Extracting an Informative Latent Representation of High-Dimensional Galaxy Spectra

Daiki Iwasaki[1], **Suchetha Cooray**[2], **and Tsutomu T. Takeuchi**[1,3]
[1] Nagoya University
[2] National Astronomical Observatory of Japan
[3] Institute of Statistical Mathematics

Abstract

To understand the fundamental parameters of galaxy evolution, we investigated the minimum set of parameters that explain the observed galaxy spectra in the local Universe. We identified four latent variables that efficiently represent the diversity of high-dimensional galaxy spectral energy distributions (SEDs) observed by the Sloan Digital Sky Survey. Additionally, we constructed meaningful latent representation using conditional variational autoencoders trained with different permutations of galaxy physical properties, which helped us quantify the information that these traditionally used properties have on the reconstruction of galaxy spectra. The four parameters suggest a view that complex SED population models with a very large number of parameters will be difficult to constrain even with spectroscopic galaxy data. Through an Explainable AI (XAI) method, we found that the region below 5000Å and prominent emission lines ([O II], [O III], and Hα) are particularly informative for predicting the latent variables. Our findings suggest that these latent variables provide a more efficient and fundamental representation of galaxy spectra than conventionally considered galaxy physical properties.

1 Introduction

A galaxy's spectral energy distribution (SED), which reflects the multi-wavelength flux observed of a galaxy, is often the only path to study the astrophysical processes within them. Instruments like the current Dark Energy Spectroscopic Instrument (DESI), upcoming Prime-Focus Spectrograph (PFS) instrument, and integral field units (IFUs) like the Multi Unit Spectroscopic Explorer (MUSE), obtain high-dimensional data for sometimes millions of galaxies, posing computational complexity challenges and issues inherent to high-dimensional analysis, such as overfitting and multicollinearity. Traditionally, we have employed various methods that make high-dimensional data manageable, such as photometry and color-magnitude diagrams [e.g., 4] and the BPT diagram [3], to characterize galaxy spectra based on their representative low dimensional features. However, these low-dimensional representations do not contain the complete spectral information. Hence, we use neural networks to extract full information from high-dimensional galaxy spectroscopic data. In doing so, this work addresses three main scientific questions: 1. How many parameters are required for adequate representation of galaxy SEDs? 2. What fundamental physical properties explain the observed SEDs? 3. Which spectral ranges are the most informative for representing observed SEDs?

This work uses the Variational Autoencoder [VAE; 8], a type of neural network to compress galaxy spectra into meaningful, lower-dimensional latent parameters without relying on explicit labels or prior assumptions. To identify the most informative physical properties in reconstructing galaxy spectra, we employ the Conditional VAE [CVAE; 7], incorporating conditional properties such as stellar mass (M_*), star formation rate (SFR), specific SFR (sSFR = SFR/M_*), and metallicity (Z). We also utilize the XAI method named SHAP [SHapley Additive exPlanations; 10] to assign importance

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

values to specific spectral features in galaxy SEDs. This enables us to identify key spectral ranges that influence the VAE's latent variables, providing insight into the fundamental characteristics of galaxies.

2 Data and Methodology

The primary dataset for this analysis is the galaxy spectroscopic data from the Sloan Digital Sky Survey [SDSS; 1]. Physical properties were derived from the GALEX-SDSS-WISE Legacy Catalog [GSWLC; 17], which used UV+optical+mid-IR SED fitting. Our final dataset includes SDSS spectra cross-matched with the GSWLC. We limit to galaxies with a redshift < 0.1 to ensure consistency in the spectral range across our dataset. For preprocessing spectra, we used the Python module `spectres` to shift each galaxy spectrum to its rest frame and to resample it to 4000 logarithmically spaced wavelength pixels within the range of 3400Å–8400Å. A similar approach was employed in previous studies [e.g., 16, 13]. We also used an iterative principal component analysis (PCA) method to handle bad data points, following a similar approach used by [21]. The dataset utilized for this study comprises a total of about 320,000 galaxy spectra.

We employ a VAE and its variant CVAE for this analysis. VAE is a two-component architecture, an encoder and a decoder, where the encoder maps high-dimensional data into a usually lower-dimensional latent space, and the decoder reconstructs the original input from lower-dimensional latent variables. The VAE loss function is the sum of two components: the reconstruction error, which is the distance metric between the input data and the reconstructed output and Kullback-Leibler divergence [9], which encourages the latent variables to usually follow a standard Gaussian distribution, $N(0, 1)$. The VAE achieves a continuous lower-dimensional latent representation of the complex, high-dimensional data by optimizing this composite loss function(refer to the Appendix A). CVAE integrates conditional data during the encoding and decoding processes, enabling more controlled output generation while retaining the same loss function as the VAE. By utilizing physical properties as conditional data, we obtain latent representations that remain unaffected by them.

The deep neural network model, often seen as a "black box," poses challenges in understanding the underlying reasons for its predictions. To address this problem, we utilized SHAP values, which quantify the influence of each feature on the model's predictions, thereby elucidating the impact of different inputs on the outcome (refer to the Appendix B). All scripts for downloading SDSS spectra, preprocessing data, training, and generating the figures are accessible here.

