Understanding Galaxy Populations and Their Evolution Through Machine Learning

Luca Marziano

King's College,
University of Cambridge

May 2024

Supervised by Dr Sinan Deger and Prof. Hiranya Peiris

Abstract

Modern surveys are presenting us with datasets of unprecedented size on the evolution of galaxies through cosmic time. In this work, we explore the application of unsupervised deep convolutional neural networks to synthetically generated galaxy spectra through the variational autoencoder framework. Our goal was to create meaningful compressed representations of the spectra. The motivation behind this lies in exploring what insights a machine can get from input data with no pre-existing labels. By applying an information bottleneck, the model is trained to learn the most important features of the spectra. In particular, the variational autoencoder model is also trained to disentangle the learnt features. This is for the goal of direct interpretation of what the model has learnt using known physics. We find that 7 out of the 32 dimensions in the compressed representations contain useful information. These informative dimensions were then explored by applying the mutual information metric between them and the physical parameters used to generate the synthetic spectra. We find that these dimensions have in many cases picked up basic notions of said parameters, without ever seeing them in training. Further exploration of the representations was carried out through the method of latent space traversal. This involves observing how the spectra change when we alter one representation dimension at a time. We find some likely physical reasoning behind the results of the traversals, but believe that those not directly interpretable may be the most interesting for novel scientific discovery. The decision of the machine about what features are most important to learn could form the foundation of machine-assisted approaches in the future.

Contents

1 Introduction

Spectroscopy has provided the foundation for our modern understanding of galaxy formation and evolution (Conroy 2013). Deep, wide-field spectroscopic and imaging surveys are presenting us with vast amounts of new data on the evolution of galaxies, such as the Dark Energy Spectroscopic Instrument (Dey et al. 2019). Upcoming surveys, such as Rubin Observatory's Legacy Survey of Space and Time (Collaboration et al. 2020), will provide orders of magnitude more data than any previous survey. With classical techniques becoming insufficient to handle this data, it is clear that we have a need to expand our analysis tool kit.

For this expansion, we turn towards the field of machine learning. Machine learning has already found rapidly increasing use in astronomy, with a review being found in Baron (2019). Indeed, these methods have been applied for the study of galaxies (e.g. Dieleman et al. 2015; Walmsley et al. 2018; de Andres et al. 2022). Much of this work, however, involves the use of labels for training, restricting the data that can be used. This motivates our use of unsupervised machine learning methods, for which an overview can be found in Chen et al. (2022).

For this work, we concern ourselves with representation learning (Bengio et al. 2014), which has seen applications in a scientific context. This involves compressing input data into a lower dimensional but meaningful representation. Applications began with principal component analysis (PCA), for which an overview of the subject can be found in Jolliffe and Cadima (2016), and have since moved onto more sophisticated methods (e.g. Iten et al. 2020; Lucie-Smith et al. 2022; Teimoorinia et al. 2022; Melchior et al. 2023; Liang et al. 2023a; Liang et al. 2023b; Lanusse et al. 2023; Lucie-Smith et al. 2024). Many of these methods utilise autoencoders (Hinton and Zemel 1993), which are a type of model trained to reconstruct their input. Our work is focused on the use of variational autoencoders (VAEs), which were introduced by Kingma and Welling (2013) as a modification of the vanilla autoencoder with the goal of achieving interpretability in the learnt representations. They have helped to advance representation learning (Tschannen et al. 2018), with recent applications for scientific purposes (Portillo et al. 2020; Sedaghat et al. 2021; Yang 2022).

VAEs were applied to galaxy spectra by Portillo et al. (2020), but their implementation was a relatively shallow network of linear layers with the primary goal of demonstrating the framework's advantages over PCA. In Teimoorinia et al. (2022), autoencoders are used to summarise the diversity of spectra by classifying them onto a position on a 15×15 map. This provides a method of estimating physical parameters much faster than traditional methods, although with the same biases. The work of Melchior et al. (2023) (and the accompanying papers Liang et al. 2023a; Liang et al. 2023b) uses a different modification of the autoencoder on galaxy spectra, with a primary focus on the redshift. Finally, Lanusse et al. (2023) stray away from the autoencoder for their representation learning purposes, with an approach that combines images of galaxies and their spectra into a shared space.

In contrast to these previous works, we adopt a methodology most similar to Sedaghat et al. (2021), where we train a deep convolutional neural network (introduced by LeCun et al. 1989) as a VAE on synthetically generated galaxy spectra. We assess the VAE's performance and then use the mutual information (MI) (Vergara and Estvez 2013) metric to search for traces of physics in the compressed representations. Despite not seeing the physical parameters used to generate the spectra in training, we find that the model is able to learn basic notions of some of these parameters within the dimensions of the representations. We explore the representations further by utilising the generative capabilities of the model to see how the spectra change as we vary one dimension of the representations at a time. It is what the model chooses to learn that is of interest to us. This makes steps towards the goal of using machine learning for scientific discovery from observational data with no prior assumptions (Iten et al. 2020). In our particular case, we believe this sort of analysis could eventually be used to find novel relationships that advance our understanding of galaxy formation and evolution.

Section 2 presents a more extensive background on galaxy spectra and the VAE framework, with section 3 discussing the data and model we used. The results are presented in section 4, beginning with an assessment of the VAE's performance, followed by the exploration and interpretation of the representations. Finally, in section 5 we share concluding remarks and potential avenues for future work.

2 Background

2.1 Spectral Energy Distributions

Spectral energy distributions (SEDs) are key observational objects that encode a wealth of information about many of the physical properties of galaxies. By measuring quantities such as stellar mass, star formation rate, metallicities, and dust and gas content, we've been able to form our current understanding of how galaxies form and evolve. Two example SEDs can be seen in Figure 1. We provide a short overview of how some physical processes are reflected in the SEDs, but a full review of the subject can be found in Conroy (2013).

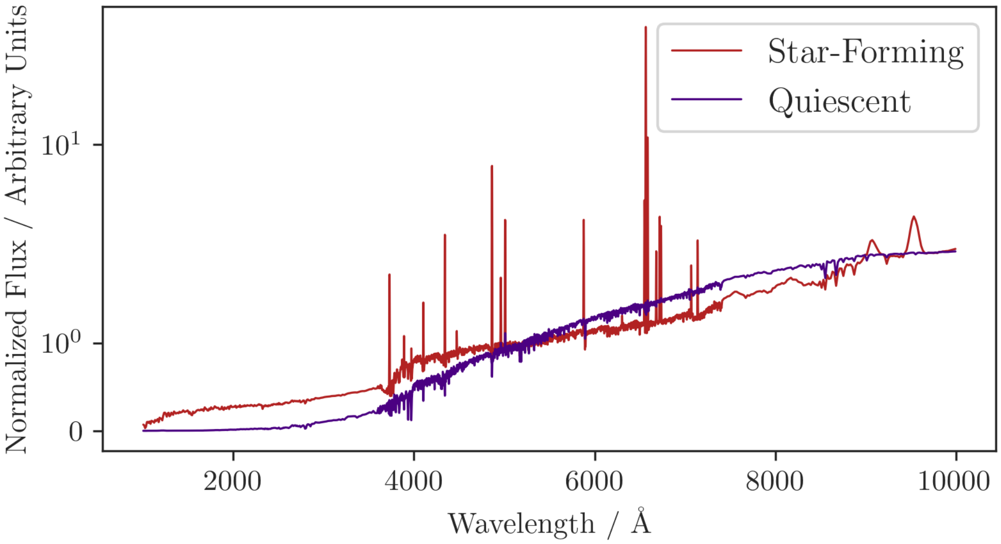

Figure 1: A sample of two spectra from our dataset to illustrate the differences between star-forming and quiescent galaxies. In particular, we see the expected stronger 4000 Å break for the quiescent galaxy and more emission lines for the star-forming galaxy.

The total stellar mass of a galaxy is believed to be fundamentally linked to many other physical properties of the galaxy. All the techniques for determining the stellar mass utilise stellar population synthesis models and generate mass-to-light ratios (Kettlety et al. 2017). It is the overall luminosity of the SED that is largely influenced. Since changing the stellar mass also changes various other parameters, specific wavelengths regions are chosen to perform this analysis.

A particularly important property that can be extracted from SEDs is the star formation rate (SFR). Many methods have been suggested for measuring it, in particular the use of intensity in the ultraviolet (UV) wavelengths (e.g. Salim et al. 2007; Schiminovich et al. 2007) and the strength of the emission lines, with Hα being a common choice (e.g. Glazebrook et al. 1999; Brinchmann et al. 2004). The underlying physics for this is that young stars emit mostly in the UV and provide the radiation required for the ionisation that leads to emission lines. It is argued by Glazebrook et al. (1999) that the Hα strength gives a better measure of the instantaneous SFR, as the massive young stars required die out very quickly. Additionally, this measurement is less impacted by dust attenuation, which has greater effects at UV wavelengths (Salim and Narayanan 2020).

Another important feature of a galaxy SED is the 4000 Å break (see e.g. Poggianti and Barbaro 1997; Kriek et al. 2006; Renard et al. 2022), which is a jump in the spectral intensities around that wavelength. It is a result of absorption features from metals and the Balmer series, which are more pronounced in the spectra of older, cooler stars. The strength of the break can provide an insight into the age of the galaxy, as well as the star formation history and the metallicity (Worthey 1994). The similar effects of age and metallicity leads to the age-metallicity degeneracy (Worthey 1999). A strong break is often a characteristic feature of a quiescent galaxy.

Finally, we consider the importance of dust and metal content. Dust will absorb radiation at UV wavelengths and emit at infrared (IR) wavelengths (Calzetti 2001). This process leads to the attenuation of intensities at lower wavelengths and an increase in intensities at higher wavelengths. The metallicity of both the stars and gas in the galaxy has effects on both the emission and absorption lines (see e.g. Kewley et al. 2019; Schady et al. 2024). In contrast to SFR, which primarily influences the intensity of the emission lines, the metallicity also influences the ratios of emission lines. This is reflective of changes in chemical composition, which are another fundamental interest in the study of galaxy evolution.

