
Unsupervised Discovery of High-Redshift Galaxy Populations with Variational Autoencoders

Aayush Saxena

Department of Physics

University of Oxford

Denys Wilkinson building, Oxford, OX1 3RH, UK

aayush.saxena@physics.ox.ac.uk

Abstract

We apply variational autoencoders to automatically discover galaxy populations using publicly available high-redshift *JWST* spectra without prior classification knowledge. Our unsupervised method identifies distinct astrophysical classes of unique and exciting galaxy types, demonstrating automated discovery capabilities for large spectroscopic surveys.

1 Introduction

Since its launch on Christmas Day in 2021, the *James Webb Space Telescope (JWST)* has rapidly been transforming our understanding of how the first galaxies form and evolve a few hundred million years after the Big Bang, when the Universe was a fraction of its current age of 13.6 billion years. This has mainly been achieved thanks to *JWST*'s imaging and spectroscopic capabilities at near-infrared wavelengths, which trace the redshifted ultraviolet and optical light from distant galaxies.

Spectroscopy of galaxies is one of the most important tools to analyze their physical and chemical properties and test theories of galaxy formation and evolution. Particularly for distant, high redshift galaxies, the shape of the continuum emission from stars, supermassive black holes and hot gas, and emission lines from nebular regions, can give insights into the dominant sources of photoionization, the physical (temperature, density) and chemical (level of enrichment from elements heavier than Hydrogen and Helium) properties of the stellar populations and the interstellar gas, and their cosmic dust content, which are all vital building blocks of galaxies.

Early *JWST* spectroscopic results have shed new light on several outstanding questions in the field, such as when and how did the first stars and supermassive black holes form in early galaxies, how are the first generation of stars different from evolved stars in our Milky Way, and what is the impact of the radiation emitted by the earliest galaxies on the cosmological evolution of the intergalactic medium. These early results have often relied on the identification of small samples of interesting galaxies ‘by eye’ from individual large observing programs. With over three years of scientific operations and a growing repository of publicly available spectroscopic data from several large and treasury observing campaigns, the need of the hour is to assemble statistically significant samples of the most interesting galaxies to better inform models of galaxy evolution.

Machine learning approaches, particularly unsupervised deep learning models, are perfectly suited to automate the discovery and classification of astrophysical objects at scale. Variational Autoencoders (VAEs; [5]) can be powerful when applied to spectroscopic data analysis as they learn compact, interpretable representations from intrinsically complex datasets, while enabling both reconstruction and generation of synthetic data. Unlike supervised methods that require a large repository of labeled data, VAEs are capable of discovering structure in the latent space in an unsupervised fashion. VAEs have previously been applied to large spectroscopic datasets of nearby, low redshift galaxies taken

from ground-based telescopes [e.g. 12, 1, 15, 11], but have never been deployed in the context of high redshift galaxy spectra from *JWST*, which represents a potent discovery space.

In this work, we leverage VAEs and latent space clustering to discover and characterize statistically significant samples of rare and exciting high redshift ($z > 4$) galaxies, probing the first 1.5 billion years of the Universe’s evolution. In Section 2 we describe the implementation of the VAE, in Section 3 we present our main results, in Section 4 we discuss future directions for this work, and in Section 5. We have made the datasets and code publicly available¹ in the spirit of reproducibility and open science.

2 Methods

2.1 Variational autoencoder (VAE) architecture

In this work, we implement a Variational Autoencoder (VAE) following the framework developed by Kingma and Welling [5]. The VAE learns a probabilistic mapping between high-dimensional astronomical spectroscopic data and a lower-dimensional latent space. The VAE optimizes the Evidence Lower Bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \quad (1)$$

where the first term represents the reconstruction accuracy and the second enforces regularization towards a prior distribution, typically a Gaussian, $p(z) = \mathcal{N}(0, I)$ via calculation of the Kullback-Leibler (KL) divergence. The encoder neural network, $q_\phi(z|x)$, learns to map the input spectra, $x \in \mathbb{R}^d$ to latent parameters, $(\mu, \sigma^2) \in \mathbb{R}^{2k}$, where d is the input dimension and k is the latent space dimension. The decoder neural network, $p_\theta(x|z)$ reconstructs the original spectra from latent variables, $z \in \mathbb{R}^k$ sampled via the reparameterization trick [5].