3 Results

The primary results of our study indicate that four latent variables can effectively represent the 320k, 4000-dimensional galaxy spectra, showing strong correlations between the observed diversity of galaxy physical properties. These latent parameters also show correlations with traditional galaxy physical properties such as stellar masses, star formation rates, specific star formation rates, and metallicity, as shown in Figure 1.

We use the Bayesian information criterion [BIC; 18] to identify the optimal number of parameters for spectral reconstruction. The BIC is a criterion used for model selection among models with varying permutations of conditional parameters. BIC is defined as $\text{BIC} = k\log(n) - 2\log(L)$ where k is the total number of parameters and L is the likelihood. BIC, derived from Bayesian principles, aims to identify the 'true' model among the candidates. This approach is distinct from other criteria based on information entropy [2]. Figure 2 presents the BIC values as a function of the total number of parameters (latent variables + conditional data). We successfully quantified the extent to which physical properties influence the reconstruction of galaxy spectra and the capture of an informative latent representation. Our results indicate that the VAE with four parameters is the best model. The fact that the VAE outperforms the CVAE suggests that the latent variables obtained provide a more efficient representation for characterizing galaxy spectra distribution compared to traditional galaxy physical properties. We also found that increasing the number of conditional parameters does not improve the results significantly. Moreover, when excluding the VAE, the CVAE model conditioned with M_* and the model conditioned with SFR performs well under our metric. The former yielded a result of 945.67, while the latter scored 945.14, indicating a difference of just approximately 0.5. The standard deviation of the BIC values for models with five parameters is 8.83, so this 0.5 difference is

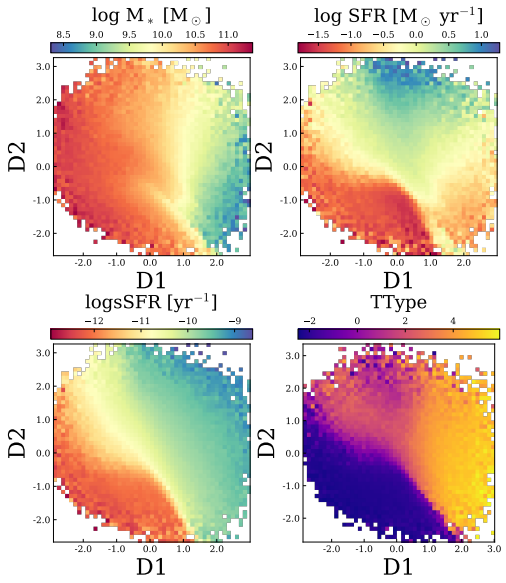

Figure 1: Distribution of average stellar masses, SFR, sSFR, and TType in the latent space. The figures from the top right to the bottom left display the distributions of M_*, SFR, sSFR, and TType in the latent space. Median values for each property are calculated within their respective bins. The morphological properties are derived from 5.

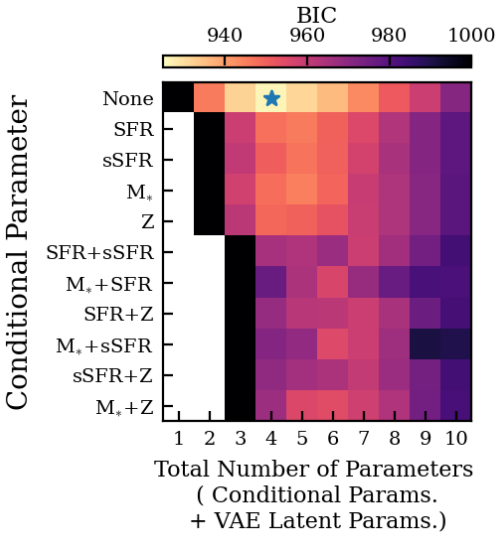

Figure 2: BIC values for combinations of latent variables and conditional data. BIC values are presented as a function of model complexity, with total parameters on the x-axis and conditional data on the y-axis. The color gradient depicts BIC values: darker shades represent higher values, and lighter shades indicate lower ones. A model with a lower BIC is typically a better fit. The blue star marks the model that achieves the optimal balance between fit and complexity.

not very significant. This result is equivalent to what would be obtained by determining the Evidence Lower Bound with an added KL divergence loss term.

4 Discussion

4.1 How many parameters are required for adequate representation of galaxy SEDs?

The latent variables D_1 to D_4 are PCA-transformed values of the VAE latent features. PCA is a lossless transformation that preserves information because it does not alter the number of parameters [14]. In the original VAE latent space, axes are arbitrary and may not correspond to meaningful directions. By applying PCA tranformation to the latent variables, the latent axes are the most informative. A similar approach was done in previous studies [16, 13]. D_1 to D_4 capture 34%, 32%, 16%, and 18% of the total variance, respectively.