2.2 Relations of Galaxy Properties

As well as summarising how changes in certain physical properties of a galaxy might affect the SED, it is also important to mention the links between these properties. A model of the galaxy population with predictions for many of these relations can be found in Alsing et al. (2024).

The star-forming sequence is the relation between the stellar mass and SFR in galaxies. While this is complex to model, a power law or similar is typically fitted (see e.g. Santini et al. 2017; Leja et al. 2022). This models the SFR increasing with stellar mass. Physical origins of this lie in the balance of gas supply in the galaxy and the efficiency at which stars can be formed from this gas (Kennicutt 1998). However, many of these studies are restricted to star-forming galaxies. This is because identifying large samples of quiescent galaxies has its difficulties (Steinhardt et al. 2020). There have been examples of massive quiescent galaxies being identified (see e.g. Girelli et al. 2019).

It is also observed that for a given mass the SFR increases with redshift up to $z \approx 2$, which is known as *cosmic noon*, before decreasing again (Madau and Dickinson 2014). This follows from a number of factors during that epoch, particularly an abundant gas supply and more frequent galaxy mergers and interactions (Frster Schreiber and Wuyts 2020).

The gas-phase metallicity is another property believed to roughly increase with stellar mass. Physical origins of this are explored in Tremonti et al. (2004). A likely cause is that more massive galaxies tend to have more extensive star formation histories. These stars have returned the metals they have produced to the interstellar medium at the end of their lives. This is part of the baryon cycle (see e.g. Angls-Alczar et al. 2017; Proux and Howk 2020). A similar relation has been observed for the stellar metallicity and stellar mass (see e.g. Panter et al. 2008). Considering the existence of both of these relations, it follows that gas-phase metallicity and stellar metallicity are also correlated (see e.g. Fraser-McKelvie et al. 2021).

Stellar mass is also found to have a weak correlation with dust attenuation in star-forming galaxies (Garn and Best 2010). This gives the dust attenuation relations with other properties. We draw attention to the increase in dust attenuation with gas-phase metallicity (see e.g. Zahid et al. 2013; Qin et al. 2023). A significant fraction of the metals released at the end of stellar lifetimes condense into dust grains (Asano et al. 2013), so this also has direct physical reasoning.

2.3 Autoencoders

In order to describe the VAE, we must first begin with a discussion of how a general autoencoder works. An autoencoder is a type of neural network trained to reconstruct its input by first compressing it into a meaningful lower dimensional representation. They were first introduced by Rumelhart et al. (1986), and an overview can be found in Bank et al. (2021).

Mathematically, it learns an encoding function $A : \mathbb{R}^n \to \mathbb{R}^m$ and a decoding function $B : \mathbb{R}^m \to \mathbb{R}^n$ such that the quantity

$$\mathcal{L} = \Delta\left(x, B \circ A(x)\right) \equiv \mathcal{L}_{\mathrm{recon}} \tag{1}$$

is minimised. Here, x is the input and Δ is the reconstruction loss function. The quantity $B \circ A(x)$ is the result of applying the encoding function A followed by the decoding function B to the input. An illustration of the process can be seen in Figure 2.

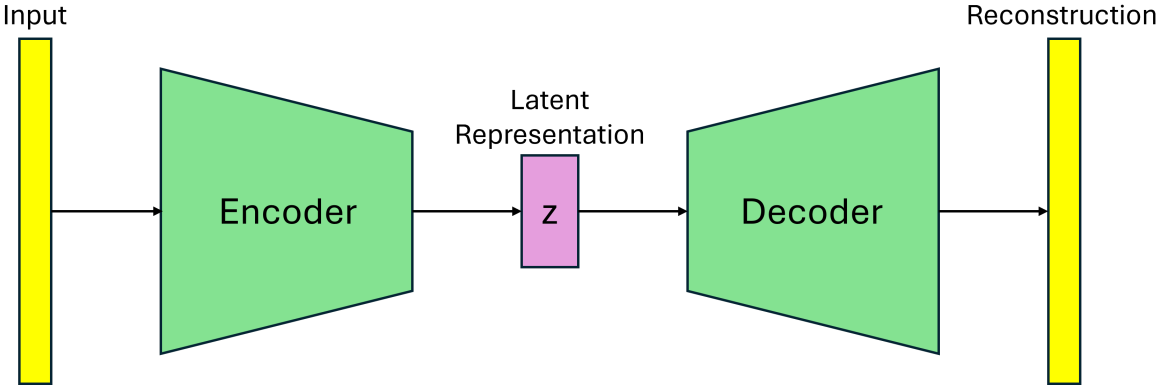

Figure 2: A schematic representation of how an autoencoder works.

Our input data has dimension n and the compressed representation, formally known as the *latent representation*, has dimension m. By taking $m < n$ we impose an information bottleneck (Tishby et al. 2000) on the network, meaning that the model must learn to extract the most important features from the input data. This ensures that the reconstructions will be faithful. It is this point that captures the essence of this work. We are particularly interested in what the model chooses to learn.

2.4 Variational Autoencoders

In representation learning, it is of great interest to *disentangle* the generative factors of the data (Bengio et al. 2014). This means that the dimensions of our representation capture particular underlying features of the input. Learning as many independent factors, while retaining enough information about the data, is important for our interpretation of what the model has learned. This is the motivation behind the VAE, which is trained to regularise its latent space, along with being able to reconstruct its input. A comprehensive overview of the motivation and theory behind VAEs can be found in Kingma and Welling (2019).

The key difference between the vanilla autoencoder and the variational version is that the latent representations are now drawn from a distribution with learnable parameters, rather than learnt directly. In order to convert to a VAE, we must add an additional term to the quantity to be minimised (1). It becomes

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta D_{KL}(q(z|x)||p(z)). \tag{2}$$

Here z is the latent representation, $q(z|x)$ is the approximate posterior distribution (which is learnt by the encoder), and $p(z)$ is the prior distribution of the latent representation, which is something we specify. Additionally, D_{KL} is the Kullback-Leibler divergence (Kullback and Leibler 1951), which is a statistical measure of the similarity of two distributions. It is equal to zero if the two distributions are the same. Finally, we specify β, which is a relative weight between the two terms (Higgins et al. 2016). Choosing an optimal value of β is of great importance (Mathieu et al. 2019). We must find the balance between enforcing disentanglement and retaining good reconstruction quality.

Neural networks learn via the process of backpropagation (Plaut et al. 1986), which involves calculating gradients over the learnt variables. Since the latent representation is now drawn from a random distribution, we must use the *reparameterization trick* to be able to calculate unbiased gradients over the now stochastic variable. In order to do this, we express the latent representations as some differentiable and invertible transformation of another random variable. This takes the form

$$z = g(\phi, x, \varepsilon), \tag{3}$$

where ϕ are the parameters learnt by the encoder and $\varepsilon \sim p(\varepsilon)$ is a stochastic variable that is independent of ϕ and x. This diverts the non-differentiable operation out of the network. An illustration of how VAEs work can be seen in Figure 3.

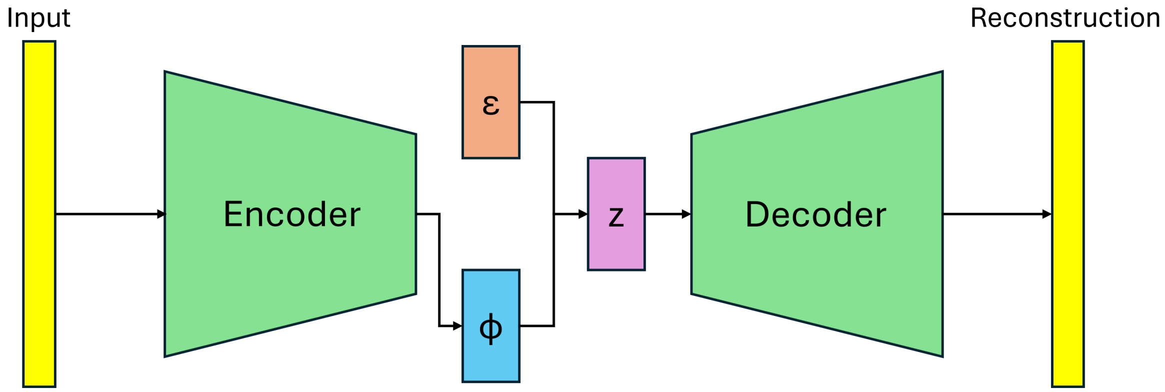

Figure 3: A schematic representation of how a VAE works. Instead of learning z directly, the models learns a set of parameters ϕ. These parameters are used in conjunction with a random variable ε to give a distribution that z is then drawn from. Since the random variable ε is not part of the network, backpropagation can successfully occur along a deterministic pathway.

We reiterate that the overarching idea is that the network learns to reconstruct its input, but also is trained to constrain the distribution of latent representations to be similar to some prior distribution, which is specified by the user. This second point is not present in the vanilla autoencoder, and is the source of the disentanglement we seek to reach the goal of interpretability.

3 Methods

3.1 Data

We use a dataset of 100000 spectra that were generated using stellar population synthesis (SPS). SPS is the method of combining star formation and metal enrichment histories of a galaxy, along with stellar evolution and attenuation by interstellar dust in order to predict its SED. In particular, our data was generated using the Flexible Stellar Population Synthesis package (Conroy et al. 2009; Conroy and Gunn 2010). A summary of some of the parameters used to generate the spectra can be seen in Table 1. These parameters are drawn from a model capable of generating realistic galaxy populations, meaning that any relations between these parameters is reflective of what was discussed in section 2.2. This consideration is important for our discussions in sections 4.3 and 4.4.