For the encoder and the decoder, we employ a deep symmetric neural network architecture with four fully-connected layers. For the encoder, the layers progressively compress the dimensionality, d of the input spectra: $d \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow k$, where $k = 16$ is the dimensionality of the latent vectors chosen in our implementation. The choice of $k = 16$ balances expressiveness of the neural network with computational efficiency. The decoder architecture mirrors that of the encoder, expanding the latent space back to the dimensionality of the input spectra: $k \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow d$.

We implement a range of regularization techniques to prevent overfitting and improve generalization: (i) we apply L2-weight regularization with $\lambda = 0.001$ to all hidden layers; (ii) we apply batch normalization after each dense layer to stabilize training; (iii) we deploy dropout layers in the encoder network with rates decreasing from 0.2 to 0.1 towards the latent bottleneck (and vice-versa in the decoder); and (iv) we apply gradient clipping for training stability. For spectral data containing missing/masked inputs, we implement a masked reconstruction loss:

$$L_{rec} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^d M_{ij} (x_{ij} - \hat{x}_{ij})^2 \quad (2)$$

where M_{ij} is a binary mask that excludes missing spectral data.

To train the networks, we implement exponentially decaying learning rates starting at 10^{-4} with a decay rate of 0.95 every 500 steps, combined with early stopping if the validation reconstruction loss stops improving after 50 steps. The training set contained 85% of the data and the validation set contained 15% of the data.

2.2 Data pre-processing

The spectroscopic data are taken from the DAWN *JWST* Archive² (DJA), which is a repository containing nearly all publicly available *JWST* datasets. DJA further provides redshift information inferred from the galaxy spectra along with quality flags. In this work we only use redshifts with

¹<https://github.com/aayush3009/learnspec>

²<https://dawn-cph.github.io/dja/index.html>

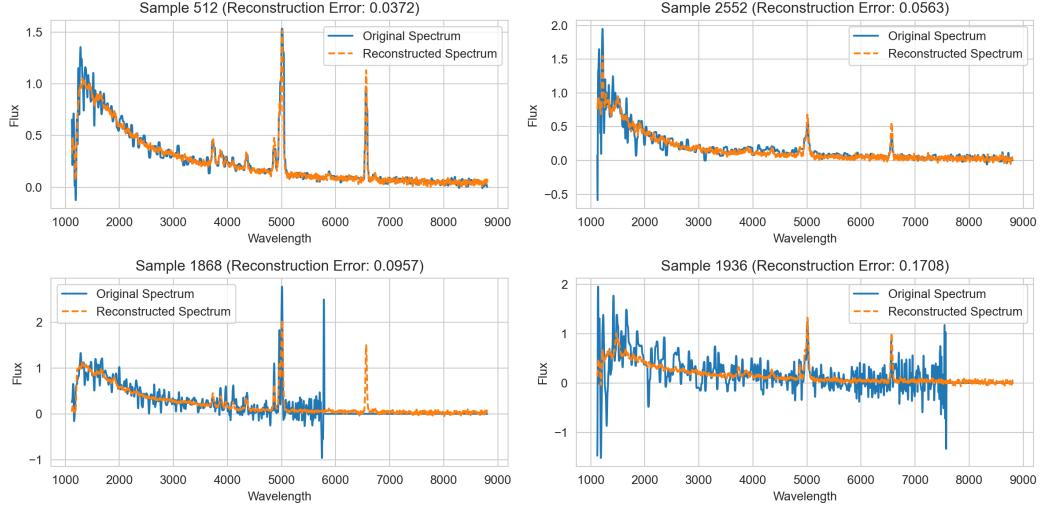


Figure 1: Comparisons of input (blue) and reconstructed (orange) spectra, drawn from four quartiles of reconstruction errors distribution, with decreasing accuracy clockwise from top-left. We note that the reconstruction often makes predictions for when the input data is missing/masked.

the highest quality flag, considering only sources with redshifts above 4 ($z > 4$; tracing the first 1.5 billion years after the Big Bang). There are 2743 objects in our final dataset.