Figure 3, shows the effect of changing the latent parameters on the reconstructed spectra. We used Mutual Information (MI) values to understand the link between latent variables and physical properties (Figure 4). As seen in Figure 3 Panel A, D_1 impacts the 4000Å break, signifying evolving stellar populations along D_1. Higher D_1 aligns with star-forming galaxies, while lower values indicate older, quiescent galaxies. D_1 also has the highest correlation with M_*, then sSFR, and lastly, metallicity. This reflects a trend from lower-mass to massive galaxies along D_1 as seen in Figure 1. Modifying D_2 (Figure 3, Panel B) alters the spectral intensity, mainly below 5000Å, while increasing emission line strength to around 6000Å. Figure 4 shows D_2 exhibits a strong correlation with SFR. D_1 and D_2 account for roughly 66% of the total variance. The low MI values between D_1 and SFR, and D2 and $M*$, combined with improved reconstructions using SFR and M as data, emphasize the importance of $M*$ and SFR in describing galaxy spectra. These patterns remained consistent across different number of latent variables settings, underlining the key roles of D_1 with $M*$ and D_2 with SFR in galaxy spectrum characterization.

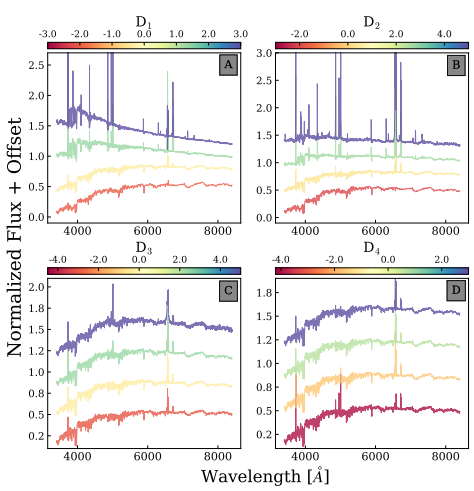

Figure 3: Effect of single latent variable change on reconstructed spectra. With all other parameters held at zero, the figure demonstrates the influence of varying just one latent variable on the generated spectra. The four panels, from A to D, correspond to different latent variables: D_1 through D_4.

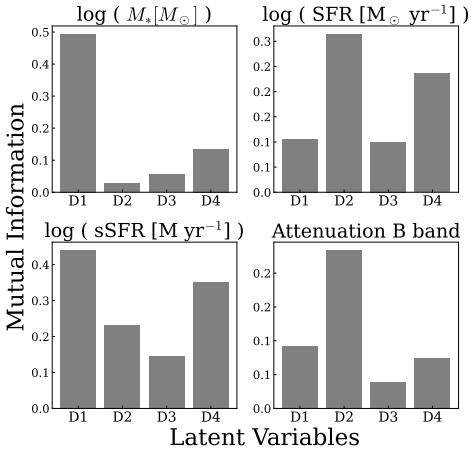

Figure 4: Mutual information between parameters and latent variables (D1 to D4). These eight bar plots depict the MI between parameters and latent variables (D1 to D4). Parameters include M_*, SFR, sSFR, Metallicity, exponential τ, Velocity Dispersion, Attenuation in B band, and V band, ordered top left to bottom right. The y-axis signifies MI value, and the x-axis presents the four latent variables.

Figure 3 (Figure 3, Panel C and D) displays how D_3 and D_4 influence emission line intensities like [O II], [O III], and Hα without altering the stellar continuum. D_3 correlates with ionized gas attributes and Active Galactic Nuclei (AGNs) presence, especially the [O III] emission line. Galaxies with AGNs or ionization from evolved stellar populations exhibit pronounced [O III] lines and higher D4000 values, indicating a larger fraction of older stars. While D_3 shows variable [O III] intensity, Hα remains stable. In D_4, oxygen lines and Hα increase simultaneously. Hα from ionized hydrogen gas, indicates active star formation, with its strength correlating with the [O II] emission due to ionization by young stars. However, AGNs can alter this relationship.

To more carefully interpret these latent representations, we use SHAP values to evaluate the significance of individual input wavelengths for predicting latent variables. A positive SHAP value indicates that increasing the feature's value increases the prediction, while a negative value decreases. In Figure 6 Panel A, SHAP value of D_3 peaks at [O II] emission lines and declines at [O III] and Hα emission lines. Similarly, Panel B shows SHAP value of D_4 rises at both [O II] and [O III] emission lines and falls at Hα emission lines. Therefore, we conclude that D_3 represents AGN strength, signifying the ratio of [O II] strength compared to [O III] and Hα, while D_4 indicates the ratio of Hα strength relative to [O III] and [O II] emission lines. This study's results are based on a single dataset and single model. There is a need for further investigation using additional datasets to verify and extend these findings.