Parameter	**Description**
$\log_{10}(M/M_\odot)$	The base 10 logarithm of the current stellar mass in units of solar masses
$\log_{10}(Z/Z_\odot)$	The base 10 logarithm of the stellar metallicity in units of solar metallicity
$\log_{10}(Z_{\text{gas}}/Z_\odot)$	The base 10 logarithm of the gas metallicity in units of solar metallicity
z	The observed redshift
D_1	A dust parameter describing the attenuation of light from stars younger than 10^7 years old
n	The power law index of the attenuation curve
D_2	A dust parameter describing the attenuation of light from stars older than 10^7 years old
$\log_{10}(\text{SFR}/M_\odot \text{yr}^{-1})$	The base 10 logarithm of the star formation rate averaged over the last 10^8 years in units of solar masses per year

Table 1: The parameters used to generate the synthetic spectra.

The motivation for using synthetically generated data lies in allowing us access to a large amount of clean data (Lu et al. 2024). The application of this sort of analysis for scientific purposes is very much in its infancy, with the full possibilities not yet known. With this in mind, it seems logical to test it on synthetic data before moving onto real observational data.

The spectra were restricted to the range 1000 Å to 10000 Å and then normalised by their median value in the range 5000 Å to 5500 Å. Normalisation is often required to reduce fluctuations in the spectra and improve the prediction ability of quantitative models (Guezenoc et al. 2019). It is also used to accelerate the training process and improve the generalisation of the model (Huang et al. 2020). Each spectrum is comprised of 4582 data points across the wavelength range. They are not, however, evenly distributed. The distribution can be seen in Figure 4. The implications of this upon the model's performance are discussed in section 4.1.1.

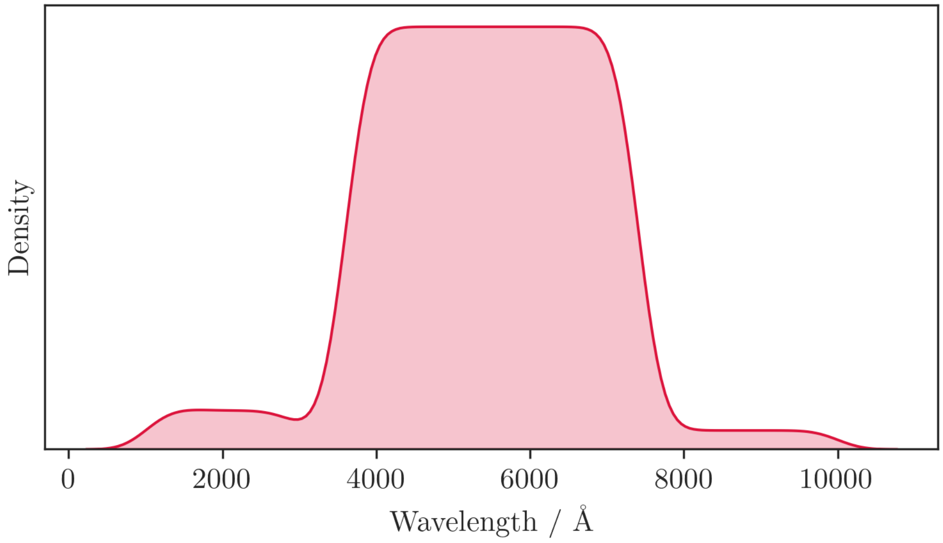

Figure 4: The distribution of data points across the wavelength range. We see that a majority of the points lie in the range 3000 Å to 8000 Å.

3.2 Model

3.2.1 Convolutional Neural Networks

The encoder uses a convolutional architecture, in which convolutional layers make up a bulk of the operations. These have found extensive use in image-driven tasks, as the convolution operation is particularly well-suited for feature extraction. A feature of an image, for example, is determined by multiple pixels near each other. The spectra can be viewed as images of 4582×1 pixels and, as in the case of images, the features are determined by the flux values at nearby wavelengths. A full introduction to convolutional neural networks can be found in O'Shea and Nash (2015).

In the one-dimensional case, a kernel of a specified size will move through the input data at a specified rate, known as the stride, and map vectors of some length to new vectors of a specified length. The number of vectors in the input mapped to one vector in the output is determined by the size of the kernel. The initial length of the vectors is referred to as the number of input channels, with the length of the output vectors being the number of output channels. It is common to select the kernel and stride such that the output contains fewer vectors than the input. In our case, we do this to systematically reduce the dimensions towards the bottleneck. A simple example is outlined in Figure 5. A reverse operation exists, known as a transposed convolutional layer, and is utilised in the decoder.

3.2.2 Architecture

The model has 11 convolutional layers that take the input from a size of 4582 to a size of 512. Two fully connected layers then reduce this to two separate outputs of size 32. This is our chosen dimensionality of the latent space. We use these two outputs to parameterize our latent representations

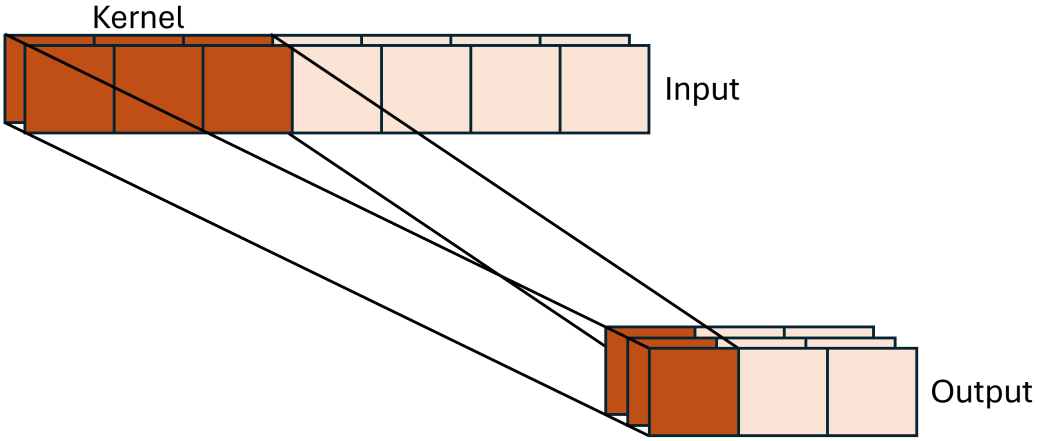

Figure 5: An illustration of how a convolutional layer works. Here we have an input of 7 vectors of length 2 (this is the number of input channels), a kernel size of 3, and a stride of 2. With 3 output channels, this gives the output as 3 vectors of length 3. The highlighted output vector is determined only by the three highlighted input vectors.

as

$$z = \sigma + \mu \odot \varepsilon \sim \mathcal{N}(\mu, \sigma). \tag{4}$$

Here, $\varepsilon \sim \mathcal{N}(0, \mathbb{I}_{32\times32})$ and $\odot$ denotes element-wise multiplication. Since 32 is the dimensionality of the latent space, z, ε, μ, and σ are all of this size[1]. This means that σ and μ are the sets of parameters learnt by the network such that z is drawn from a multivariate Gaussian distribution.

The Gaussian distribution has multiple advantages. Firstly, it is a simple implementation of the reparameterization trick (3), allowing us to successfully backpropagate through the network during training. Secondly, it leads to an analytic result for the Kullback-Leibler divergence, which we present in section 3.2.3. It also allows for simple generation of new spectra, which we utilise in section 4.4.

A reverse operation with transposed convolutional layers returns the output of the original size. A schematic representation of the model can be seen in Figure 6.

3.2.3 Training

In order to train our model, we must define a loss function. We return to the quantity $\mathcal{L}$, which we defined earlier (2). For the reconstruction loss, we choose the total squared error, which is given by

$$\mathcal{L}_{\text{recon}} = \sum_{i=1}^{M} (x_i - \hat{x}_i)^2, \tag{5}$$

where we have denoted the reconstruction of a particular input x as $\hat{x}$ and the subscript denotes the i^{th} component. Additionally, M is the total number of data points in each spectrum (4582 in our case). For the other part of our loss function, we must select a prior distribution for z and calculate the Kullback-Leibler divergence. For continuous random variables, it is given by

$$D_{KL}(q(z|x)||p(z)) = \int_{-\infty}^{\infty} q(z|x) \ln\left(\frac{q(z|x)}{p(z)}\right) \mathrm{d}z. \tag{6}$$

[1] In general, σ is a 32×32 matrix. The model learns a vector of size 32 that contains the diagonal entries of this matrix, while all other entries are 0. We do not make this distinction and simply label the vector as σ.

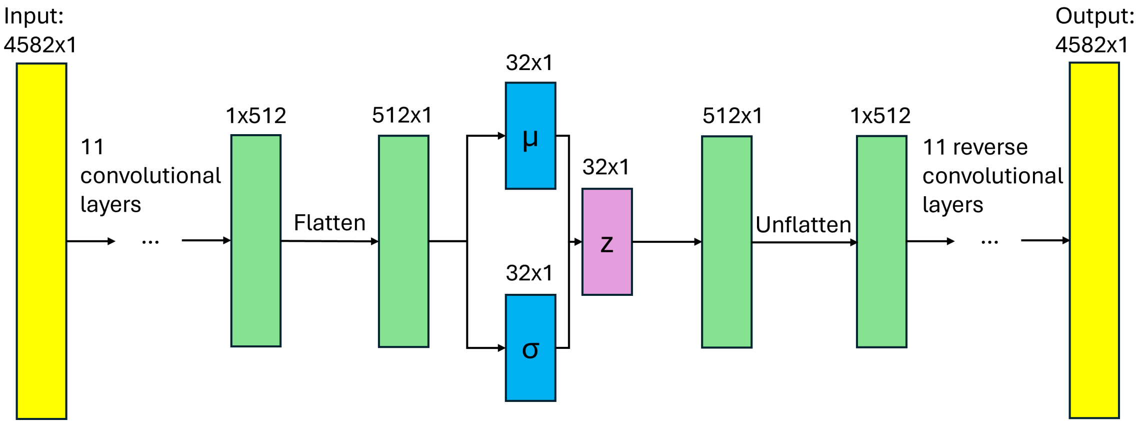

Figure 6: A schematic representation of the model architecture. The network learns a mean and standard deviation vector, both of size 32, that parameterize a multivariate Gaussian from which the latent representation, which is also of size 32, is drawn from. The flatten and unflatten processes serve only to align the dimensions correctly in our case[2].