A number of pre-processing steps must be applied to prepare the datasets before feeding them into the VAE. The first step involves de-redshifting the spectroscopic data to resample the spectra into the rest-frame wavelength ($\lambda_{\text{rest}} = \lambda_{\text{obs}}/(1+z)$). A uniformly spaced rest-wavelength grid was determined based on the median redshift of the sample. Since the observed wavelength range of *JWST/NIRSpec* is fixed from $\sim 7500 - 53000$ Å, spectra at different redshifts will sample different rest-frame wavelength ranges. Any resulting missing spectroscopic flux on the common rest-frame wavelength grid was masked.

Each de-redshifted spectrum was then normalized by scaling its continuum flux at rest-frame 1500 Å to 1.0. Spectroscopic data contain strong nebular emission lines and lower flux continuum tracing starlight. Therefore, to properly leverage the discerning power within the dynamic range of fluxes after normalization, we used a novel arcsinh transformation ($\text{arcsinh}(x) = \ln(x + \sqrt{x^2 + 1})$), which is approximately linear for small values of x (continuum), and log for large values of x (emission lines). This helps preserve information from both the continuum shape and emission lines, which are important features for galaxy classification task at hand.

3 Results

3.1 Reconstruction accuracy

Our VAE model performs excellently at reconstructing the vast majority of spectra. The reconstruction error distribution measured as the mean squared error (MSE) between the original and reconstructed spectra has a median value of 0.122 and is one-sided, long-tailed Gaussian with a standard deviation of 0.124. In Figure 1 we show representative examples from four quartiles of the error distribution. The VAE is able to reconstruct masked/missing flux that often leads to higher reconstruction errors. High-error reconstructions ($\text{MSE} > 0.1$) typically trace noisy spectra that contain artifacts, or spectra with extremely faint continua.

3.2 Clustering in latent space

To identify interesting galaxy types from the latent space while breaking the ‘curse of dimensionality’, we collapse the 16D latent space to a 2D representation using UMAP dimensionality reduction [8]. We apply Gaussian Mixture Modeling with increasing number of components in the range [5, 15] to the 2D embeddings 100 times and record the clustering solution that returns the maximum Silhouette

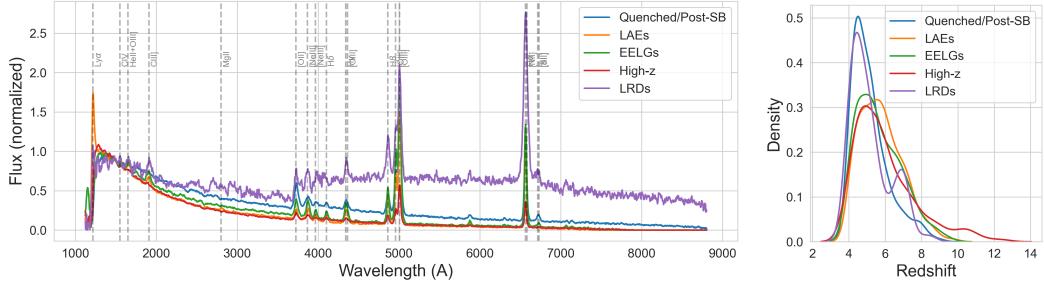


Figure 2: *Left:* Observed, median-combined spectra of five exciting high redshift galaxy types identified using our VAE and clustering approach. The diversity of the galaxy spectra demonstrates the various physical processes that shape the continua and emission lines, enabling insights into galaxy evolution. *Right:* The redshift distribution of the clusters indicating a redshift correlation between galaxy types identified naturally by the VAE.

score [13]. With a Silhouette score of 0.44, we identify 12 well-separated clusters that include clusters containing noisy and artifact-dominated data. The number of galaxies across clusters ranges from 63 to 334 with no clear dominant class, demonstrating that our model is capable of capturing diverse galaxy populations across redshifts.