4.2 What fundamental physical properties explain the observed SEDs?

Figure 2 shows that the CVAE model conditioned with M_* and the model conditioned with SFR ranks higher in the metric, implying that M_* and SFR contribute most significantly to describing the galaxy spectral distribution while penalizing for the increasing model complexity. Figure 5 shows the spectra reconstructed by the CVAEs when changing each physical property while keeping other latent variables constant at zero. When altering SFR, the intensity of the reconstructed spectra below 5000Å decreases while the continuum above 5000Å remains mostly the same. As expected, we see significant variation of [O II], [O III], and Hα emission lines with SFRs. This means that SFR is relatively independent of the continuum shape. However, that is not the case when M_*, sSFR, and metallicity are altered. M_* and metallicity are expected to be correlated [e.g., 20]. Therefore, increasing M_* and metallicity show similar changes to the reconstructed spectra. Decreasing the sSFR shows the

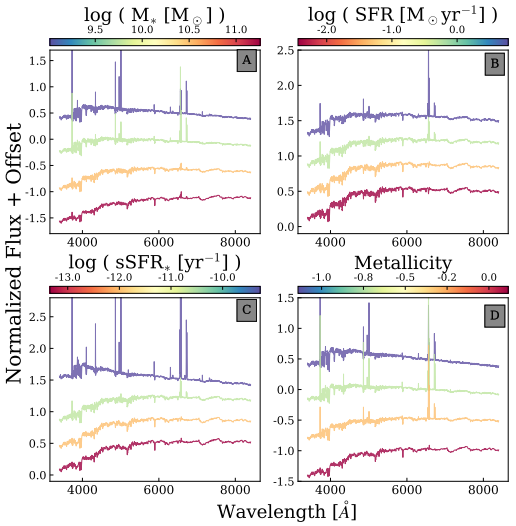

Figure 5: This figure consists of four panels (labeled A to D), each representing a different conditional data parameter: SFR, M, sSFR, and Metallicity. The four spectra within each panel are generated by a CVAE model, with one latent parameter held at zero while varying the values of the corresponding conditional parameter. The color of each spectrum represents the value of the respective conditional parameter.

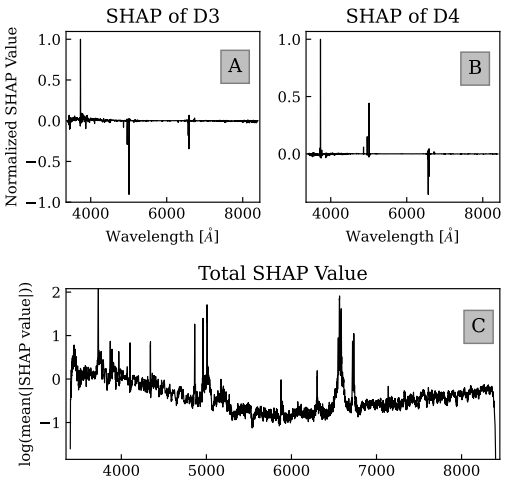

Figure 6: SHAP values as a function of wavelength for predicting latent variables. These figures display SHAP values as a function of wavelength, representing the importance of each wavelength in predicting latent variables. Panel A displays the SHAP values for D_3 for the galaxy (Plate1173-Mjd52790-FiberID111). Panel B presents the SHAP values for D_4 for the same galaxy. Panel C shows the logarithmically averaged absolute SHAP values of four latent variable.

continuous change from spectra dominated by young stellar populations with prominent nebular emissions to older stellar population emissions with minimal nebular emissions. These changes are in line with spectra changes expected for star-forming to quiescent stages of a galaxy. These results show that SFR affects the reconstructed spectra differently from M_*, sSFR, and metallicity.

4.3 Which spectral ranges are the most informative for representing observed SEDs?

Figure 6 shows the SHAP values in predicting the latent variables as a function of wavelength. SHAP values provide a measure of the contribution of each feature in predicting the latent variables, thereby indicating their importance. Panel C displays the average importance of each input wavelength in predicting latent variables, as measured by average absolute SHAP values. By taking the absolute value of these SHAP values, we emphasize the magnitude of a feature's impact on the predictions Our analysis reveals the follwing pattern: wavelengths above 5000Å, with the exception of a few specific emission lines, exhibit minimal influence on our predictions. Therefore, it appears that an adequate characterization of galaxy spectra relies primarily on the data below 5000Å and selected emission lines.

5 Conclusion

This study employed VAE and CVAE to extract four fundamental parameters from high-dimensional galaxy spectra. Our aim was to assess how these parameters, representing physical properties, affect the reconstruction of galaxy spectra and the formation of informative latent features. Specifically, D_1 is associated with M_*, and D_1 correlates with the SFR. D_3 indicates the ratio of [O II] to [O III] and Hα intensities, whereas D_4 represents the ratio of Hα to [O III] and [O II] emissions. Our findings show that incorporating M_* and SFR into the CVAE model enhances the accuracy of galaxy spectra reconstruction. Additionally, SHAP analysis identified that wavelengths below 5000Å and certain emission lines are particularly influential in these spectral ranges.