We select the prior to be $p(z) = \mathcal{N}(0, \mathbb{I}_{32\times32})$. Combined with our previously mentioned learnt distribution of z, this gives the analytic result

$$D_{KL}(\mathcal{N}(\mu,\sigma)||\mathcal{N}(0,\mathbb{I}_{32\times32})) = \frac{1}{2}\sum_{i=1}^{32}(\mu_i^2 + \sigma_i^2 - \ln(\sigma_i^2) - 1). \tag{7}$$

This gives our overall loss function to be minimised as

$$\mathcal{L} = \sum_{i=1}^{M}(x_i - \hat{x}_i)^2 + \frac{\beta}{2}\sum_{i=1}^{32}(\mu_i^2 + \sigma_i^2 - \ln(\sigma_i^2) - 1). \tag{8}$$

We experiment with the value of β, with details of this being discussed in section 4.1.

We implement the model using the PyTorch library for Python (Paszke et al. 2019). We split the data into 80000 spectra for training and 20000 for testing. The model is trained for up to 100 epochs, with early stopping occurring if the validation loss (the performance accessed on the testing data) does not improve over 20 epochs. We then return the model from the epoch with the lowest validation loss. This helps to reduce overfitting (Li et al. 2024) and should provide the best general performance. We also set a weight decay (Andriushchenko et al. 2023) of 0.1 as an additional method for preventing overfitting. For optimisation, we use Adam (Kingma and Ba 2017) with a learning rate (Wu et al. 2019) that begins at 3×10^{-4} and decreases by a factor of 0.5 every 10 epochs. Finally, we use a mini-batch size (Masters and Luschi 2018) of 32. These parameters were chosen based on preliminary experiments with training the model.

[2] In general, they can compress a matrix into a single vector, which is needed for the latent space, and then undo this process in the decoder.

4 Results

4.1 Choosing β

As was mentioned in section 2.4, choosing an optimal value of β is of great importance, as it controls the level of disentanglement the VAE is trained to exhibit (Burgess et al. 2018). Too high a value, however, will result in poor reconstruction quality, as enforcing disentanglement is prioritised over the effectiveness of the encoding and decoding functions. For our purposes we need sufficiently good reconstructions to ensure that the model is learning useful latent representations that capture the important features of the input data. To this end, we decide that $\beta = 0.3$ is an appropriate balance. The reasoning behind this decision is presented in sections 4.1.1 and 4.1.2. Beginning with section 4.2, all subsequent results discussed are for the model trained with this value.

4.1.1 Reconstruction Quality

Our first assessment of the model's performance as a VAE is looking at the reconstruction quality. While the VAE framework introduces the goal of disentanglement, the reconstruction quality still remains essential as it is a measure of how useful the latent representations are. Poor reconstruction quality indicates that the model has not learnt the key features of the training data, and would render it unfit for our purposes.

The relative weight β in (8) plays an important role in the reconstruction quality. We expect a higher value to result in poorer reconstructions, as disentanglement is enforced more. The model was trained with multiple values and we present the results for $\beta = 0.1$, $\beta = 0.3$, and $\beta = 0.9$. A qualitative comparison of their performance can be seen in Figure 7, where we do indeed see the expected behaviour.

To perform a quantitative analysis of how well the different trainings of the model reconstruct the spectra, we use the mean squared error for each point in the spectra. We mentioned in section 3.1 that each spectrum is comprised of 4582 data points. For the i^{th} data point we compute the mean squared error as

$$\Delta_i = \frac{1}{N}\sum_{j=1}^{N}\left(x_i^{(j)} - \hat{x}_i^{(j)}\right)^2, \tag{9}$$

where N is the total number of spectra (100000 in our case) and $x_i^{(j)}$ denotes the i^{th} component of the j^{th} spectrum. We use this metric as it mimics the reconstruction loss the model is trained with (5).

The comparison of these results can be seen in Figure 8. We now quantitatively see that the higher values of β do indeed lead to larger errors. Returning to our discussion in section 3.1, we observe that the errors are larger at the extremes of the continuum. This reflects the fact that the resolution of the data in these regions is lower, as depicted in Figure 4.

4.1.2 Informative Dimensions

In a similar vein to Sedaghat et al. (2021), we want to quantify how informative the latent dimensions are. They use the median absolute deviation (Arachchige and Prendergast 2019) of the latent dimensions and suggest that it can provide insight into the level of disentanglement being achieved. We instead use the covariance (Fissler and Pohle 2023) between the input data and the latent dimensions. Our motivation for this is that it is a measure of the joint variability of two variables. Intuitively, if a latent dimension contains useful information about the input, we expect it to show a non-zero covariance.

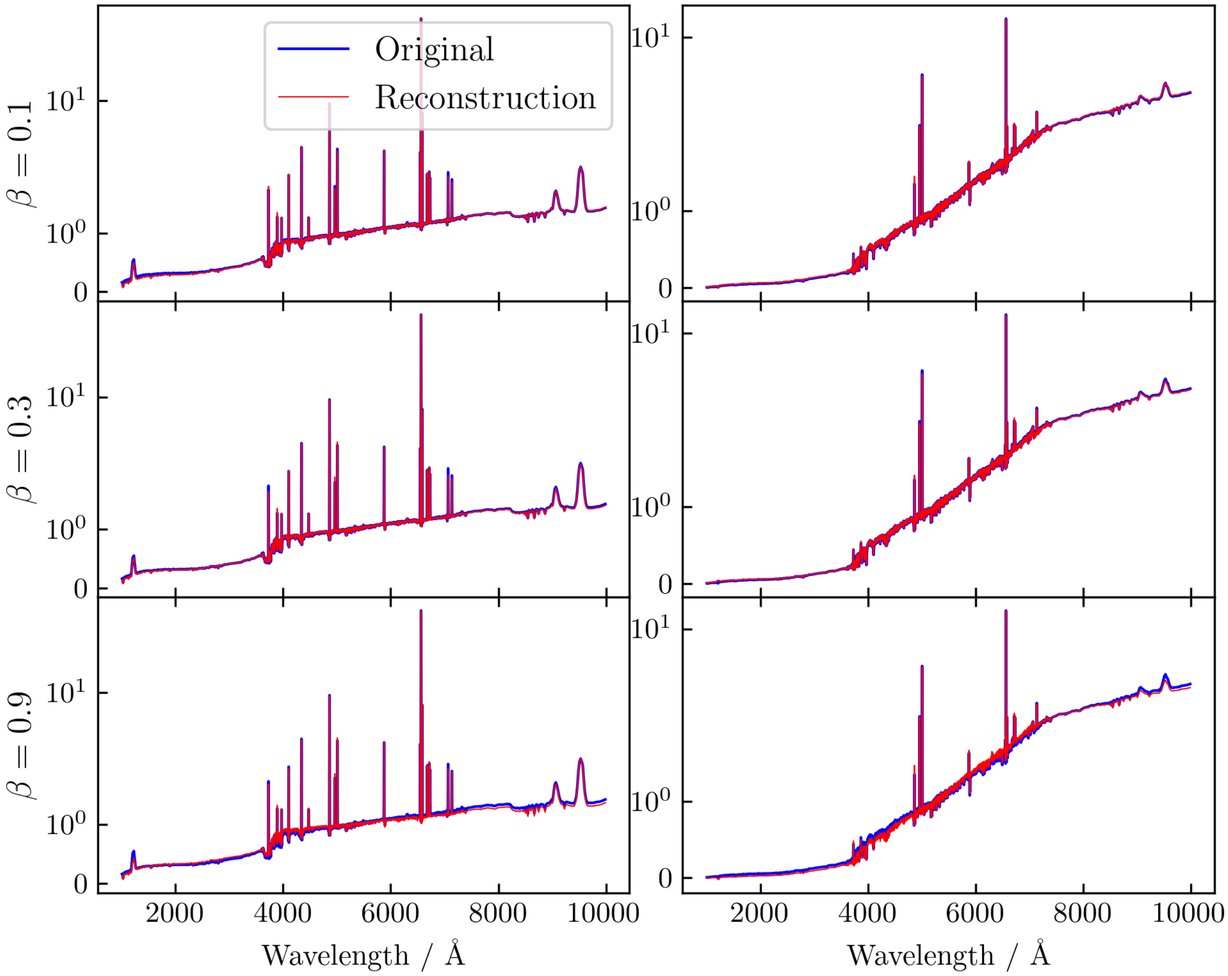

Figure 7: A comparison of the reconstructions of two different spectra for the model trained with three different values of β. We observe the expected behaviour of increased β leading to poorer reconstructions.