3.3 Astrophysical insights

To explore which kinds of astrophysical sources have been clustered together, we create a median spectrum for each cluster, and then compare the resulting properties of the galaxies in each cluster with known galaxy types in the literature, tracing rare and unique phenomena in the early Universe. These labels are effectively assigned using prior knowledge of the expected spectroscopic properties of known distant galaxies, which up until now have largely been visually classified. In this work, we focus on five exciting classes of objects and briefly describe the importance of each galaxy type below:

Quenched/Post-starburst (SB) galaxies: We identify 326 galaxies that can be classified as being in their (mini-) quenched or post-starburst phase [e.g. 6], which is when a galaxy has recently undergone a burst of star-formation and is currently in its ‘lull’ phase. Above a redshift of 4, our newly discovered sample nearly doubles the number of known such galaxies.

Lyman- α Emitters (LAEs): Characterized by their strong Lyman- α emission at rest-frame wavelength of 1216 Å, LAEs trace intense star-formation. At $z > 6$, a strong Lyman- α line emerges from regions of the Universe that have been ‘reionized’ due to UV photons from young stars, charting the phase transition of the intergalactic medium from a completely neutral to ionized state within a billion years after the Big Bang [e.g. 14]. We identify 213 strong LAEs, doubling the number of LAEs currently known at $z > 4$ [16, 4].

Extreme Emission Line Galaxies (EELGs): These galaxies are characterized by extremely strong emission lines, tracing some of the highest star-formation rates in the Universe, driven by young, massive stars forming in short bursts. We identify 180 EELGs, more than doubling the number of such galaxies currently known at these redshifts [e.g. 2].

High-redshift (High-z): Galaxies at the highest redshifts trace galaxy formation immediately after the Big Bang. These galaxies typically lack heavier elements as evidenced by the weaker emission lines in their spectra [e.g. 3, 10]. With individual spectra lacking significant signal to enable a robust analysis of the underlying stars and gas, our identification of 320 sources that exhibit properties similar to some of the first galaxies, including some of the highest redshift galaxies in our parent sample, significantly expands the sample statistics for studying their properties in detail.

Little Red Dots (LRDs): An exciting discovery made using *JWST* has been that of a handful of so-called Little Red Dots, which are extremely compact galaxies with a puzzling ‘V-shaped’ continuum and strong emission lines, tracing both star-formation and supermassive black hole activity [7]. Current models are unable to self-consistently explain the observed properties of LRDs without invoking exotic astrophysical phenomena [e.g. 9]. Our new sample of 142 LRDs will enable detailed spectroscopic analyses along with robust model comparison for these puzzling objects.

The median-combined original input spectra for these classes of objects are shown in the left-hand panel of Figure 2, with the right-hand panel showing the redshift distribution of each class. Although the signal-to-noise ratio of individual galaxies that make up these combined spectra vary, it is clear from the unweighted median-combined spectra that individual galaxies within each cluster exhibit highly correlated spectral shapes and properties.

4 Future Work

Since the original draft of this paper, the number of galaxy spectra available in public archives has grown substantially. The logical next step for the framework introduced here would be to retrain the model on these larger datasets and re-identify clusters of interesting galaxy types. Larger datasets also increase the probability of finding truly anomalous galaxy spectra, thereby expanding the discovery space. Additionally, ‘truth’ labels assigned from smaller training samples can be leveraged to further automate the isolation of these interesting galaxy types from larger datasets.

Further improvements could be made to the clustering methodology. At present, we identify clusters using GMMs in the collapsed 2D UMAP representation of the 16D latent space. Experiments could be performed on the performance of clustering directly in the latent space, as well as by implementing other clustering algorithms (such as DBSCAN, OPTICS, or hierarchical clustering) on the UMAP representation. Additionally, given the nature of the input data, there are likely to be degeneracies between the latent space parameters as the same galaxy spectrum could theoretically belong to multiple classes of known objects. Exploring these degeneracies further could help make the identification of clusters more robust.

Inclusion of multi-modal data, such as 2D imaging (or photometry) from *JWST*, or accompanying spectra with higher spectral resolution (albeit with over limited wavelength ranges) capable of resolving finer spectroscopic features could add significant value to the clustering power by adding focus on additional important galaxy features. This would require changes to the VAE architecture to account for the increased complexity of the input data.

5 Conclusions

In this work, we have leveraged a Variational Autoencoder (VAE) architecture combined with clustering algorithms deployed on 2D representations of the learned latent space to identify unique and exciting classes of distant galaxies from publicly available *JWST* spectroscopic data. Our approach has yielded significantly increased the number of objects belonging to known classes of interesting galaxies tracing unique physical phenomena in the early Universe in a highly automatic fashion. Increased samples of distant galaxies with interesting physical properties are desperately needed to test and refine theories of star and black hole formation in some of the first galaxies that formed after the Big Bang.

With publicly available *JWST* spectroscopic datasets steadily growing, our model architecture and implementation enables training and deployment at scale, providing a vital tool for astronomers to automate the identification of known galaxy types from large datasets in addition to discovering unknown, anomalous spectra that may trace new astrophysical phenomena. Our model can potentially be integrated into existing *JWST* spectroscopic data pipelines and repositories to rapidly speed up automatic classification of interesting and/or anomalous galaxy spectra. Our input (processed) dataset and code is publicly available³

References

- [1] Böhm, V., Kim, A. G., and Juneau, S. (2023). Fast and efficient identification of anomalous galaxy spectra with neural density estimation. *Monthly Notices of the Royal Astronomical Society*, 526(2):3072–3087.
- [2] Boyett, K., Bunker, A. J., Curtis-Lake, E., Chevallard, J., Cameron, A. J., Jones, G. C., Saxena, A., Charlot, S., Curti, M., Wallace, I. E. B., Arribas, S., Carniani, S., Willott, C., Alberts, S., Eisenstein, D. J., Hainline, K., Hausen, R., Johnson, B. D., Rieke, M., Robertson, B., Stark, D. P.,

³<https://github.com/aayush3009/learnspec>

- Tacchella, S., Williams, C. C., Chen, Z., Egami, E., Endsley, R., Kumari, N., Laseter, I., Looser, T. J., Maseda, M. V., Scholtz, J., Shvarei, I., Simmonds, C., Smit, R., Übler, H., and Witstok, J. (2024). Extreme emission line galaxies detected in JADES JWST/NIRSpec - I. Inferred galaxy properties. *Monthly Notices of the Royal Astronomical Society*, 535(2):1796–1828.
- [3] Carniani, S., Hainline, K., D'Eugenio, F., Eisenstein, D. J., Jakobsen, P., Witstok, J., Johnson, B. D., Chevallard, J., Maiolino, R., Helton, J. M., Willott, C., Robertson, B., Alberts, S., Arribas, S., Baker, W. M., Bhatawdekar, R., Boyett, K., Bunker, A. J., Cameron, A. J., Cargile, P. A., Charlot, S., Curti, M., Curtis-Lake, E., Egami, E., Giardino, G., Isaak, K., Ji, Z., Jones, G. C., Kumari, N., Maseda, M. V., Parlanti, E., Pérez-González, P. G., Rawle, T., Rieke, G., Rieke, M., Del Pino, B. R., Saxena, A., Scholtz, J., Smit, R., Sun, F., Tacchella, S., Übler, H., Venturi, G., Williams, C. C., and Willmer, C. N. A. (2024). Spectroscopic confirmation of two luminous galaxies at a redshift of 14. *Nature*, 633(8029):318–322.
- [4] Jones, G. C., Bunker, A. J., Saxena, A., Arribas, S., Bhatawdekar, R., Boyett, K., Cameron, A. J., Carniani, S., Charlot, S., Curtis-Lake, E., Hainline, K., Johnson, B. D., Kumari, N., Maseda, M. V., Rix, H.-W., Robertson, B. E., Tacchella, S., Übler, H., Williams, C. C., Willott, C., Witstok, J., and Zhu, Y. (2025). JADES: measuring reionization properties using Lyman-alpha emission. *Monthly Notices of the Royal Astronomical Society*, 536(3):2355–2380.
- [5] Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114.
- [6] Looser, T. J., D'Eugenio, F., Maiolino, R., Tacchella, S., Curti, M., Arribas, S., Baker, W. M., Baum, S., Bonaventura, N., Boyett, K., Bunker, A. J., Carniani, S., Charlot, S., Chevallard, J., Curtis-Lake, E., Lola Danhaive, A., Eisenstein, D. J., de Graaff, A., Hainline, K., Ji, Z., Johnson, B. D., Kumari, N., Nelson, E., Parlanti, E., Rix, H.-W., Robertson, B., Del Pino, B. R., Sandles, L., Scholtz, J., Smit, R., Stark, D. P., Übler, H., Williams, C. C., Willott, C., and Witstok, J. (2025). JADES: Differing assembly histories of galaxies: Observational evidence for bursty star formation histories and (mini-)quenching in the first billion years of the Universe. *Astronomy & Astrophysics*, 697:A88.
- [7] Matthee, J., Naidu, R. P., Brammer, G., Chisholm, J., Eilers, A.-C., Goulding, A., Greene, J., Kashino, D., Labbe, I., Lilly, S. J., Mackenzie, R., Oesch, P. A., Weibel, A., Wuyts, S., Xiao, M., Bordoloi, R., Bouwens, R., van Dokkum, P., Illingworth, G., Kramarenko, I., Maseda, M. V., Mason, C., Meyer, R. A., Nelson, E. J., Reddy, N. A., Shvarei, I., Simcoe, R. A., and Yue, M. (2024). Little Red Dots: An Abundant Population of Faint Active Galactic Nuclei at $z \sim 5$ Revealed by the EIGER and FRESCO JWST Surveys. *The Astrophysical Journal*, 963(2):129.
- [8] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, page arXiv:1802.03426.
- [9] Naidu, R. P., Matthee, J., Katz, H., de Graaff, A., Oesch, P., Smith, A., Greene, J. E., Brammer, G., Weibel, A., Hvizing, R., Chisholm, J., Labb'e, I., Simcoe, R. A., Witten, C., Atek, H., Baggen, J. F. W., Belli, S., Bezanson, R., Boogaard, L. A., Bose, S., Covelo-Paz, A., Dayal, P., Fudamoto, Y., Furtak, L. J., Giovinazzo, E., Goulding, A., Gronke, M., Heintz, K. E., Hirschmann, M., Illingworth, G., Inoue, A. K., Johnson, B. D., Leja, J., Leonova, E., McConachie, I., Maseda, M. V., Natarajan, P., Nelson, E., Setton, D. J., Shvarei, I., Sobral, D., Stefanon, M., Tacchella, S., Toft, S., Torralba, A., van Dokkum, P., van der Wel, A., Volonteri, M., Walter, F., Wang, B., and Watson, D. (2025a). A “Black Hole Star” Reveals the Remarkable Gas-Enshrouded Hearts of the Little Red Dots. *arXiv e-prints*, page arXiv:2503.16596.
- [10] Naidu, R. P., Oesch, P. A., Brammer, G., Weibel, A., Li, Y., Matthee, J., Chisholm, J., Pollock, C. L., Heintz, K. E., Johnson, B. D., Shen, X., Hvizing, R. E., Leja, J., Tacchella, S., Ganguly, A., Witten, C., Atek, H., Belli, S., Bose, S., Bouwens, R., Dayal, P., Decarli, R., de Graaff, A., Fudamoto, Y., Giovinazzo, E., Greene, J. E., Illingworth, G., Inoue, A. K., Kane, S. G., Labbe, I., Leonova, E., Marques-Chaves, R., Meyer, R. A., Nelson, E. J., Roberts-Borsani, G., Schaerer, D., Simcoe, R. A., Stefanon, M., Sugahara, Y., Toft, S., van der Wel, A., van Dokkum, P., Walter, F., Watson, D., Weaver, J. R., and Whitaker, K. E. (2025b). A Cosmic Miracle: A Remarkably Luminous Galaxy at $z_{\text{spec}} = 14.44$ Confirmed with JWST. *arXiv e-prints*, page arXiv:2505.11263.

- [11] Nicolaou, C., Nathan, R. P., Lahav, O., Palmese, A., Saintonge, A., Aguilar, J., Ahlen, S., Allende Prieto, C., Bailey, S., BenZvi, S., Bianchi, D., Brodzeller, A., Brooks, D., Claybaugh, T., de la Macorra, A., Della Costa, J., Dey, A., Doel, P., Forero-Romero, J. E., Gaztañaga, E., Gontcho, S. G. A., Gutierrez, G., Honscheid, K., Howlett, C., Ishak, M., Kehoe, R., Kirkby, D., Kisner, T., Kremin, A., Lambert, A., Landriau, M., Le Guillou, L., Meisner, A., Miquel, R., Moustakas, J., Nadathur, S., Prada, F., Pérez-Ràfols, I., Rossi, G., Sanchez, E., Schubnell, M., Siudek, M., Sprayberry, D., Tarlé, G., Weaver, B. A., and Zou, H. (2025). Identifying Anomalous DESI Galaxy Spectra with a Variational Autoencoder. *arXiv e-prints*, page arXiv:2506.17376.
- [12] Portillo, S. K. N., Parejko, J. K., Vergara, J. R., and Connolly, A. J. (2020). Dimensionality Reduction of SDSS Spectra with Variational Autoencoders. *The Astronomical Journal*, 160(1):45.
- [13] Rousseeuw, P. (1987). Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *comput. appl. math.* 20, 53–65. *Journal of Computational and Applied Mathematics*, 20:53–65.
- [14] Saxena, A., Bunker, A. J., Jones, G. C., Stark, D. P., Cameron, A. J., Witstok, J., Arribas, S., Baker, W. M., Baum, S., Bhatawdekar, R., Bowler, R., Boyett, K., Carniani, S., Charlot, S., Chevallard, J., Curti, M., Curtis-Lake, E., Eisenstein, D. J., Endsley, R., Hainline, K., Helton, J. M., Johnson, B. D., Kumari, N., Looser, T. J., Maiolino, R., Rieke, M., Rix, H.-W., Robertson, B. E., Sandles, L., Simmonds, C., Smit, R., Tacchella, S., Williams, C. C., Willmer, C. N. A., and Willott, C. (2024). JADES: The production and escape of ionizing photons from faint Lyman-alpha emitters in the epoch of reionization. *Astronomy & Astrophysics*, 684:A84.
- [15] Scourfield, M., Saintonge, A., de Mijolla, D., and Viti, S. (2023). De-noising of galaxy optical spectra with autoencoders. *Monthly Notices of the Royal Astronomical Society*, 526(2):3037–3050.
- [16] Tang, M., Stark, D. P., Topping, M. W., Mason, C., and Ellis, R. S. (2024). JWST/NIRSpec Observations of Lyman α Emission in Star-forming Galaxies at $6.5 \lesssim z \lesssim 13$. *The Astrophysical Journal*, 975(2):208.