Acknowledgments

We express our deep appreciation to the five anonymous reviewers whose thorough reviews and insightful suggestions have greatly improved the quality of this manuscript. SC is supported by the Japan Society for the Promotion of Science (JSPS) under Grant No. 21J23611. This work has been supported by JSPS Grants-in-Aid for Scientific Research (TT: JP19H05076). This work has been supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (21H01128 and JP21J23611). This work has also been supported in part by the Collaboration Funding of the Institute of Statistical Mathematics "New Perspective of the Cosmology Pioneered by the Fusion of Data Science and Physics". Additionally, We also acknowledge the Center for Computational Astrophysics at the National Astronomical Observatory of Japan for providing access to their GPU cluster, which significantly supported our study. This research utilized several software packages, including Astropy [19], Matplotlib [6], NumPy [11], PyTorch [12], scikit-learn [15].

Checklist for Authors

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   See Chapter 4 for a discussion on the limited architectural experimentation and the analysis conducted on a single dataset. Further details will be elaborated in the main paper, which is being prepared for submission to an astronomical journal.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] All the code for downloading spectra, preprocessing, the model, training the model, and creating the figures used in this paper is publicly available in my GitHub repository: iwasakida.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes], Training details are also publicly available
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   [Yes], refer to Section 5 for details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets?
   [No] I used publicly accessible Galaxy data from SDSS with no specific license
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   [Yes] The process of downloading SDSS data details is also publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, R. Arcodia, E. Armengaud, M. Aubert, S. Avila, V. Avila-Reese, C. Badenes, C. Balland, K. Barger, J. K. Barrera-Ballesteros, S. Basu, J. Bautista, R. L. Beaton, T. C. Beers, B. I. T. Benavides, C. F. Bender, M. Bernardi, M. Bershady, F. Beutler, C. M. Bidin, J. Bird, D. Bizyaev, G. A. Blanc, M. R. Blanton, M. Boquien, J. Borissova, J. Bovy, W. N. Brandt, J. Brinkmann, J. R. Brownstein, K. Bundy, M. Bureau, A. Burgasser, E. Burtin, M. Cano-Díaz, R. Capasso, M. Cappellari, R. Carrera, S. Chabanier, W. Chaplin, M. Chapman, B. Cherinka, C. Chiappini, P. D. Choi, S. D. Chojnowski, H. Chung, N. Clerc, D. Coffey, J. M. Comerford, J. Comparat, L. d. Costa, M.-C. Cousinou, K. Covey, J. D. Crane, K. Cunha, G. d. S. Ilha, Y. S. Dai, S. B. Damsted, J. Darling, J. W. Davidson, R. Davies, K. Dawson, N. De, A. d. l. Macorra, N. D. Lee, A. B. d. A. Queiroz, A. D. Machado, S. d. l. Torre, F. Dell'Agli, H. d. M. d. Bourboux, A. M. Diamond-Stanic, S. Dillon, J. Donor, N. Drory, C. Duckworth, T. Dwelly, G. Ebelke, S. Eftekharzadeh, A. D. Eigenbrot, Y. P. Elsworth, M. Eracleous, G. Erfanianfar, S. Escoffier, X. Fan, E. Farr, J. G. Fernández-Trincado, D. Feuillet, A. Finoguenov, P. Fofie, A. Fraser-McKelvie, P. M. Frinchaboy, S. Fromenteau, H. Fu, L. Galbany, R. A. Garcia, D. A. García-Hernández, L. A. G. Oehmichen, J. Ge, M. A. G. Maia, D. Geisler, J. Gelfand, J. Goddy, V. Gonzalez-Perez, K. Grabowski, P. Green, C. J. Grier, H. Guo, J. Guy, P. Harding, S. Hasselquist, A. J. Hawken, C. R. Hayes, F. Hearty, S. Hekker, D. W. Hogg, J. A. Holtzman, D. Horta, J. Hou, B.-C. Hsieh, D. Huber, J. A. S. Hunt, J. I. Chitham, J. Imig, M. Jaber, C. E. J. Angel, J. A. Johnson, A. M. Jones, H. Jönsson, E. Jullo, Y. Kim, K. Kinemuchi, C. C. K. IV, G. W. Kite, M. Klaene, J.-P. Kneib, J. A. Kollmeier, H. Kong, M. Kounkel, D. Krishnarao, I. Lacerna, T.-W. Lan, R. R. Lane, D. R. Law, J.-M. L. Goff, H. W. Leung, H. Lewis, C. Li, J. Lian, L. Lin, D. Long, P. Longa-Peña, B. Lundgren, B. W. Lyke, J. T. Mackereth, C. L. MacLeod, S. R. Majewski, A. Manchado, C. Maraston, P. Martini, T. Masseron, K. L. Masters, S. Mathur, R. M. McDermid, A. Merloni, M. Merrifield, S. Mészáros, A. Miglio, D. Minniti, R. Minsley, T. Miyaji, F. G. Mohammad, B. Mosser, E.-M. Mueller, D. Muna, A. Muñoz-Gutiérrez, A. D. Myers, S. Nadathur, P. Nair, K. Nandra, J. C. d. Nascimento, R. J. Nevin, J. A. Newman, D. L. Nidever, C. Nitschelm, P. Noterdaeme, J. E. O'Connell, M. D. Olmstead, D. Oravetz, A. Oravetz, Y. Osorio, Z. J. Pace, N. Padilla, N. Palanque-Delabrouille, P. A. Palicio, H.-A. Pan, K. Pan, J. Parker, R. Paviot, S. Peirani, K. P. Ramírez, S. Penny, W. J. Percival, I. Perez-Fournon, I. Pérez-Ràfols, P. Petitjean, M. M. Pieri, M. Pinsonneault, V. J. Poovelil, J. T. Povick, A. Prakash, A. M. Price-Whelan, M. J. Raddick, A. Raichoor, A. Ray, S. B. Rembold, M. Rezaie, R. A. Riffel, R. Riffel, H.-W. Rix, A. C. Robin, A. Roman-Lopes, C. Román-Zúñiga, B. Rose, A. J. Ross, G. Rossi, K. Rowlands, K. H. R. Rubin, M. Salvato, A. G. Sánchez, L. Sánchez-Menguiano, J. R. Sánchez-Gallego, C. Sayres, A. Schaefer, R. P. Schiavon, J. S. Schimoia, E. Schlafly, D. Schlegel, D. P. Schneider, M. Schultheis, A. Schwope, H.-J. Seo, A. Serenelli, A. Shafieloo, S. J. Shamsi, Z. Shao, S. Shen, M. Shetrone, R. Shirley, V. S. Aguirre, J. D. Simon, M. F. Skrutskie, A. Slosar, R. Smethurst, J. Sobeck, B. C. Sodi, D. Souto, D. V. Stark, K. G. Stassun, M. Steinmetz, D. Stello, J. Stermer, T. Storchi-Bergmann, A. Streblyanska, G. S. Stringfellow, A. Stutz, G. Suárez, J. Sun, M. Taghizadeh-Popp, M. S. Talbot, J. Tayar, A. R. Thakar, R. Theriault, D. Thomas, Z. C. Thomas, J. Tinker, R. Tojeiro, H. H. Toledo, C. A. Tremonti, N. W. Troup, S. Tuttle, E. Unda-Sanzana, M. Valentini, J. Vargas-González, M. Vargas-Magaña, J. A. Vázquez-Mata, M. Vivek, D. Wake, Y. Wang, B. A. Weaver, A.-M. Weijmans, V. Wild, J. C. Wilson, R. F. Wilson, N. Wolthuis, W. M. Wood-Vasey, R. Yan, M. Yang, C. Yèche, O. Zamora, P. Zarrouk, G. Zasowski, K. Zhang, C. Zhao, G. Zhao, Z. Zheng, Z. Zheng, G. Zhu, and H. Zou. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *ApJS*, 249(1):3, June 2020. Publisher: The American Astronomical Society.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Feb. 1974.

[3] J. A. Baldwin, M. M. Phillips, and R. Terlevich. CLASSIFICATION PARAMETERS FOR THE EMISSION-LINE SPECTRA OF EXTRAGALACTIC OBJECTS. *Publications of the Astronomical Society of the Pacific*, 93(551):5, Feb. 1981.

[4] E. F. Bell, C. Wolf, K. Meisenheimer, H.-W. Rix, A. Borch, S. Dye, M. Kleinheinrich, L. Wisotzki, and D. H. McIntosh. Nearly 5000 Distant Early-Type Galaxies in COMBO-17: A Red Sequence and Its Evolution since z ~ 1. *ApJ*, 608(2):752, June 2004. Publisher: IOP Publishing.

[5] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. L. Fischer. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, May 2018.

[6] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering*, 9:90–95, May 2007. ADS Bibcode: 2007CSE.....9...90H.

[7] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-Supervised Learning with Deep Generative Models, June 2014.

[8] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, Feb. 2013.

[9] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, Feb. 2017. Curran Associates Inc.

[11] T. E. Oliphant. *NumPy: A guide to NumPy*. NumPy, 2006. Accessed: 2023-11-28.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[13] F. Pat, S. Juneau, V. Böhm, R. Pucha, A. G. Kim, A. S. Bolton, C. Lepart, D. Green, and A. D. Myers. Reconstructing and Classifying SDSS DR16 Galaxy Spectra with Machine-Learning and Dimensionality Reduction Algorithms, Nov. 2022. arXiv:2211.11783 [astro-ph].

[14] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, Jan. 1901.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[16] S. K. N. Portillo, J. K. Parejko, J. R. Vergara, and A. J. Connolly. Dimensionality Reduction of SDSS Spectra with Variational Autoencoders. *The Astronomical Journal*, 160(1):45, June 2020.

[17] S. Salim, J. C. Lee, S. Janowiecki, E. d. Cunha, M. Dickinson, M. Boquien, D. Burgarella, J. J. Salzer, and S. Charlot. GALEX-SDSS-WISE Legacy Catalog (GSWLC): Star Formation Rates, Stellar Masses, and Dust Attenuations of 700,000 Low-redshift Galaxies. *The Astrophysical Journal Supplement Series*, 227(1):2, Jan. 2016.

[18] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[19] The Astropy Collaboration, A. M. Price-Whelan, B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva, A. Ginsburg, J. T. VanderPlas, L. D. Bradley, D. Pérez-Suárez, M. De Val-Borro, (Primary Paper Contributors), T. L. Aldcroft, K. L. Cruz, T. P. Robitaille, E. J. Tollerud, (Astropy Coordination Committee), C. Ardelean, T. Babej, Y. P. Bach, M. Bachetti, A. V. Bakanov, S. P. Bamford, G. Barentsen, P. Barmby, A. Baumbach, K. L. Berry, F. Biscani, M. Boquien, K. A. Bostroem, L. G. Bouma, G. B. Brammer, E. M. Bray, H. Breytenbach, H. Buddelmeijer, D. J. Burke, G. Calderone, J. L. C. Rodríguez, M. Cara, J. V. M. Cardoso, S. Cheedella, Y. Copin, L. Corrales, D. Crichton, D. D'Avella, C. Deil, E. Depagne, J. P. Dietrich, A. Donath, M. Droettboom, N. Earl, T. Erben, S. Fabbro, L. A. Ferreira, T. Finethy, R. T. Fox, L. H. Garrison, S. L. J. Gibbons, D. A. Goldstein, R. Gommers, J. P. Greco, P. Greenfield, A. M. Groener, F. Grollier, A. Hagen, P. Hirst, D. Homeier, A. J. Horton, G. Hosseinzadeh, L. Hu, J. S. Hunkeler, Živezić., A. Jain, T. Jenness, G. Kanarek, S. Kendrew, N. S. Kern, W. E. Kerzendorf, A. Khvalko, J. King, D. Kirkby, A. M. Kulkarni, A. Kumar, A. Lee, D. Lenz, S. P. Littlefair, Z. Ma, D. M. Macleod, M. Mastropietro, C. McCully, S. Montagnac, B. M. Morris, M. Mueller, S. J. Mumford, D. Muna, N. A. Murphy, S. Nelson, G. H. Nguyen, J. P. Ninan, M. Nöthe, S. Ogaz, S. Oh, J. K. Parejko, N. Parley, S. Pascual, R. Patil, A. A. Patil, A. L. Plunkett, J. X. Prochaska, T. Rastogi, V. R. Janga, J. Sabater, P. Sakurikar, M. Seifert, L. E. Sherbert, H. Sherwood-Taylor, A. Y. Shih, J. Sick, M. T. Silbiger, S. Singanamalla, L. P. Singer, P. H. Sladen, K. A. Sooley, S. Sornarajah, O. Streicher, P. Teuben, S. W. Thomas, G. R. Tremblay, J. E. H. Turner, V. Terrón, M. H. V. Kerkwijk, A. De La Vega, L. L. Watkins, B. A. Weaver, J. B. Whitmore, J. Woillez, V. Zabalza, and (Astropy Contributors). The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *AJ*, 156(3):123, Aug. 2018.

[20] C. A. Tremonti, T. M. Heckman, G. Kauffmann, J. Brinchmann, S. Charlot, S. D. M. White, M. Seibert, E. W. Peng, D. J. Schlegel, A. Uomoto, M. Fukugita, and J. Brinkmann. The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey. *The Astrophysical Journal*, 613:898–913, Oct. 2004. ADS Bibcode: 2004ApJ...613..898T.

[21] C. W. Yip, A. J. Connolly, A. S. Szalay, T. Budavári, M. SubbaRao, J. A. Frieman, R. C. Nichol, A. M. Hopkins, D. G. York, S. Okamura, J. Brinkmann, I. Csabai, A. R. Thakar, M. Fukugita, and \. Ivezi\'c. Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey. *The Astronomical Journal*, 128(2):585, Aug. 2004.

A VAE

The main objective of VAE is to approximate the posterior distribution of latent variables given the input data. In VAE, first, we determine the architecture of the neural network (Figure 7) and then find the best parameters to approximate the probability distribution. This is done by maximizing the Evidence Lower Bound (ELBO). The ELBO is expressed as:

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

where $\mathbb{E}$ denotes the expectation, D_{KL} is the Kullback-Leibler divergence, θ and ϕ are neural network parameters, $\boldsymbol{z}$ are latent variables, and $\boldsymbol{x}$ is the observed data. The first term $\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ is the reconstruction loss, which encourages the decoded samples to be close to the original inputs. The second term $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))$ is the KL divergence, which measures the difference between the learned distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the prior distribution $p(\boldsymbol{z})$, typically a standard normal distribution. Each dimension is independent $p(\boldsymbol{z}) = \Pi_j p(z_j)$. Thus, this regularization constrains each element of the latent variable to be independent, which allows us to get disentangle representation. A VAE not only learns the reconstruction but also the representation $z \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$. In deep generative models, representation learning is equivalent to inference. That is why VAE is known as a good method for representation learning and we use it.

To enable gradient-based optimization, VAE uses the reparameterization trick. Latent variables are expressed as a deterministic function of the input and some random noise:

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}) \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation of the latent distribution predicted by the encoder, and $\odot$ represents element-wise multiplication.

The VAE loss function combines the reconstruction loss (typically Mean Squared Error for continuous data) and the KL divergence:

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}^{(l)}) - D_{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})] \tag{3}$$

with $\boldsymbol{z}^{(l)}$ being samples drawn from the latent distribution using the reparameterization trick. During training, a VAE aims to minimize the negative ELBO by adjusting its parameters, θ and ϕ. This approach is effectively equivalent to maximizing the ELBO. Through this process, the VAE trains the encoder to generate meaningful latent representations and the decoder to accurately reconstruct the input data from these representations.

B The SHAP values analysis

We employ the DeepExplainer function from Python's `shap` library to analyze SHAP values. These values, based on cooperative game theory, explain the output of machine learning models like our encoder g_ϕ. They measure the significance of each feature in a prediction by comparing its impact to a baseline value. This analysis reveals how individual features influence the model's decision-making. In a model f (our Encoder g_ϕ), with M input features, the SHAP value ϕ_i for feature i quantifies this influence based on cooperative game theory principles.The SHAP value $\phi_i(x)$ is calculated using the formula:

$$\phi_i(x) = \sum_{S \subseteq N \setminus i} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup i) - f_x(S)], \tag{4}$$

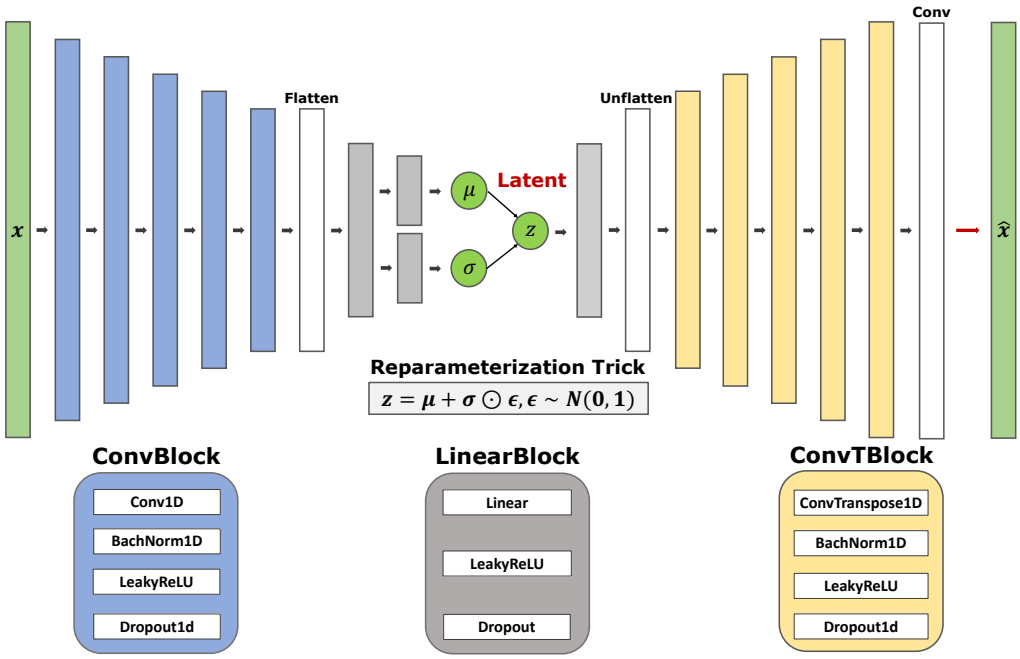

Figure 7: The illustration of the VAE architecture. The VAE architecture is depicted using color-coded blocks for different operations: ConvBloc in blue (Conv1D, BatchNorm1D, Dropout1D), LinearBlock in gray (Linear, LeakyReLU, Dropout), and ConvTBlock in yellow (ConvTranspose1D, BatchNorm1D, LeakyReLU, Dropout1D). Central white blocks represent Flatten and Unflatten operations for reshaping data, and an additional white block combines Conv1D, nn.ReLU, Flatten, Linear, and ReLU for matching input shape. These components collectively facilitate encoding of input $\boldsymbol{x}$, decoding of output $\hat{\boldsymbol{x}}$, and generation of patterns $\boldsymbol{z}$. Key processes include feature extraction, data processing, reconstruction, and data reshaping, with the Reparameterization Trick introducing learnable parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for the latent space distribution $\boldsymbol{z}$.

where $f_x(S)$ is the model output with a specific set of features S, and N is the total feature set. The formula considers all possible subsets of features excluding feature i, calculating the change in model output when i is included versus excluded. The term $\frac{|S|!(M-|S|-1)!}{M!}$ weights these changes to reflect the numerous combinations of features. This method quantifies the importance of each feature in the model's prediction, ensuring a fair distribution of impact among features, a concept rooted in cooperative game theory.