The covariance between two jointly distributed variables X and Y is given by

$$\mathrm{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{10}$$

where $\mathbb{E}$ denotes the expectation value of a quantity. To apply this metric to our data, we define a quantity to express the informativeness of a dimension i as

$$\frac{1}{M}\sum_{j=1}^{M} \left|\mathrm{cov}(\{z_i\}, \{x_j\})\right|, \tag{11}$$

where M is again the number of data points in each spectrum, $\{z_i\}$ denotes the set of the i^{th} value of the latent representations, and $\{x_j\}$ denotes the set of the j^{th} data point of the spectra. We take an absolute value since the covariance can be positive or negative. This can be seen as a mean absolute covariance of the latent dimension with the input data. Put simply, we are quantifying

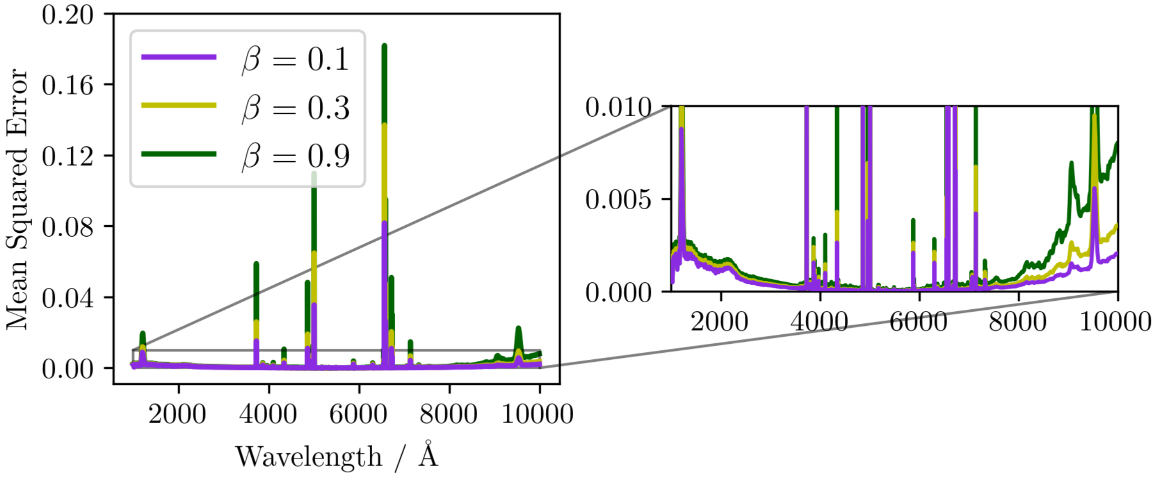

Figure 8: A comparison of the mean squared errors across the spectra for the three values of β. We see the expected behaviour of increased β leading to larger errors. Additionally, we highlight that the continuum errors are in accordance with the resolution of the data found in Figure 4.

how a latent dimension changes with each data point in the spectra and then averaging over the data points to give a single score for each dimension. Taking the approach of Burda et al. (2016), we label a dimension as *informative* if this score is greater than 0.01.

The results of this analysis for the three values of β can be seen in Figure 9. We see that enforcing more disentanglement leads to fewer informative dimensions. With too few informative dimensions, the model is not learning enough useful features of the input data, leading to the poorer reconstructions seen in the previous section. However, too many informative dimensions suggests that information is being leaked between these dimensions. We want the latent dimensions to form an orthogonal basis encoding the important features of the input data (Sarhan et al. 2020). Too many informative dimensions suggests that this is not occurring sufficiently, and hence more disentanglement is required.

Our chosen value of $\beta = 0.3$ displays seven informative dimensions. These are nodes 2, 6, 12, 13, 18, 26, and 28. Subsequent discussion will be concerned only with these dimensions.

4.2 Mutual Information of Latent Dimensions

To get a closer look at the disentanglement of the informative dimensions, we calculate the MI between them, which quantifies their mutual dependence. It is more versatile that the Pearson correlation coefficient, which only measures linear dependencies (Schober et al. 2018), and is equal to zero if, and only if, the two variables are statistically independent (Cover and Thomas 2006). For two continuous random variables X and Y with values over the space $\mathcal{X} \times \mathcal{Y}$, the MI is calculated as

$$I(X,Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} P_{(X,Y)}(x,y) \ln \left(\frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \right) \mathrm{d}x\mathrm{d}y, \tag{12}$$

where p_X and p_Y are the marginal distributions of the variables and $p_{(X,Y)}$ is their joint distribution. The use of the natural logarithm indicates that we are working in natural units (nat).

Estimating the MI from samples of two variables, however, is a non-trivial challenge. This is because it requires good knowledge or a good estimation of the joint distribution (Paninski 2003).

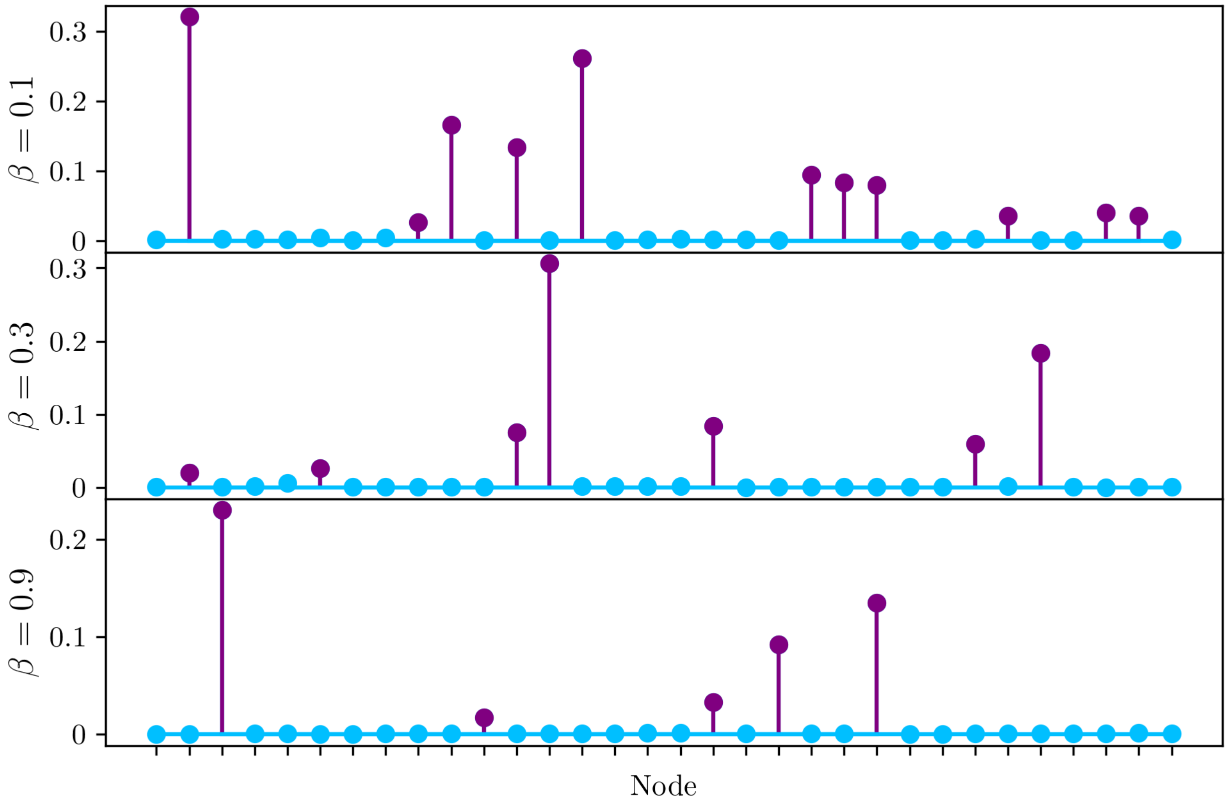

Figure 9: The results of our informativeness calculations. Informative nodes are shown in purple. We observe that a higher value of β results in fewer informativeness dimensions as disentanglement is pushed further.

To estimate the MI we use the GMM-MI algorithm (Piras et al. 2023), which is based on Gaussian mixture models. It is computationally efficient and has demonstrated strong performance in the context of representation learning.

The results of our estimations can be seen in Figure 10. We see that most of the informative dimensions have little MI, but one pair stands out. Nodes 13 and 28 have an MI of approximately 1.73 nat, which is significantly higher than any other pair. We plot density maps to show the mutual behaviour of the informative dimensions in Figure 11 and see that these nodes appear to have a strong linear correlation. The less structured plots seen for other pairs of nodes suggest more successful disentanglement. We conclude that nodes 13 and 28 have likely not been disentangled and we must consider this in subsequent analysis. A slightly higher value of β is likely to remedy this issue, but we delay this to a subsequent work, as its effects should not be too profound.

4.3 Mutual Information with Physical Parameters

Our primary interest lies in finding out what the model has learnt. We begin this endeavour by searching for traces of physics in the latent dimensions. Each of the spectra that the model was trained on have corresponding values of the physical parameters used to generate them. These were presented in Table 1. It is very important to note that the model has not seen these parameters during training, but we are interested in if it has been able to pick up notions of them by only

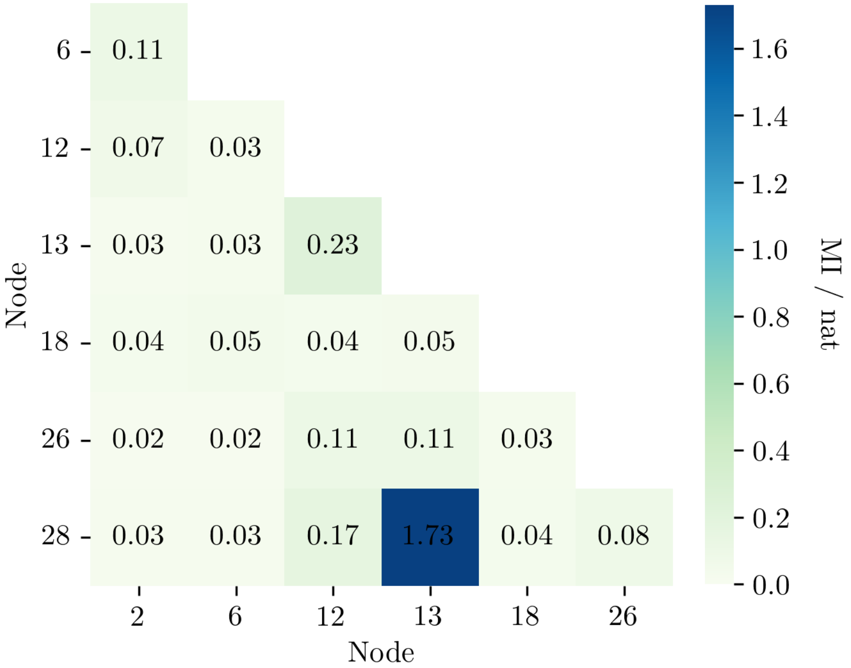

Figure 10: The results of our MI estimations between the informative dimensions shown to two decimal places. We observe a much higher mutual dependence between nodes 13 and 28 than any other pair. This leads us to the conclusion that these dimensions have not been disentangled successfully.

looking at the spectra.

To perform this analysis we again utilise the MI metric. The results of our estimations can be seen in Figure 12. We see that basic notions of the parameters are indeed being picked up. Referring back to section 3.1, we understand why most of the nodes convey a non-trivial amount of information about more than one parameter. The spectra were generated to reflect a realistic distribution of parameters. If two parameters show some correlation, then there is no need for the model to dedicate independent nodes to them. Only the results with the informative dimensions are shown, but the MI was also estimated for the remaining dimensions. We find, perhaps reassuringly, that none of the other dimensions have an MI greater than 0.005 nat. This validates our discussion in section 4.1.2. We take a closer look at the relationships by plotting the density maps in Figure 13. This gives us a better idea of how the parameters change when we vary the node values, which is important for the discussion in section 4.4. We observe some mostly linear relations and, particularly for the gas metallicity, some that are non-linear.

Both the results of the MI estimations and the form of the density plots provide more evidence that nodes 13 and 28 have not been disentangled successfully. They have picked up notions of the same parameters, and these notions take near-identical forms.

4.4 Latent Space Traversal

To further our understanding of what the network has learnt, we use the method of *latent space traversal*. Put simply, we want to observe directly how changing the latent dimensions affects the

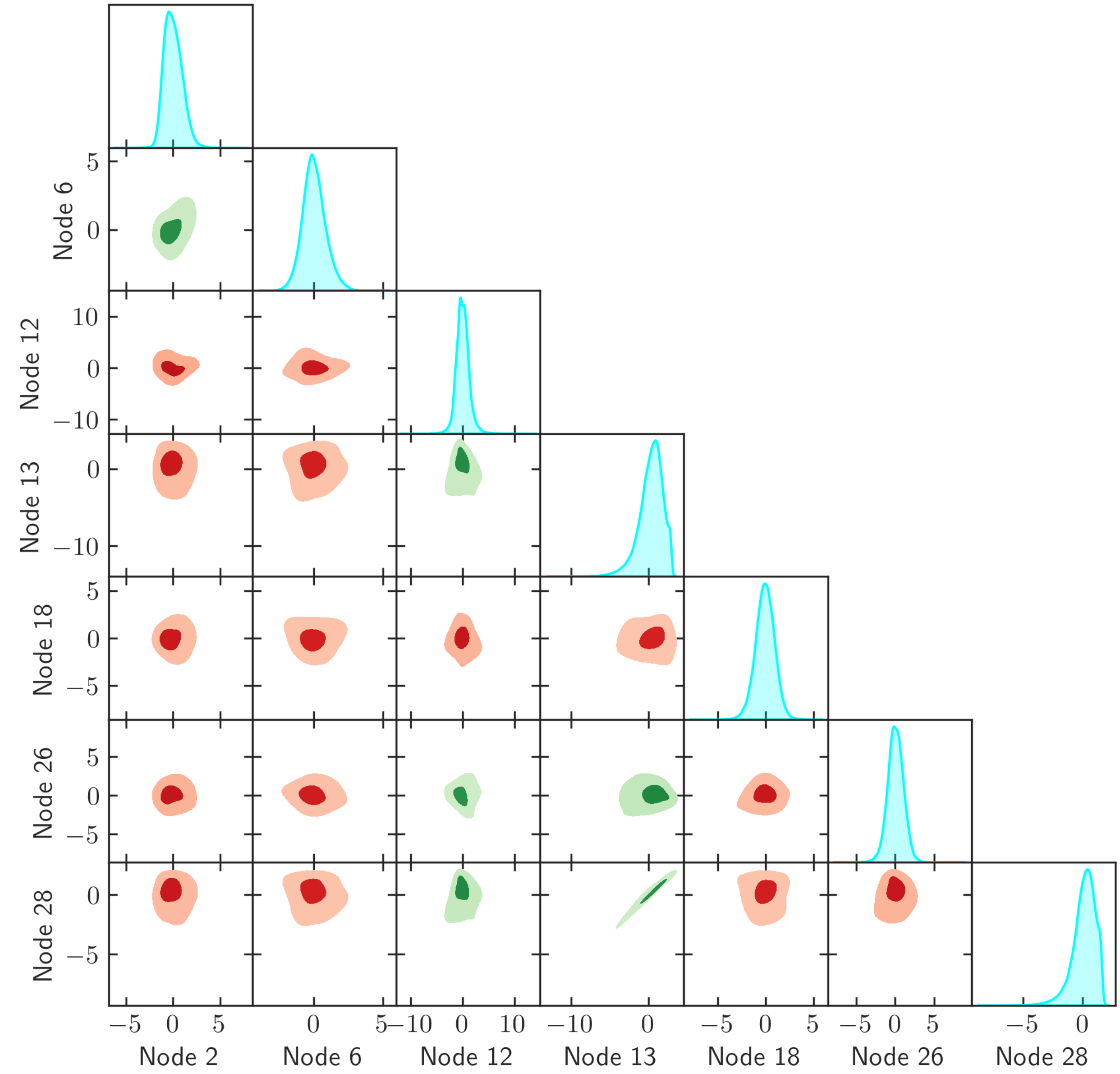

Figure 11: Density plots showing the mutual behaviour of the informative dimensions. Relations with an MI greater than 0.1 nat are shown in green. Nodes 13 and 28 show a strong linear correlation.

decoded spectra. While traversing the latent space in a meaningful manner is a non-trivial task (see e.g. Zuo et al. 2018; Song et al. 2023), we take the direct approach of varying one latent dimension at a time with the others fixed (Portillo et al. 2020; Sedaghat et al. 2021; Lucie-Smith et al. 2022). For this, we get the latent representation of a spectrum, select the dimension we are interested in, and traverse its value. The generative capabilities of the VAE are then utilised to decode the new latent values into new spectra. We traverse from the original value of the latent dimension towards the minimum and maximum values found in the full set of latent representations, stopping early if the decoded spectrum has flux values below zero. Of course, we do not expect the decoded

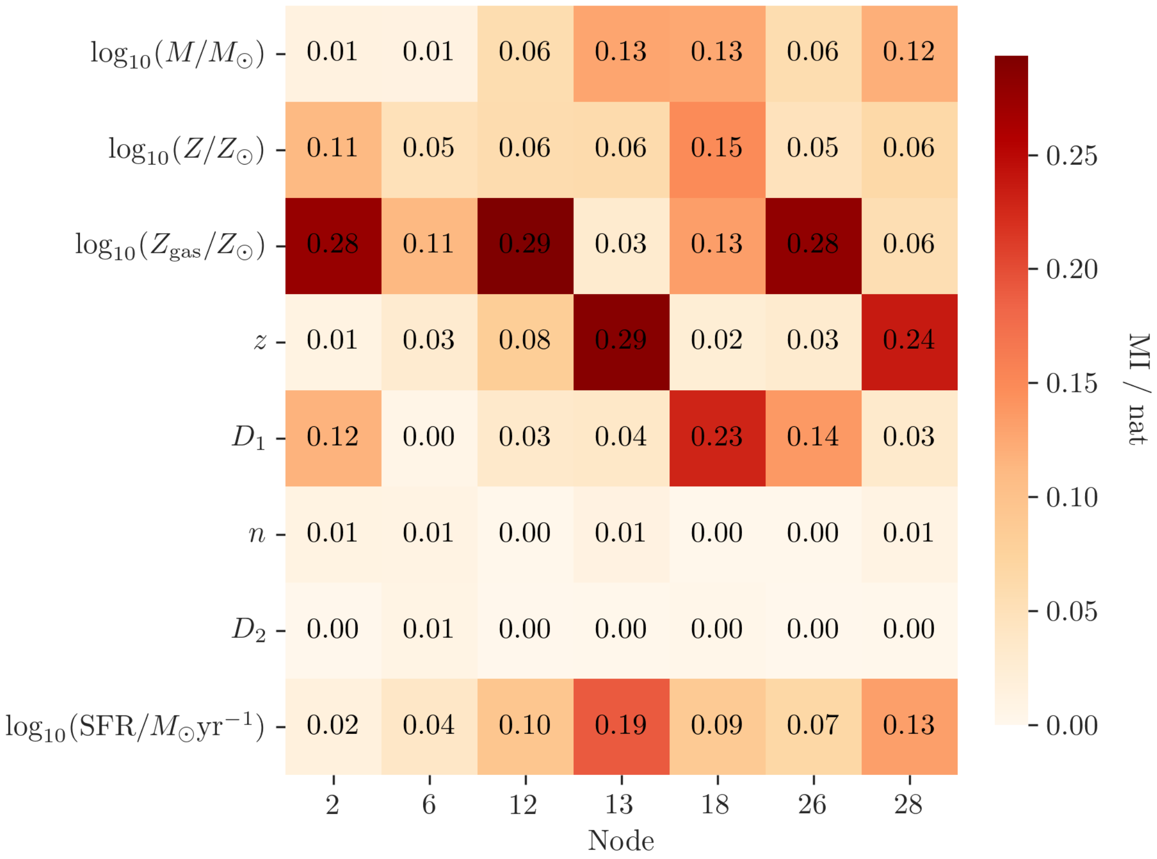

Figure 12: The results of our MI estimations between the latent dimensions and the physical parameters shown to two decimal places. We do indeed see some dimensions conveying information about the physical parameters.

spectra to necessarily be physical, but we find that restricting the traversal to positive flux values makes the interpretation easier. In Figure 14 we display the results for a spectrum. The subsequent discussion will be referring back to the theory in sections 2.1 and 2.2 and our results in section 4.3.

As shown in Figure 12, node 2 had the strongest relation with the gas metallicity, with weaker notions of the stellar metallicity and the dust attenuation of young stellar light. These parameters are expected to all show a slight positive correlation, so it makes sense that the node would learn some idea of all three. Interestingly, increasing the value of the node roughly decreases the gas and stellar metallicities, but roughly increases the dust parameter. The interplay of these parameters leads to a non-trivial traversal. Higher values of the node largely lead to some less pronounced emission lines, reflective of the lower metallicities. We also observe that the lowest values of the node have higher intensities at longer wavelengths than the highest. This can be attributed to higher gas metallicity environments producing more dust grains, leading to thermal emission at these wavelengths. Finally, we see that increasing the value of the node leads to lower intensities at the UV wavelengths. Young stars emit mostly in the UV, so it seems logical that increasing the attenuation of this light will lead to the observed effect.

Node 6 only shows a weak relation with the gas metallicity. Higher values of the node somewhat decrease the value of this parameter. Varying the node only leads to small changes in the decoded

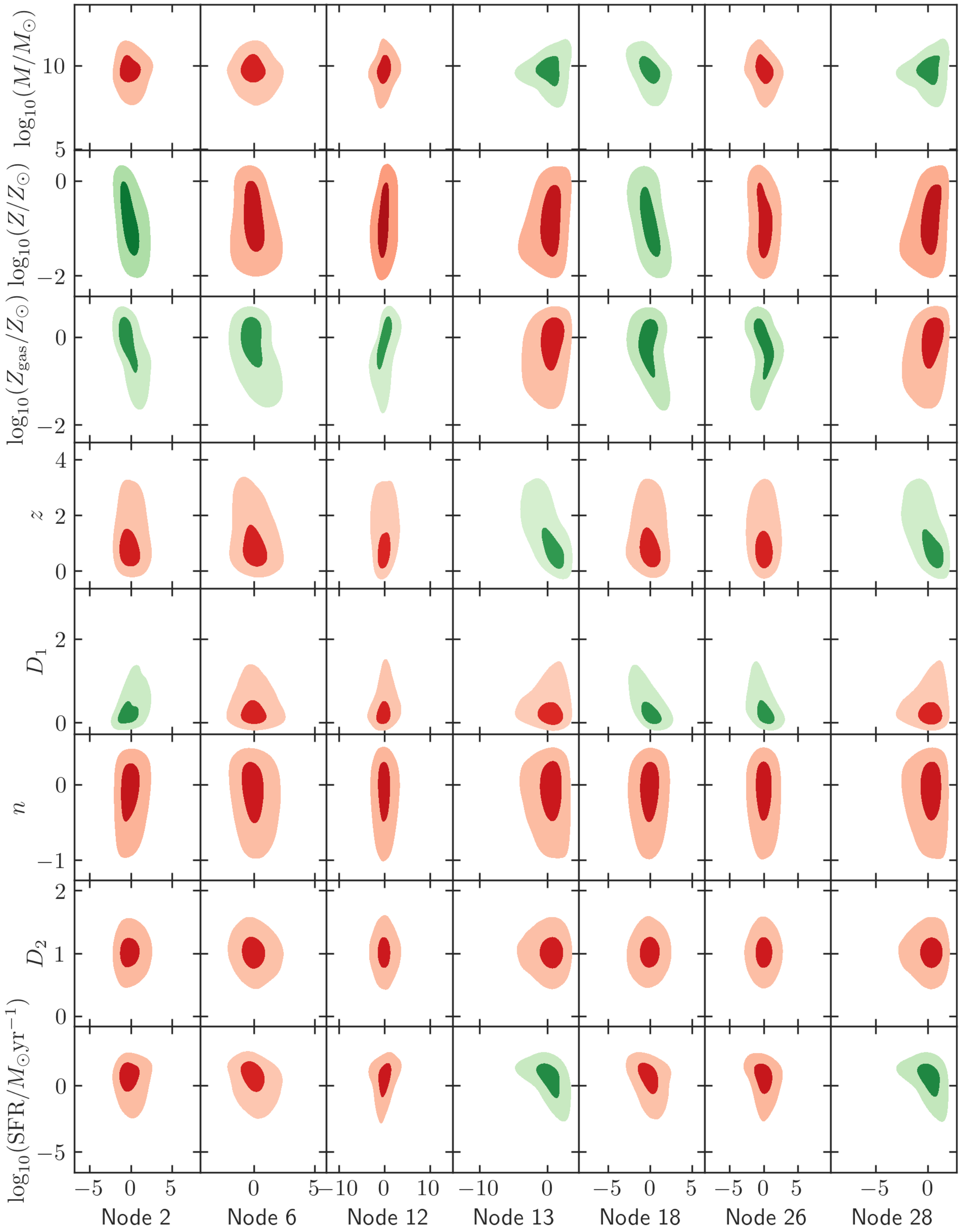

Figure 13: Density plots showing the notions of physical parameters picked up by the informative dimensions. Relations with an MI greater than 0.1 nat are again shown in green. We see some dimensions picking up weak relationships with the physical parameters.

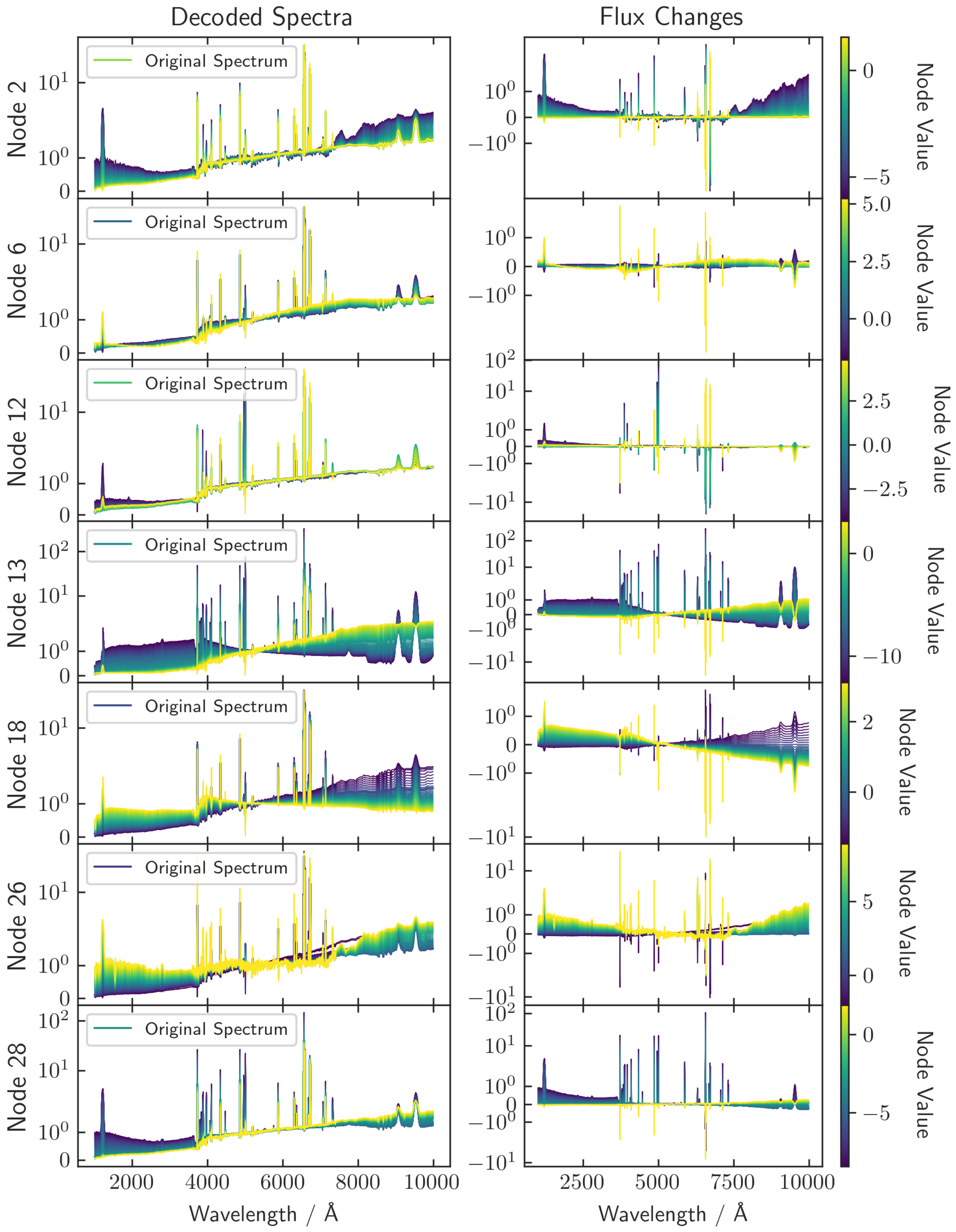

Figure 14: The results of our latent space traversal for a spectrum. The left column shows how the spectra change and the right column quantifies how much each data point changes.

spectrum, with it appearing to largely encode information about the continuum. This is an example of the model learning features that are the result of a complicated web of physical processes, rather than something that can be directly explained by a small number of parameters. It is perhaps no surprise that this node does not contain clear information about the generative factors of the spectrum, yet still encodes important information about its shape.

Increasing node 12 roughly increases the gas metallicity, with no clear relations with the other parameters. In the latent space traversal we observe that the most profound effects are on the strength of the emission lines. This indeed a feature that has some dependency on the gas metallicity. Similarly to node 6, this appears to be the model learning distinct features of the input data, but in this case there is a clearer link to a physical parameter.

In sections 4.2 and 4.3 we have highlighted that nodes 13 and 28 appear not to be disentangled. Our traversal of these nodes presents the final piece of evidence of this being the case, as their traversals affected nearly identical parts of the spectrum in the same way. The major difference is that node 13 had more pronounced effects on the continuum, making the results easier to interpret. For this reason, we restrict our discussion to this node, but remind the reader that the same points apply to node 28. Node 13 encoded information about the redshift, SFR, and stellar mass, with the strength of the relations decreasing in that order. These parameters are all expected to show a positive correlation. The node has a negative correlation with the redshift, and a weaker negative correlation with the SFR. There is no clear correlation with the stellar mass, but our MI calculations showed a non-negligible result. Our spectra are normalised and in the rest-frame, meaning that the relations with the redshift and stellar mass are likely due to their correlations with the SFR, which will be clearly reflected in our data. Traversal of the node in the positive direction leads to a consistent decrease in the emission lines and a shift in the continuum from more pronounced UV regions to more pronounced IR regions. This is in-line with a decreased SFR, as many of the physical processes that lead to these features are dependent on the presence of hot young stars. A lower SFR suggests a lesser presence of said stars. The fact that increasing the value of the node leads to a decrease in all the emission lines is important. This is in contrast to nodes 2, 6, and 12, where we saw a mixture of increases and decreases while traversing in a given direction. These nodes picked up notions of the metallicities, which have a particular effect on the ratios of emission lines. SFR only affects the intensities, leading to the consistent changes across the spectrum that we observe. Traversal of the node can be interpreted as the transition from a star-forming to quiescent galaxy.

Interestingly, traversal of node 18 also seems to show characteristics of a transition from a star-forming to quiescent galaxy, although in the opposite direction to nodes 13 and 28. This conclusion is drawn from the decrease in intensities in the UV and increase in intensities in the IR. This is interesting as the node only picked up a very weak relation with the SFR. The node did pick up a notion of the dust attenuation of young stellar light, as well as slightly weaker notions of the stellar mass, stellar metallicity, and the gas metallicity. This suggests that this apparent transition is the consequence of the age-metallicity degeneracy. The relations the model picked up all show a mostly negative correlation. This consistency makes for a smoother traversal when compared with node 2, which picked up more muddled notions of all these parameters, except the stellar mass. The physical reasons behind the traversal are the same as for node 2, but displayed in a clearer way.

Node 26 picked up its strongest relation with the gas metallicity, although a particularly non-linear one. A weak negative correlation with the dust attenuation of young stellar light is also present. This correlation leads to the expected observation that higher values of the node lead to greater intensities in the UV region. However, the non-linearity of the relation with the gas metallicity makes interpreting the other observations more difficult. Despite this, traversal shows a smooth increase in the IR intensities, as well as a consistent increase in the emission line strengths.

This is likely the model learning features of the input data purely based on shape, rather than something with a clear physical link.

5 Conclusion

We trained a deep convolutional neural network to learn compressed representations of synthetic galaxy spectra using the variational autoencoder framework. The network was trained using only the input data and chose what features were important to learn in order to convey information about the spectra. Out of 32 dimensions in the compressed representations, 7 were found to contain significant information about the input data. Two of these, however, were found to be strongly correlated, suggesting that the model failed to disentangle them. We then explored these informative dimensions by first searching for notions of physical parameters. By using the mutual information metric, we were able to form a quantitative link between what the model had learnt and the generative parameters of the spectra, which encode information about the complex web of underlying physical relations. Despite not seeing the physical parameters during training, the model was able to pick up basic notions of them. This is quite remarkable, and supports the idea that these types of models can extract physics from observational data.

We then further explored the learnt representations by exploiting the generative capabilities of the variational autoencoder to observe directly how the spectra change as we vary the values of the informative dimensions in the representations. The variation of some of these dimensions provided clear links with underlying physics, while some changed the shape of the spectra in a way that could not be directly explained. We reiterate that the model was trained in a completely physics-independent process, so the failure of a dimension to capture something physically-interpretable is not an inherently bad result. In fact, such dimensions may be the most interesting for facilitating novel discovery. The constituent components of galaxy spectra are very complicated. Analysis of what a model like this chooses to learn could be the basis for machine-assisted discovery of new scientific understanding.

While the model did pick up notions of the physical parameters, they were quite weak compared with those in Sedaghat et al. (2021). Although their work is concerned with stellar spectra, which are much simpler, it's clear that future work on full galaxy spectra may require a revision of the model used. Alterations of the hyperparameters, particularly higher β (8), did not significantly improve the strength of the learnt relations. To achieve a model effective enough for robust use in the future, a potential avenue for exploration could be in the proposed improvements over the standard variational autoencoder framework (e.g. Burda et al. 2016; Hou et al. 2019; Zhang et al. 2022; Wang et al. 2024).

References

Alsing, J., Thorp, S., Deger, S., Peiris, H., Leistedt, B., Mortlock, D., and Leja, J. (2024). pop-cosmos: A comprehensive picture of the galaxy population from cosmos data.

Andriushchenko, M., D'Angelo, F., Varre, A., and Flammarion, N. (2023). Why do we need weight decay in modern deep learning?

Angls-Alczar, D., Faucher-Gigure, C.-A., Kere, D., Hopkins, P. F., Quataert, E., and Murray, N. (2017). The cosmic baryon cycle and galaxy mass assembly in the fire simulations. *Monthly Notices of the Royal Astronomical Society*, 470(4):46984719.

Arachchige, C. N. P. G. and Prendergast, L. A. (2019). Confidence intervals for median absolute deviations.

Asano, R. S., Takeuchi, T. T., Hirashita, H., and Inoue, A. K. (2013). Dust formation history of galaxies: A critical role of metallicity dust mass growth by accreting materials in the interstellar medium. *Earth, Planets and Space*, 65(3):213222.

Bank, D., Koenigstein, N., and Giryes, R. (2021). Autoencoders.

Baron, D. (2019). Machine learning in astronomy: a practical overview.

Bengio, Y., Courville, A., and Vincent, P. (2014). Representation learning: A review and new perspectives.

Brinchmann, J., Charlot, S., White, S. D. M., Tremonti, C., Kauffmann, G., Heckman, T., and Brinkmann, J. (2004). The physical properties of star-forming galaxies in the low-redshift universe. *Monthly Notices of the Royal Astronomical Society*, 351(4):11511179.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-vae.

Calzetti, D. (2001). The dust opacity of starforming galaxies. *Publications of the Astronomical Society of the Pacific*, 113(790):14491485.

Chen, Y., Mancini, M., Zhu, X., and Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey.

Collaboration, R. O. L. S. S. S., Jones, R. L., Bannister, M. T., Bolin, B. T., Chandler, C. O., Chesley, S. R., Eggl, S., Greenstreet, S., Holt, T. R., Hsieh, H. H., Ivezi, Z., Juri, M., Kelley, M. S. P., Knight, M. M., Malhotra, R., Oldroyd, W. J., Sarid, G., Schwamb, M. E., Snodgrass, C., Solontoi, M., and Trilling, D. E. (2020). The scientific impact of the vera c. rubin observatory's legacy survey of space and time (lsst) for solar system science.

Conroy, C. (2013). Modeling the panchromatic spectral energy distributions of galaxies. *Annual Review of Astronomy and Astrophysics*, 51(1):393455.

Conroy, C. and Gunn, J. E. (2010). The propagation of uncertainties in stellar population synthesis modeling. iii. model calibration, comparison, and evaluation. *The Astrophysical Journal*, 712(2):833857.

Conroy, C., Gunn, J. E., and White, M. (2009). The propagation of uncertainties in stellar population synthesis modeling. i. the relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *The Astrophysical Journal*, 699(1):486506.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

de Andres, D., Yepes, G., Sembolini, F., Martnez-Muoz, G., Cui, W., Robledo, F., Chuang, C.-H., and Rasia, E. (2022). Machine learning methods to estimate observational properties of galaxy clusters in large volume cosmological n-body simulations. *Monthly Notices of the Royal Astronomical Society*, 518(1):111129.

Dey, A., Schlegel, D. J., Lang, D., Blum, R., Burleigh, K., Fan, X., Findlay, J. R., Finkbeiner, D., Herrera, D., Juneau, S., Landriau, M., Levi, M., McGreer, I., Meisner, A., Myers, A. D., Moustakas, J., Nugent, P., Patej, A., Schlafly, E. F., Walker, A. R., Valdes, F., Weaver, B. A., Yche, C., Zou, H., Zhou, X., Abareshi, B., Abbott, T. M. C., Abolfathi, B., Aguilera, C., Alam, S., Allen, L., Alvarez, A., Annis, J., Ansarinejad, B., Aubert, M., Beechert, J., Bell, E. F., BenZvi, S. Y., Beutler, F., Bielby, R. M., Bolton, A. S., Briceo, C., Buckley-Geer, E. J., Butler, K., Calamida, A., Carlberg, R. G., Carter, P., Casas, R., Castander, F. J., Choi, Y., Comparat, J., Cukanovaite, E., Delubac, T., DeVries, K., Dey, S., Dhungana, G., Dickinson, M., Ding, Z., Donaldson, J. B., Duan, Y., Duckworth, C. J., Eftekharzadeh, S., Eisenstein, D. J., Etourneau, T., Fagrelius, P. A., Farihi, J., Fitzpatrick, M., Font-Ribera, A., Fulmer, L., Gnsicke, B. T., Gaztanaga, E., George, K., Gerdes, D. W., A Gontcho, S. G., Gorgoni, C., Green, G., Guy, J., Harmer, D., Hernandez, M., Honscheid, K., Huang, L. W., James, D. J., Jannuzi, B. T., Jiang, L., Joyce, R., Karcher, A., Karkar, S., Kehoe, R., Kneib, J.-P., Kueter-Young, A., Lan, T.-W., Lauer, T. R., