3%)
QuasarNet reclassifies as QSO	167 (14.9%)
BLUE diagram: AGN	97 (8.6%)
Ne V AGN	85 (7.6%)
Redrock QSO template	81 (7.2%)
MgII broad line (afterburner)	46 (4.1%)
WHAN: weak AGN	34 (3.0%)
He II BPT AGN	22 (2.0%)

Notes. Number and percentage (relative to 1121 total sources classified as AGNs by our model but labeled as star-forming or other/passive) of AGNs detected by each diagnostic method from the DESI DR1 AGN/galaxy VAC. Overlaps between methods are expected. Summary counts for multi-diagnostic identifications are included, along with details of AGNs detected by non-BPT methods and robust BPT.

[NII]–BPT. Of these 1121 sources, 1048 (93.5%) are identified as AGNs by at least one diagnostic method in the AGN/galaxy VAC. Among these, 885 (78.9%) are supported by at least two independent diagnostics and 523 (46.6%) are supported by three or more, highlighting the robustness of the classification. A breakdown of the contributing diagnostic methods is provided in Table 4.

The majority of identified AGNs, 974 (86.9%), correspond to non-BPT optical diagnostics. Among these, 822 (73.3%) are classified through the mass–excitation diagram (Juneau et al. 2014) and 792 (70.6%) as strong AGNs according to the WHAN method (Cid Fernandes et al. 2011). Additional optical diagnostics include the kinematics–excitation diagram (Zhang & Hao 2018), the blue diagram (Lamareille et al. 2004; Lamareille 2010), the [Ne V] line diagnostic (Schmidt et al. 1998), and the He II BPT diagram (Shirazi & Brinchmann 2012). In addition to the classical [N II]–BPT diagram, the VAC also incorporates [SII]- and [OI]-based BPT diagnostics following Law et al. (2021); Kewley et al. (2006). In the infrared, 317 (28.3%) sources are identified through WISE color criteria, satisfying at least one of the AGN selection schemes proposed in Jarrett et al. (2011); Mateos et al. (2012); Stern et al. (2012); Assef et al. (2018); Yao et al. (2020); Hviding et al. (2022). QuasarNet (Busca & Balland 2018; Chaussidon et al. 2023), an ML classifier originally developed for SDSS spectra and subsequently adapted for DESI, was used to identify QSOs and to reclassify ambiguous targets. Finally, 216 (19.3%) sources exhibit broad emission lines.

5. Discussion

5.1. AGN identification

Multiwavelength AGN samples selected in the X-ray, infrared, optical/UV, or radio occupy partially disjoint regions of classical

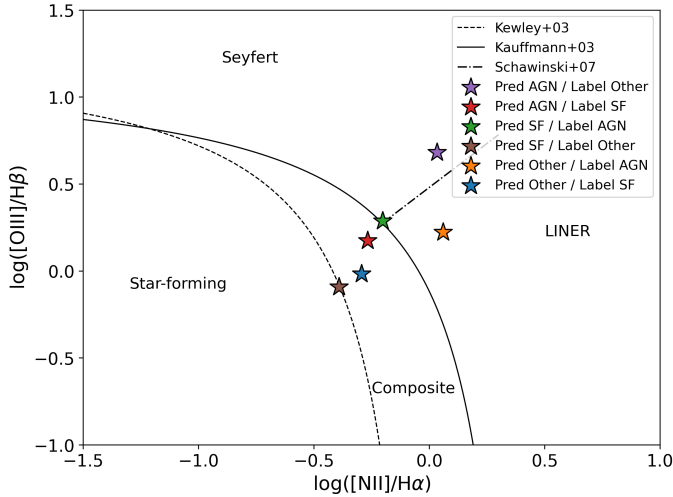


Fig. 6. Location of the stacked spectra corresponding to sources with discrepant classifications on the [NII]–BPT diagram. Each symbol represents the emission-line ratios measured from the FastSpecFit model of the corresponding stacked spectrum. The classical demarcation lines separating star-forming, composite, and AGN (Seyfert and LINER) regions are shown for reference.

optical diagnostic diagrams, so neither a single criterion nor combinations of criteria can capture the full diversity of AGN populations (Juneau et al. 2013; Trump et al. 2015; Padovani et al. 2017; Sánchez et al. 2025). In particular, BPT-based classifications (the most widely used classifications) can miss a substantial fraction of bona fide AGNs due to weak or obscured emission lines or intrinsically low accretion states (Cid Fernandes et al. 2010; Padovani et al. 2017; Sánchez et al. 2025); for instance, up to $\sim 40\%$ of X-ray–selected AGNs cannot be placed on the BPT diagram because of weak lines (Agostino & Salim 2019). As a result, traditional selections tend to underestimate the AGN fraction and motivate approaches that exploit the full spectral information instead of a limited set of emission-line ratios.

Our model performed strongly in AGN identification, achieving an accuracy of 0.952 and an AUC of 0.976 in the three-class definition (AGN, star-forming, and other/passive; Section 4.1). Direct comparisons with previous work are challenging because datasets, class definitions, and evaluation strategies are different because most autoencoder-based approaches primarily rely on qualitative inspection of latent spaces. Of the quantitative studies, the supervised GaSNet-II model (Zhong et al. 2024) reported an AGN accuracy of 0.93 on SDSS spectra. While our semi-supervised approach attains a slightly higher accuracy, comparisons should be interpreted with caution given the differing training setups and datasets.

Moreover, our approach recovers AGN signatures from low-S/N spectra that are missed by standard BPT-based classifications. This is illustrated by the stacked spectrum of sources predicted as AGNs but labeled as other/passive, which lies firmly in the AGN region of the [NII]–BPT diagram (Fig. 6; see also Section 4.4.1). Although individual spectra often lack a sufficient S/N for classification (Appendix C), the stacked spectrum reveals clear AGN-like line ratios. Moreover, 34.3% of these sources are classified as BL in the extended scheme, and their stacked spectra (Figure E.1) show clear broad Balmer components, demonstrating that the model can recover NL and BL AGNs that are missed by traditional line-based selection due to S/N limitations.

This interpretation is reinforced by the cross-match with the VAC (Section 4.4.2), where 93.5% of the 1121 sources predicted as AGNs but labeled as star-forming or other/passive are identified as AGNs by at least one independent diagnostic, and nearly half by three or more. Most of these detections rely on optical diagnostics beyond the [NII]–BPT diagram, which is known to be sensitive to metallicity effects (Storchi-Bergmann et al. 2006; Carvalho et al. 2020). This confirms that many of these sources are genuine AGNs missed by single-diagram approaches, and it implies that our reported performance metrics represent conservative lower limits on the true AGN identification capability of the model. The high precision of the AGN predictions of our model indicates that when a source is classified as an AGN, there is strong confidence in its validity, suggesting that the resulting AGN sample is both reliable and more complete than those produced by traditional approaches.

Conversely, for sources labeled as AGNs but predicted as other/passive (Appendix D.1), the stacked spectrum lies in the LINER region of the [NII]–BPT diagram and in the weak-AGN regime of the WHAN diagram, close to the retired boundary ($\text{EW}(\text{H}\alpha) \approx 3 \text{ \AA}$). This suggests that a fraction of these systems might be ionized by evolved stellar populations and not by AGNs. In this regime, BPT ratios alone can misclassify retired galaxies as AGNs (Stasińska et al. 2015; Sánchez et al. 2025). In this sense, the model prediction reduces contamination of the AGN class introduced by labeling based on the line ratio.

In the seven-class experiment (Section 4.2), the model also performed strongly in identifying BL sources. Most disagreements occur between BL and NL AGNs (Figure 3), highlighting the intrinsically ambiguous nature of defining broad emission lines through fixed-width thresholds. Consistently, 62% of sources labeled as NL AGNs but predicted as BL by our model are identified as BL in the AGN/Galaxy VAC of DESI DR1, which adopts broader line-width criteria (Table F.1). This result highlights the flexibility of our approach in recovering BL AGNs beyond restrictive single-line definitions (further details are provided in Appendix F).

From a methodological perspective, our approach can be regarded as semi-supervised because it combines unsupervised spectral representation learning with a label-based classification within a single workflow. The autoencoder operates in a fully data-driven manner, learning a low-dimensional representation of the full optical spectrum without using labels. When it is trained, a k-d tree is constructed in the latent space to exploit local proximity and identify similar spectra. The workflow becomes semi-supervised at the classification stage, where external labels (derived here from FastSpecFit emission-line measurements) are used as reference classifications to assign class predictions and compute quantitative performance metrics. Importantly, the learned latent representations are not explicitly optimized to reproduce any specific diagnostic classification and therefore do not inherit the biases associated with a given labeling scheme, unlike supervised models trained directly on labeled data. As demonstrated by the stacking analysis and the cross-matching with independent diagnostics, based on this separation, the resulting predictions can mitigate known limitations of classical methods, such as incompleteness at low S/N or contamination by retired galaxies. This design also makes it possible to project alternative classification schemes onto the same latent space, including those based on different optical diagnostics or multiwavelength AGN selections, without retraining the representation itself. Moreover, when it is trained on a large spectroscopic sample, class assignments can be calibrated using a limited high-quality labeled subset and propagated

to the remaining sources through proximity in latent space, reducing the reliance on traditional line-based diagnostics and enabling a scalable AGN classification for future DESI data releases.

5.2. Latent space structure and physical interpretation

In Figure 4, the learned latent space clearly separates NL AGNs and BL sources, validating the discriminative power of our model. Some overlap exists, particularly in specific regions, such as the left side. A notable overlap is also observed between star-forming and composite galaxies. While we expect this overlap to be significantly reduced when scaling the model to larger datasets from future DESI releases, part of it will inevitably persist due to physical causes, such as the smooth transition between AGN-dominated and host-galaxy-dominated spectra (Cid Fernandes et al. 2010).

The UMAP projection also shows that galaxies classified as other are widely distributed and overlap several other classes (see Figure 4). This behavior, together with the relatively low performance of the model for this category (e.g., a recall of 0.327 in the seven-galaxy subpopulation scheme; see Table 3), is expected because these sources do not share a well-defined or homogeneous spectroscopic signature in the latent space. Objects are labeled as other when their nature cannot be robustly determined from the available emission-line measurements, often because of low-S/N diagnostic lines, ambiguous spectral features, or incomplete measurements. Since the model learns directly from the full optical spectrum and not from a limited set of emission-line ratios, many of these galaxies can be associated with physically similar populations in the latent space. This is particularly evident for AGN sources recovered within the population classified as other (Section 4.4), supporting the interpretation that latent representation can recover physically meaningful structures beyond the discrete categories defined by emission-line diagnostics.

In the case of passive and retired galaxies (whose classification metrics are comparatively poor), these classes are typically defined using thresholds in emission-line strengths or equivalent widths, particularly $H\alpha$, as adopted in this work, making them highly sensitive to observational quality (Cid Fernandes et al. 2011). In the latent space, these populations separate into two coherent clusters that do not align with their reference labels, suggesting underlying physical differences not captured by the current classification scheme. Galaxies in the upper left cluster tend to be more massive and exhibit lower SFRs (see Figure 5), indicating that stellar population properties might provide a more physically meaningful basis for classification than the $EW_{H\alpha}$ -based definition of passive and retired galaxies. A similar separation has been reported in other dimensionality-reduction studies (Siudek et al. 2018b,a; Portillo et al. 2020; Siudek et al. 2022a).

Visual inspection of the latent space is a powerful feature of these models, helping to break the so-called black box nature commonly associated with other ML applications (Andrianomena & Tang 2024). The latent space structure provides astrophysical insight into the spectral and population diversity of galaxies, enabling the exploration of transitions and the discovery of new groupings beyond classical labels.

5.3. Implications, challenges, and future work

Traditional diagnostics for identifying AGNs face two main challenges: different techniques often select different AGN samples, leading to incomplete and inconsistent catalogs; and these

methods typically rely on a detailed spectral analysis, which is difficult to scale to the large volumes of data produced by modern surveys (Padovani et al. 2017). While ML algorithms can address scalability, supervised methods risk propagating the incompleteness by relying on labels based on arbitrary decisions (Ball & Brunner 2010; Baron 2019). Unsupervised representation learning overcomes this issue by extracting information from the entire spectrum, not just specific lines, and does so without requiring a priori definitions. However, these models do not directly provide galaxy classifications or quantitative metrics.

Our method addresses this limitation by combining an autoencoder with a k-d tree nearest-neighbor search in the latent space to assign labels. Although the final predictions still depend on predefined labels (and thus, can inherit their biases) our findings in Section 4.4 show that the method can recover sources that are missed by the original diagnostics, but are supported by alternative diagnostics not used during labeling, resulting in more complete catalogs. Additionally, the approach offers flexibility and interpretability: the k-d tree algorithm can be applied to different label definitions without retraining the network, and the latent space visualization provides insight into physical trends and ambiguous regions.

Nevertheless, to truly address the challenge of incompleteness, we plan to integrate unsupervised spectroscopic algorithms with multiwavelength data, such as mid-IR, X-ray, or photometric observations. This integration will be essential for the building of more comprehensive AGN catalogs and to capture the continuous transition between AGN-dominated and host-dominated spectral features (Kewley et al. 2006; Hickox & Alexander 2018; Pfeifle et al. 2024). A key challenge will be to manage missing data because not all sources have matched observations at all wavelengths. To address this, we will explore multi-branch architectures (Zhang et al. 2019), feature-level fusion (Ross & Govindarajan 2005; Sharma et al. 2025; Yang et al. 2025), and strategies for handling missing data, such as masking, imputation, or robust models such as XGBoost (Chen & Guestrin 2016).

The interpretability of the latent space opens several scientific avenues for future work, such as exploring how it traces dust and stellar population age, and whether slope changes reflect reddening or intrinsic AGN properties. A promising but challenging extension would be to separate the contributions by AGN and host, particularly in low-mass galaxies. Such analyses might clarify whether AGNs preferentially inhabit star-forming or evolved systems, with implications for feedback and quenching (Piotrowska et al. 2022).

Moreover, as mentioned in Section 1, broad emission lines can also be produced by transient stellar phenomena, particularly in low-mass galaxies, and follow-up spectroscopy is often required to disentangle these cases from genuine AGN activity (Baldassare et al. 2016; Panda & Śniegowska 2024). In future work, we plan to investigate whether our model can help break this degeneracy and distinguish BL signatures originating from AGNs from those produced by stellar processes using single-epoch spectra. Specifically, training in more carefully curated samples might be required to fully address this question.

Finally, based on the promising results of this study, we are now applying our method to a subset of DESI DR2 (Abdul Karim et al. 2025), comprising over 14 million galaxies and QSOs from three years of observations. This is nearly 280 times larger than the 50 000 sources analyzed in our current sample, enabling significantly improved statistical precision and more robust constraints on AGN populations and their

connection to galaxy evolution. The flexibility of our framework, together with the straightforward preprocessing and computational efficiency of the SPENDER model, makes it well suited for scaling to large datasets. Training on larger samples is expected to further improve accuracy and robustness, enabling the construction of more complete AGN catalogs. Furthermore, this approach provides a pathway to investigating the physical processes underlying AGN activity and their connection to host-galaxy properties.

6. Conclusions

Traditional AGN identification methods often yield incomplete and inconsistent samples because different diagnostics select different subsets of sources and require a detailed spectral analysis that does not scale well with modern surveys. Our results demonstrated that advanced data-driven approaches can help mitigate these biases. We showed that autoencoders can learn latent representations of galaxy spectra that correlate strongly with physical galaxy classes. Unlike previous studies that relied on a visual inspection of the latent space (Portillo et al. 2020), our work introduced a semi-supervised framework for galaxy classification that combines unsupervised feature extraction with a k-d tree classifier, using labels derived from FastSpecFit.

Our main conclusions are summarized below:

- The proposed model achieved high performance in AGN identification in large spectroscopic surveys, with an accuracy of 0.952 and an AUC of 0.976 in a three-class scheme;
- The model reliably recovered AGNs in low-S/N spectra, overcoming the limitations of traditional diagnostics that depend on strict emission-line thresholds, as confirmed by stacking analyses and multi-diagnostic validation;
- BL AGNs were effectively identified (accuracy 0.965 and AUC 0.973), including sources missed by classical methods due to weak or undetected broad components in individual spectra, highlighting the sensitivity of the approach to BL features;
- Stacking analyses and cross-matching with independent diagnostics showed that many sources predicted as AGNs but unlabeled by classical criteria are bona fide AGNs; in particular, 93.5% of these sources were confirmed by at least one diagnostic in the AGN/galaxy VAC. This implies that the reported performance metrics represent conservative lower limits on the true model capability;
- Conversely, the model reduced contamination in the AGN class by identifying sources whose emission was consistent with retired or weakly ionized galaxies. In this regime, line-ratio diagnostics alone can be misleading;
- The learned latent space encodes physically meaningful structure, with clear correlations with galaxy classes and properties such as stellar mass and SFR, providing interpretability and insight into ambiguous or transition populations, and revealing, for instance, two distinct populations of passive galaxies;
- The framework is robust to biases inherent to specific labeling schemes and was validated using independent diagnostics from the DESI AGN/galaxy VAC.

Overall, this work demonstrates that unsupervised spectral representation learning, combined with flexible labeling strategies, provides a powerful and scalable pathway toward more complete and reliable AGN catalogs. By leveraging the full spectral information, this approach overcomes key limitations of traditional

emission-line diagnostics, particularly in low-S/N and ambiguous regimes. The semi-supervised framework is well suited for integration with multiwavelength data and for application to upcoming DESI data releases, enabling improved AGN demographics and more robust constraints on the connection between nuclear activity and galaxy evolution.

Data availability

The data supporting the findings of this study are publicly available at <https://doi.org/10.5281/zenodo.20490011>. The code used to reproduce the analyses presented in this work is publicly available at <https://doi.org/10.5281/zenodo.21069095>.

Acknowledgements. M.S. acknowledges support by the State Research Agency of the Spanish Ministry of Science and Innovation under the grants ‘Galaxy Evolution with Artificial Intelligence’ (PGC2018-100852-A-I00) and ‘BASALT’ (PID2021-126838NB-I00) and the Polish National Agency for Academic Exchange (Bekker grant BPN/BEK/2021/1/00298/DEC/1). This work was partially supported by the European Union’s Horizon 2020 Research and Innovation program under the Maria Skłodowska-Curie grant agreement (No. 754510). M.M. acknowledges support from the Spanish Ministry of Science and Innovation through the project PID2021-124243NBC22. This work was partially supported by the program Unidad de Excelencia María de Maeztu CEX2020-001058-M. M.E. acknowledges support from the Spanish Ministry of Science and Innovation through the project PID2023-152069NA-I00. This work has received funding from the ‘Severo Ochoa Centres of Excellence’ programme (grant CEX2024-001441-S), funded by MICIU/AEI/10.13039/501100011033. R.P. is currently supported by the University of Utah, and was previously supported by the University of Arizona and, in part, by U.S. NSF NOIRLab. The research of S.J. is supported by U.S. NSF NOIRLab, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. S.P. is supported by the International Gemini Observatory, a program of NSF NOIRLab, which is managed by AURA under a cooperative agreement with the U.S. National Science Foundation on behalf of the Gemini partnership of Argentina, Brazil, Canada, Chile, the Republic of Korea, and the United States of America. This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF’s National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONACYT); the Ministry of Science, Innovation and Universities of Spain (MICIU/AEI/10.13039/501100011033), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies. The authors are honored to be permitted to conduct scientific research on I’oligam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation.

References

- Abdul Karim, M., Aguilar, J., Ahlen, S., et al. 2025, *Phys. Rev. D*, **112**, 083514
- Agostino, C. J., & Salim, S. 2019, *ApJ*, **876**, 12
- Anand, A., Guy, J., Bailey, S., et al. 2024, *ApJ*, **168**, 124
- Andrianomena, S., & Tang, H. 2024, *JCAP*, **06**, 034
- Angeloudi, E., Audenaert, J., Bowles, M., et al. 2024, in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*
- Antonucci, R. 1993, *ARA&A*, **31**, 473
- Assef, R. J., Stern, D., Noirot, G., et al. 2018, *ApJS*, **234**, 23
- Baldassare, V. F., Reines, A. E., Gallo, E., et al. 2016, *ApJ*, **829**, 57
- Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, *Publ. Astron. Soc. Pac.*, **93**, 5

- Ball, N. M., & Brunner, R. J. 2010, *Int. J. Mod. Phys. D*, **19**, 1049
- Baron, D. 2019, *MNRAS*, **487**, 5450
- Beckmann, V., & Shrader, C. R. 2012, *The Central Engine* (John Wiley & Sons, Ltd), 35
- Bentley, J. L. 1975, *Commun. ACM*, **18**, 509
- Böhm, V., Kim, A. G., & Juneau, S. 2023, *MNRAS*, **526**, 3072
- Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, *A&A*, **622**, A103
- Busca, N., & Balland, C. 2018, arXiv e-prints [arXiv:1808.09955]
- Cann, J. M., Satyapal, S., Abel, N. P., et al. 2019, *ApJ*, **870**, L2
- Carvalho, S. P., Dors, O. L., Cardaci, M. V., et al. 2020, *MNRAS*, **492**, 5675
- Chaussidon, E., Yèche, C., Palanque-Delabrouille, N., et al. 2023, *ApJ*, **944**, 107
- 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
- Chilingarian, I. V., & Zolotukhin, I. Y. 2011, *MNRAS*, **419**, 1727
- Cid Fernandes, R., Stasińska, G., Schlickmann, M. S., et al. 2010, *MNRAS*, **403**, 1036
- Cid Fernandes, R., Stasińska, G., Mateus, A., & Vale Asari, N. 2011, *MNRAS*, **413**, 1687
- De Jong, T. S., Agertz, O., Berbel, A. A., et al. 2019, *The Messenger*, **175**, 3
- DESI Collaboration (Aghamousa, A., et al.) 2016, arXiv e-prints [arXiv:1611.00037]
- DESI Collaboration (Abareshi, B., et al.) 2022, *ApJ*, **164**, 207
- DESI Collaboration (Adame, A. G., et al.) 2024, *AJ*, **168**, 58
- DESI Collaboration (Karim, M. A., et al.) 2026a, *AJ*, **171**, 285
- DESI Collaboration 2026b, <https://github.com/desihub/desigal>
- Dey, A., Schlegel, D. J., Lang, D., et al. 2019, *ApJ*, **157**, 168
- Fawcett, V. A., Alexander, D. M., Brodzeller, A., et al. 2023, *MNRAS*, **525**, 5575
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press), <http://www.deeplearningbook.org>
- Greene, J. E., & Ho, L. C. 2005, *ApJ*, **630**, 122
- Guo, H., Shen, Y., & Wang, S. 2018, PyQSOFit: Python code to fit the spectrum of quasars, Astrophysics Source Code Library [record ascl:1809.008]
- Guo, W.-J., Zou, H., Fawcett, V. A., et al. 2024, *ApJS*, **270**, 26
- Guy, J., Bailey, S., Kremin, A., et al. 2023, *ApJ*, **165**, 144
- Hahn, C., Gontcho, S. G. A., Melchior, P., et al. 2025, arXiv e-prints [arXiv:2506.18986]
- Harrison, C. M., & Ramos Almeida, C. 2024, *Galaxies*, **12**
- Heckman, T. M., & Best, P. N. 2014, *ARA&A*, **52**, 589
- Hickox, R., & Alexander, D. 2018, *ARA&A*, **56**, 625
- Huertas-Company, M., & Lanusse, F. 2023, *PASA*, **40**, e001
- Hviding, R. E., Hainline, K. N., Goulding, A. D., & Greene, J. E. 2024, *ApJ*, **167**, 169
- Hviding, R. E., Hainline, K. N., Rieke, M., et al. 2022, *AJ*, **163**, 224
- Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, *ApJ*, **735**, 112
- Jespersen, C. K., Melchior, P., Spergel, D. N., et al. 2026, *ApJ*, **1002**, 132
- Jin, S., Trager, S. C., Dalton, G. B., et al. 2024, *MNRAS*, **530**, 2688
- Johansson, J., Woods, T. E., Gilfanov, M., et al. 2016, *MNRAS*, **461**, 4505
- Juneau, S., Dickinson, M., Bournaud, F., et al. 2013, *ApJ*, **764**, 176
- Juneau, S., Bournaud, F., Charlot, S., et al. 2014, *ApJ*, **788**, 88
- Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, *MNRAS*, **346**, 1055
- Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, *ApJ*, **556**, 121
- Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, *MNRAS*, **372**, 961
- Kristensen, M. T., Pimbblet, K., & Penny, S. 2020, *MNRAS*, **496**, 2577
- Lamareille, F. 2010, *A&A*, **509**, A53
- Lamareille, F., Mouhcine, M., Contini, T., Lewis, I., & Maddox, S. 2004, *MNRAS*, **350**, 396
- Law, D. R., Ji, X., Belfiore, F., et al. 2021, *ApJ*, **915**, 35
- Liang, Y., Melchior, P., Hahn, C., et al. 2023a, *ApJ*, **956**, L6
- Liang, Y., Melchior, P., Lu, S., Goulding, A., & Ward, C. 2023b, *ApJ*, **166**, 75
- Lisiecki, K., Malek, K., Siudek, M., et al. 2023, *A&A*, **669**, A95
- Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. 2012, *MNRAS*, **426**, 3271
- McInnes, L., Healy, J., & Melville, J. 2020, arXiv e-prints [arXiv:1802.03426]
- McInnes, L., Healy, J., Saul, N., & Grossberger, L. 2018, *J. Open Source Softw.*, **3**, 861
- Melchior, P., Liang, Y., Hahn, C., & Goulding, A. 2023, *ApJ*, **166**
- Mezcua, M. 2017, *Int. J. Mod. Phys. D*, **26**, 1730021
- Mezcua, M., Siudek, M., Suh, H., et al. 2023, *ApJ*, **943**, L5
- Miller, T. N., Doel, P., Gutierrez, G., et al. 2024, *ApJ*, **168**, 95
- Moran, E. C., Shahinyan, K., Sugarman, H. R., Vélez, D. O., & Eracleous, M. 2014, *AJ*, **148**, 136
- Moustakas, J. 2023, FastSpecFit: Fast spectral synthesis and emission-line fitting of DESI spectra, Astrophysics Source Code Library [record ascl:2308.005]
- Netzer, H. 2015, *ARA&A*, **53**, 365
- Nicolaou, C., Nathan, R. P., Lahav, O., & et al. 2026, *MNRAS*, **547**
- Padovani, P., Alexander, D., Assef, R., & et al. 2017, *A&AR*, **25**
- Pan, Z., Jiang, L., Guo, W.-J., et al. 2025, *ApJ*, **987**, 48
- Panda, S., & Śniegowska, M. 2024, *ApJSS*, **272**, 13
- Pat, F., Juneau, S., Böhm, V., et al. 2022, arXiv e-prints [arXiv:2211.11783]
- Pattanaik, R., Kartaltepe, J. S., & Binu, C. 2025, *ApJ*, **988**, 139
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, **12**, 2825
- Pfeifle, R., Weaver, K., Secrest, N., & Rothberg, B. 2024, in *American Astronomical Society Meeting Abstracts*, **243**, 218.03
- Piotrowska, J. M., Bluck, A. F. L., Maiolino, R., & Peng, Y. 2022, *MNRAS*, **512**, 1052
- Poppett, C., Tyas, L., Aguilar, J., et al. 2024, *AJ*, **168**, 245
- Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, *ApJ*, **160**, 45
- Pucha, R., Juneau, S., Dey, A., et al. 2025, *ApJ*, **982**, 10
- PyTorch. 2025, torch.optim.lr_scheduler.OneCycleLR Documentation, https://docs.pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.OneCycleLR.html
- Pérez-Díaz, V. S., Martínez-Galarza, J. R., Caicedo, A., & D'Abrusco, R. 2024, *MNRAS*, **528**, 4852
- Ramasawmy, J., Stevens, J., Martin, G., & Geach, J. E. 2019, *MNRAS*, **486**, 4320
- Risaliti, G., & Lusso, E. 2019, *Nat. Astron.*, **3**, 272
- Rosito, M. S., Pedrosa, S. E., Tissera, P. B., et al. 2021, *A&A*, **652**, A44
- Ross, A. A., & Govindarajan, R. 2005, *SPIE Conf. Ser.*, **5779**, 196
- Salpeter, E. E. 1964, *ApJ*, **140**, 796
- Sánchez, S. F., Muñoz-Tuñón, C., Sánchez Almeida, J., González-Martín, O., & Pérez, E. 2025, *A&A*, **704**, A145
- Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, *MNRAS*, **382**, 1415
- Schlafly, E. F., Kirkby, D., Schlegel, D. J., et al. 2024, arXiv e-prints [arXiv:2306.06309]
- Schmidt, M., Hasinger, G., Gunn, J., et al. 1998, *A&A*, **329**, 495
- Sharma, A., Varman, U., Bharti, V., et al. 2025, *IEEE J. Biomed. Health Inform.*, **1**
- Shen, Y., Hwang, H.-C., Zakamska, N., & Liu, X. 2019, *ApJ*, **885**, L4
- Shirazi, M., & Brinchmann, J. 2012, *MNRAS*, **421**, 1043
- Silber, J. H., Fagrelus, P., Fanning, K., et al. 2023, *AJ*, **165**, 9
- Siudek, M., Malek, K., Pollo, A., et al. 2018a, arXiv e-prints [arXiv:1805.09905]
- Siudek, M., Malek, K., Pollo, A., et al. 2018b, *A&A*, **617**, A70
- Siudek, M., Lisiecki, K., Mezcua, M., et al. 2022a, arXiv e-prints [arXiv:2211.11792]
- Siudek, M., Mezcua, M., & Krywult, J. 2022b, *MNRAS*, **518**, 724
- Siudek, M., Pucha, R., Mezcua, M., et al. 2024, *A&A*, **691**, A308
- Siudek, M., Mezcua, M., Circosta, C., et al. 2025, *A&A*, **700**, A209
- Stasińska, G., Costa-Duarte, M. V., Vale Asari, N., Cid Fernandes, R., & Sodré, L., J. 2015, *MNRAS*, **449**, 559
- Stern, D., Assef, R. J., Benford, D. J., et al. 2012, *ApJ*, **753**, 30
- Storchi-Bergmann, T., Schmitt, R. A., Schmitt, H. R., & Dors, O. L. 2006, *MNRAS*, **371**, 1559
- Suh, H., Civano, F., Trakhtenbrot, B., et al. 2020, *ApJ*, **889**, 32
- Tamura, N., Takato, N., Shimono, A., et al. 2016, in *Ground-based and Airborne Instrumentation for Astronomy VI*, 9908, eds. C. J. Evans, L. Simard, & H. Takami (SPIE), 99081M
- Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., & Bluck, A. F. L. 2022, *ApJ*, **163**, 71
- Tous, J. L., Solanes, J. M., & Perea, J. D. 2025, *MNRAS*, **537**, 1459
- Trump, J. R., Sun, M., Zeimann, G. R., et al. 2015, *ApJ*, **811**, 26
- Urry, C. M., & Padovani, P. 1995, *PASP*, **107**, 803
- Vogelsberger, M., Genel, S., Springel, V., et al. 2014, *Nature*, **509**, 177
- Wu, Y., Tao, Y., Fan, D., Cui, C., & Zhang, Y. 2023, *MNRAS*, **527**, 1163
- Yang, B., Jiang, Z., Pan, D., et al. 2025, *Inform. Fusion*, **113**, 102639
- Yao, H. F. M., Jarrett, T. H., Cluver, M. E., et al. 2020, *ApJ*, **903**, 91
- Zel'dovich, Y. B. 1964, *Sov. Phys. Dokl.*, **9**, 195
- Zhang, K., & Hao, L. 2018, *ApJ*, **856**, 171

- Zhang, H., Shao, J., & Salakhutdinov, R. 2019, in *Proc. Mach. Learn. Res.*, 89, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, eds. K. Chaudhuri, & M. Sugiyama (PMLR), 1099
- Zhong, F., Napolitano, N. R., Heneka, C., et al. 2024, *MNRAS*, 532, 643
-
- ¹ Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Magrans, 08193 Barcelona, Spain
 - ² Instituto de Astrofísica de Canarias (IAC), Departamento de Astrofísica, Universidad de La Laguna (ULL), 38200 La Laguna, Tenerife, Spain
 - ³ Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, 08193 Bellaterra (Barcelona), Spain
 - ⁴ Port d'Informació Científica (PIC), Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
 - ⁵ Institut d'Estudis Espacials de Catalunya (IEEC), Edifici RDIT, Campus UPC, 08860 Castelldefels (Barcelona), Spain
 - ⁶ Department of Physics and Astronomy, University of Utah, 115 South 1400 East, Salt Lake City, UT 84112, USA
 - ⁷ Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
 - ⁸ NSF NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA
 - ⁹ University of Virginia, Department of Astronomy, Charlottesville, VA 22904, USA
 - ¹⁰ International Gemini Observatory/NSF NOIRLab, Casilla 603, La Serena, Chile
 - ¹¹ Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
 - ¹² Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA
 - ¹³ Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
 - ¹⁴ INAF – Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
 - ¹⁵ Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK
 - ¹⁶ Instituto de Física, Universidad Nacional Autónoma de México, Circuito de la Investigación Científica, Ciudad Universitaria, Cd. de México C. P. 04510, Mexico
 - ¹⁷ Department of Astronomy & Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
 - ¹⁸ Department of Physics & Astronomy and Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, 3941 O'Hara Street, Pittsburgh, PA 15260, USA
 - ¹⁹ University of California, Berkeley, 110 Sproul Hall #5800 Berkeley, CA 94720, USA
 - ²⁰ Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys, 23, 08010 Barcelona, Spain
 - ²¹ Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, CP 111711, Bogotá, Colombia
 - ²² Observatorio Astronómico, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio H, CP 111711 Bogotá, Colombia
 - ²³ Institute of Cosmology and Gravitation, University of Portsmouth, Dennis Sciamia Building, Portsmouth PO1 3FX, UK
 - ²⁴ Fermi National Accelerator Laboratory, PO Box 500, Batavia, IL 60510, USA
 - ²⁵ Department of Astronomy, University of Texas at Austin, 2515 Speedway, TX 78712, USA
 - ²⁶ Institut d'Astrophysique de Paris. 98 bis boulevard Arago., 75014 Paris, France
 - ²⁷ IRFU, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France
 - ²⁸ Department of Physics, Southern Methodist University, 3215 Daniel Avenue, Dallas, TX 75275, USA
 - ²⁹ Department of Physics and Astronomy, University of California, Irvine 92697, USA
 - ³⁰ The Ohio State University, Columbus, 43210 OH, USA
 - ³¹ Sorbonne Université, CNRS/IN2P3, Laboratoire de Physique Nucléaire et de Hautes Energies (LPNHE), 75005 Paris, France
 - ³² Department of Physics and Astronomy, Siena University, 515 Loudon Road, Loudonville, NY 12211, USA
 - ³³ Department of Physics and Astronomy, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
 - ³⁴ Perimeter Institute for Theoretical Physics, 31 Caroline St. North, Waterloo, ON N2L 2Y5, Canada
 - ³⁵ Waterloo Centre for Astrophysics, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
 - ³⁶ Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, s/n, E-18008 Granada, Spain
 - ³⁷ Departament de Física, EEBE, Universitat Politècnica de Catalunya, c/Eduard Maristany 10, 08930 Barcelona, Spain
 - ³⁸ Department of Physics and Astronomy, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea
 - ³⁹ CIEMAT, Avenida Complutense 40, 28040 Madrid, Spain
 - ⁴⁰ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
 - ⁴¹ University of Michigan, 500 S. State Street, Ann Arbor, MI 48109, USA
 - ⁴² National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Road, Chaoyang District, Beijing 100101, PR China

Appendix A: Performance metrics

In this work, the performance metrics are defined as follows:

- Accuracy: The proportion of correct predictions, both true positives and true negatives, among all predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (\text{A.1})$$

where TP , TN , FP , and FN denotes the number of true positives, true negatives, false positives, and false negatives, respectively.

- Precision: The proportion of true positives among all instances predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (\text{A.2})$$

- Recall (also known as true positive rate, TPR): The proportion of true positives among all actual positive instances:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (\text{A.3})$$

- False positive rate (FPR): The proportion of false positives among all actual negatives:

$$\text{FPR} = \frac{FP}{FP + TN} \quad (\text{A.4})$$

- F1 score: The harmonic mean of precision and recall, providing a balanced measure of both:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (\text{A.5})$$

- Receiver Operating Characteristic (ROC) Curve: A plot of the true positive rate against the false positive rate as the classification threshold is varied. The implementation details are provided in Appendix B.
- Area Under the Curve (AUC): A scalar metric representing the area under the ROC curve, which summarizes the model's performance across all classification thresholds. An AUC of 1 indicates perfect classification, while an AUC of 0.5 corresponds to random guessing.

Appendix B: ROC curves

To construct the ROC curves, since the autoencoder does not provide direct probabilities of belonging to each class, we use the proportion of neighbors in the k-d tree voting for a given class as a proxy for class probability. For instance, if six of the neighbors are labeled AGNs, we interpret this as a 60% probability of being AGNs. By varying the threshold on this estimated probability, we obtain different TPR–FPR pairs to construct the ROC curve.

In panel (a) of Figure B.1, which shows the ROC curves for three galaxy definitions, the performance in balancing completeness and contamination remains good across different thresholds for all classes, with the AGN curve performing slightly better (higher AUC). In panel (b), where the extended galaxy classification is used, there is more variation between classes. BL performs the best, followed by NL AGNs and star-forming, while galaxies classified as other show the curve closest to the diagonal. Passive and retired galaxies, which are often misclassified between each other, reach high completeness only at the expense of considerable contamination, with their ROC curves peaking further to the right compared to the rest of the classes.

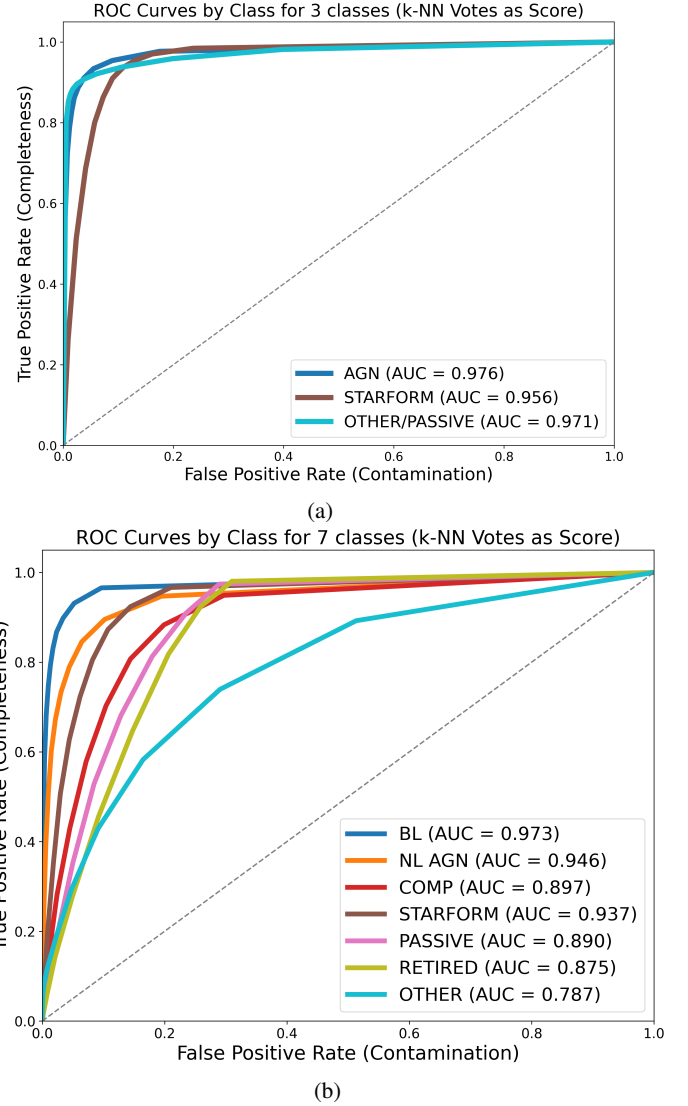


Fig. B.1: ROC curves for the classifier predictions, with the AUC value for each class given in parentheses. (a) ROC curves for the three galaxy classes: AGNs, star-forming, and other/passive. (b) ROC curves for the extended galaxy classes: NL AGNs, BL, star-forming, composite, retired, passive, and other.

Appendix C: S/N limitations of BPT diagnostics

Among the 286 sources labeled as other/passive but predicted as AGNs examined in Section 4.4.1, 144 (50.3%) fail the $S/N > 3$ requirement in at least one of the four BPT lines. Notably, 121 of these 144 sources (84.0%) fail the S/N threshold in either [O III] and/or $H\beta$, highlighting the dominant role of the vertical BPT axis. Individually, $H\beta$ is the most frequently limiting line, falling below the S/N threshold in 104 sources (36.4%), followed by [O III] in 81 sources (28.3%), [N II] in 72 sources (25.2%), and $H\alpha$ in 61 sources (21.3%).

Several alternative optical diagnostic diagrams have been proposed to mitigate the limitations of the classical BPT, particularly those associated with the reliance on the [O III]/ $H\beta$ ratio. Among them, the WHAN diagram (Cid Fernandes et al. 2010; Cid Fernandes et al. 2011), which combines the equivalent width of $H\alpha$ with the [N II]/ $H\alpha$ line ratio, was developed explicitly to address the large population of emission-line galaxies left

unclassified by BPT selections. By requiring only two emission lines and incorporating equivalent width information, WHAN is more sensitive to weak AGNs and low-luminosity systems, for which $H\beta$ is often undetectable or intrinsically weak (Cann et al. 2019; Sánchez et al. 2025).

However, studies comparing BPT and WHAN classifications have shown that while most BPT-selected AGNs are also identified as AGNs in WHAN, the converse is not true: a substantial fraction of WHAN-selected AGNs are classified as star-forming or composite in the BPT diagram (e.g., Kristensen et al. 2020). This does not imply that the WHAN classification is unreliable; rather, it reflects the fact that WHAN is designed to probe weaker or diluted AGNs, which may be missed by the more stringent four-line requirements of the BPT. Such systems can be affected by star-formation dilution, intrinsically weak emission lines, or environmental effects, all of which tend to suppress classical AGN signatures (Cid Fernandes et al. 2010; Cid Fernandes et al. 2011; Sánchez et al. 2025).

Previous studies have shown that a significant fraction of AGNs cannot be robustly classified using BPT diagnostics even when the spectroscopy is sufficiently deep, as a result of intrinsically weak or obscured narrow-line regions, radiatively inefficient accretion flows, or short AGN duty cycles; indeed, Agostino & Salim (2019) find that $\sim 40\%$ of X-ray-selected AGNs exhibit emission lines too weak to be placed on the BPT diagram and are often associated with low specific SFRs.

The above considerations highlight that no single optical diagnostic diagram can provide a complete and unbiased AGN census (Juneau et al. 2013; Padovani et al. 2017; Sánchez et al. 2025). This motivates the use of approaches that exploit the full spectral information, rather than relying on a small number of emission-line ratios.

Appendix D: Stack of sources with discrepant prediction–label combinations

Building on the analysis presented in Section 4.4.1, this section extends the study to all discrepant prediction–label combinations. While the main text focuses on sources predicted as AGNs but labeled as other/passive, here we examine the remaining cases. Figure D.1 shows the stacked spectra for each off-diagonal population in the confusion matrix. To complement the BPT-based classification discussed in the main text (Figure 6), we also place the line measurements of the stacked spectra on the WHAN diagram (Fig. D.2), which incorporates $H\alpha$ equivalent width and enables a more robust distinction between weak AGN and retired galaxies in low-ionization regimes. Table D.1 summarizes the classification of each stack according to both diagnostics. The following subsections examine each prediction–label combination in detail, providing a consistent interpretation of their physical nature.

D.1. Predicted as other/passive – labeled as AGNs

For sources originally labeled as AGNs but predicted as other/passive by our model, the stacked spectrum lies in the LINER region of the [NII]–BPT diagram, while its location in the WHAN diagram corresponds to the weak-AGN regime, very close to the retired boundary. In particular, the measured $H\alpha$ equivalent width of the stacked spectrum is $EW(H\alpha) \approx 3.0\text{\AA}$, which marks the transition between weak AGN and retired regions of the WHAN diagram.

This indicates that a fraction of the sources in this population may not host ongoing nuclear activity, but are instead ionized by old stellar populations. As emphasized by Stasińska et al. (2015), emission-line ratios alone can misclassify retired galaxies as LINERs if line equivalent widths are not taken into account, as the ionization in these systems is driven by hot low-mass evolved stars (HOLMES). By explicitly incorporating equivalent widths, the WHAN diagram provides a more robust classification in this low-ionization regime. In this context, the model prediction may effectively reduce contamination of the AGN class by retired galaxies.

D.2. Predicted as star-forming – labeled as other/passive

Similarly, stacking the 2011 sources predicted as star-forming by our model (including both star-forming and composite classes) but labeled as other/passive, and placing the resulting spectrum on the [NII]–BPT diagram (Figure D.1), shows that the stack lies firmly within the star-forming region. Stacking effectively recovers the characteristic star-formation-driven line ratios that are buried in the individual low-S/N spectra. The stacked spectrum is located very close to the Kewley et al. (2001) demarcation line separating star-forming and composite regions, indicating that, on average, these sources occupy the transition between pure star formation and composite excitation.

In the WHAN diagram, the stack is placed in the strong AGN region but falls close to the star-forming boundary. As discussed by Sánchez et al. (2025), late-type galaxies that are genuinely dominated by star formation can be classified as AGNs in equivalent-width-based diagnostics under certain conditions, despite their overall ionization being driven primarily by star formation. Taken together, the BPT and WHAN diagnostics suggest that this population is predominantly star-forming, consistent with the model prediction.

D.3. Predicted as star-forming – labeled as AGNs

The stacked spectrum of galaxies predicted as star-forming but labeled as AGNs shows clear AGN signatures in both diagnostic diagrams. In the [NII]–BPT diagram, the stack lies very close to the boundaries separating the Seyfert, LINER, and composite regions, while the WHAN diagram classifies it as a strong AGN. It is important to note that our star-forming class also includes composite galaxies; therefore, part of this apparent disagreement may reflect sources with mixed ionization.

D.4. Predicted as AGNs – labeled as star-forming

For sources originally labeled as star-forming but predicted as AGNs by our model, the stacked spectrum occupies the composite region of the [NII]–BPT diagram, while the WHAN diagram classifies it as a strong AGN. This behavior is consistent with well-known limitations of the BPT diagram, which can be biased against AGN identification when NL AGN signatures are overpowered by host galaxy star formation. This effect is particularly relevant in low-mass, blue galaxies with high specific star-formation rates, where AGN emission lines can be diluted by stellar processes (Agostino & Salim 2019; Trump et al. 2015). In addition, BPT line ratios are sensitive to metallicity, which can further bias the classification toward the star-forming or composite regions in such systems (Storchi-Bergmann et al. 2006; Carvalho et al. 2020).

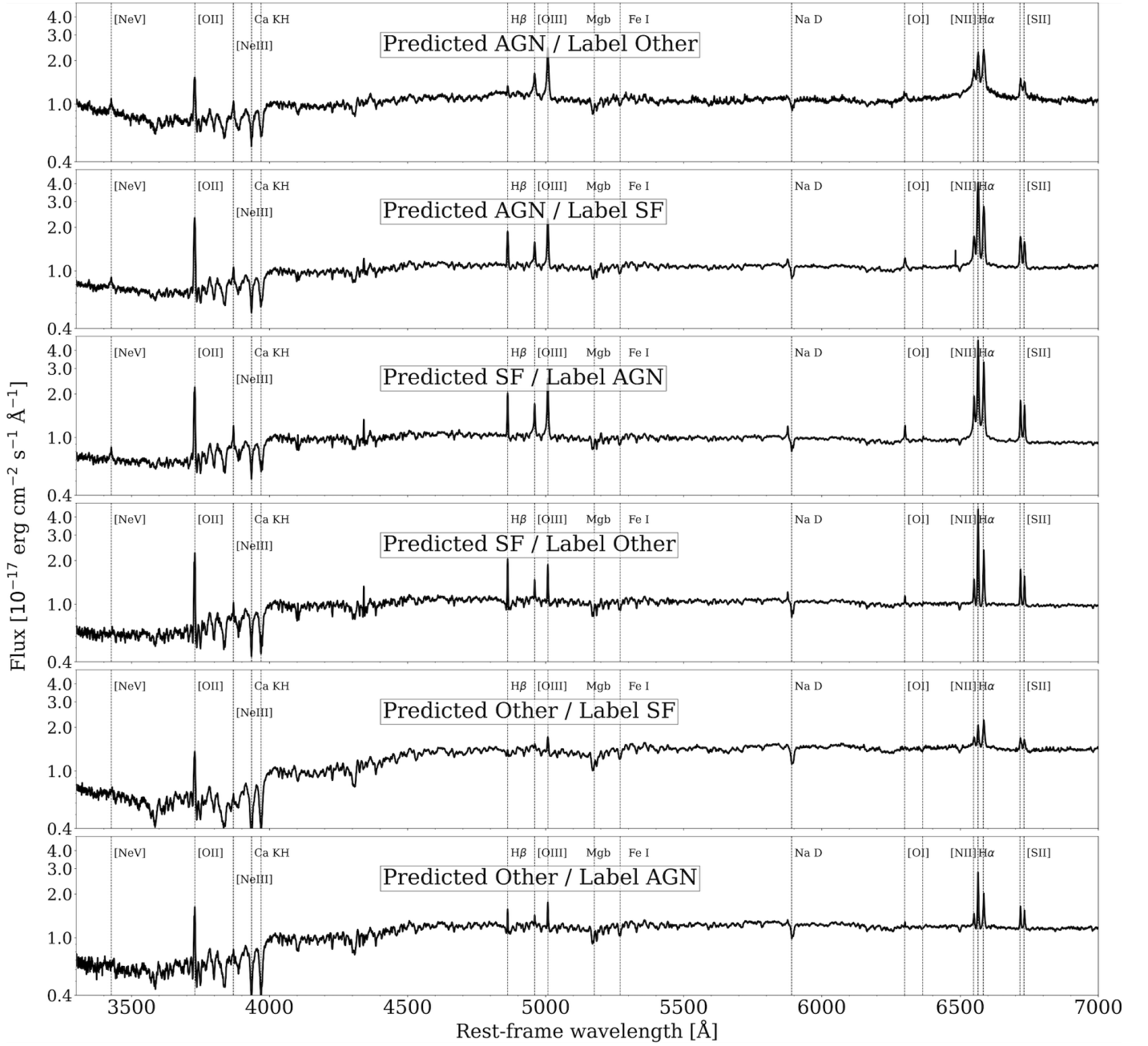


Fig. D.1: Stacked rest-frame optical spectra corresponding to the different sources with discrepant prediction–label combinations analyzed in this work. Key emission and absorption features are indicated with vertical dashed lines.

Therefore, despite their composite BPT classification, these sources are likely genuine AGNs. This interpretation is supported not only by the WHAN classification of the stacked spectrum, but also by the additional multi-diagnostic analysis presented in Section 4.4.2, where several independent indicators consistently favor an AGN interpretation.

D.5. Predicted as other/passive – labeled as star-forming

For sources originally labeled as star-forming and predicted as other/passive by our model, the stacked spectrum lies in the composite region of the [NII]–BPT diagram, while the WHAN diagram places it close to the strong-AGN regime. However, this stacked spectrum shows the weakest emission-line strengths

among all off-diagonal populations, indicating that the classification is intrinsically uncertain. This population likely represents a transition regime, where low-level ionization and star formation coexist, and small differences in line strength can lead to discrepant diagnostic classifications.

Appendix E: BL signatures in the stacked spectra

To illustrate the presence of broad emission-line components in the stacked spectra of galaxies predicted as AGNs but labeled as other/passive, we performed a multi-Gaussian decomposition using the PyQSOFit package (Guo et al. 2018). While the spectral measurements used throughout this work are obtained

Table D.1: Summary of the stacked spectra for sources with discrepant prediction–label combinations.

Original label	Predicted	BPT (stack)	WHAN (stack)	Notes
AGN	Other/passive	LINER	Weak AGN–Retired	Line ratios consistent with LINER-like ionization; low $H\alpha$ equivalent width places the stack close to the retired region in the WHAN diagram.
AGN	Star-forming	Seyfert (almost LINER/composite)	Strong AGN	Clear AGN signatures in both diagnostics despite the star-forming prediction; high equivalent width indicates ongoing nuclear activity.
Star-forming	Other/passive	Composite	Strong AGN (almost weak)	Composite location in the BPT diagram but classified as strong AGN in WHAN.
Star-forming	AGN	Composite	Strong AGN	Classified as composite by the BPT but robust AGN classification in WHAN, likely driven by high $H\alpha$ equivalent width.
Other/passive	AGN	Seyfert	Strong AGN	Clear Seyfert classification in both BPT and WHAN; the presence of a broad $H\alpha$ component confirms AGN activity.
Other/passive	Star-forming	Star-forming (almost composite)	Strong AGN (almost SF)	Located close to the star-forming/composite boundary; the WHAN classification suggests a transition regime between star formation and AGN activity.

Notes. For each stack, we report the original label, the model prediction, and the classification inferred from the BPT and WHAN diagnostic diagrams applied to the stacked spectrum.

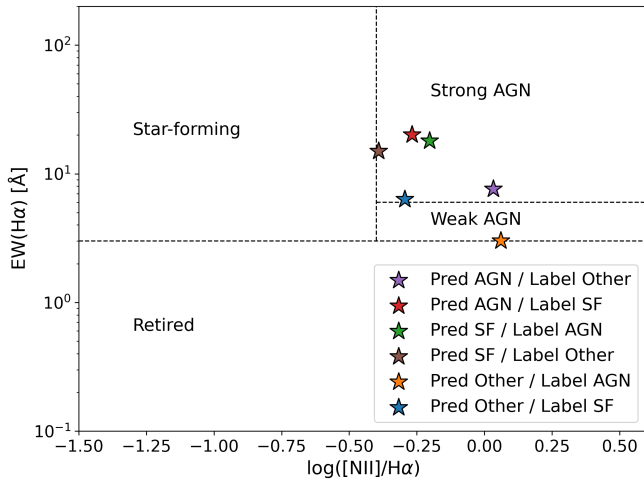


Fig. D.2: Location on the WHAN diagram of the stacked spectra corresponding to sources with discrepant classifications. Each symbol represents the emission-line measurements derived from the FastSpecFit model of the corresponding stacked spectrum. The classical demarcation lines separating star-forming, strong and weak AGN, and retired galaxies are shown for reference.

with FastSpecFit, which provides automated emission-line fits for the full sample, PyQSOFit is employed here solely to perform a more detailed decomposition of the stacked spectrum and to visually separate the narrow and broad components of the Balmer lines in Fig. E.1. PyQSOFit is specifically designed for AGN spectral modeling and allows simultaneous fitting of narrow and broad components, making it well suited for illustrating the line-profile structure in stacked spectra.

The inclusion of a broad component is required to reproduce the extended wings of the Balmer lines and leads to a significantly improved fit, as indicated by the reduced χ^2 values shown in the figure. The resulting broad component exceeds the typical BL AGN threshold of FWHM of $\sim 1200 \text{ km s}^{-1}$, consistent with the presence of BL emission in the stacked spectrum.

However, we do not attempt to derive precise kinematic measurements such as an exact FWHM or black hole masses from the stacked spectrum. Stacking spectra with different systemic velocities and intrinsic line widths can artificially broaden strong emission lines (such as $H\alpha$) while simultaneously suppressing the wings of weaker lines (such as $H\beta$), potentially biasing quantitative measurements. Therefore, the Gaussian decomposition is presented here only to illustrate the presence of broad emission, rather than to provide detailed BL region properties.

Appendix F: Validation of BL classifications with the DESI AGN/galaxy VAC

To further investigate discrepancies between predicted and reference BL classifications, we cross-matched these sources with the AGN/galaxy VAC of DESI DR1, which identifies BL sources using broader criteria than those adopted in this work, considering multiple emission lines ($H\alpha$, $H\beta$, and/or $MgII$) with FWHM $\geq 1200 \text{ km s}^{-1}$.

Table F.1 compares reference labels, model predictions, and the BL flag in the AGN/galaxy VAC. Among the 906 sources predicted as BL by our model but not labeled as such, 530 (58%) are identified as BL in the AGN/galaxy VAC. Similarly, of the 585 sources labeled as NL AGNs but predicted as BL, 364 (62%) are classified as BL in the AGN/galaxy VAC.

These discrepancies likely arise from the more restrictive BL identification criteria adopted in our labeling scheme,

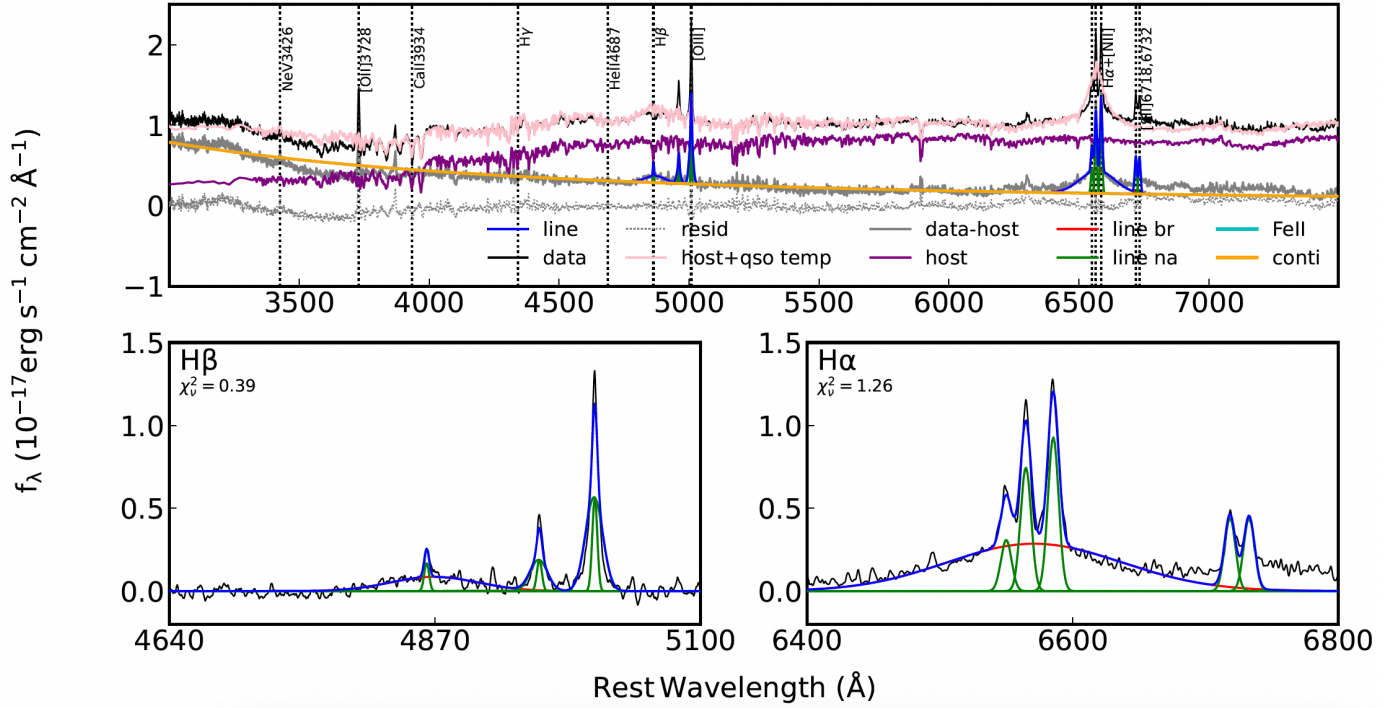


Fig. E.1: Top panel: stacked rest-frame optical spectrum of the 286 sources predicted as AGNs by our model but labeled as other/passive by traditional diagnostics. Bottom panels: zoom-in views of the $H\beta$ (left) and $H\alpha$ (right) regions, with the PyQSOFit model overplotted. The fits highlight the separation between narrow and broad emission-line components, providing independent evidence for the presence of BL emission in the stacked spectrum.

Table F.1: Comparison of reference labels, model predictions, and the BL flag in the DESI DR1 AGN/galaxy VAC.

Category	Number of Sources	Of which BL in VAC
Labeled as BL but not predicted as BL	867	648 (75%)
Not labeled as BL but predicted as BL	906	530 (59%)
Labeled as BL, predicted as NL AGN	643	505 (79%)
Labeled as NL AGN, predicted as BL	585	364 (62%)

Notes. The lower block highlights mismatches between the BL and NL AGN classifications.

which rely on the AON and S/N of the fitted broad and narrow components only of $H\alpha$ (see Section 2.4), whereas the AGN/galaxy VAC incorporates broader line-width criteria and multiple emission-line diagnostics. The high level of agreement with the AGN/galaxy VAC therefore supports the ability of our model to recover BL sources beyond the restrictive definitions adopted in the reference labeling.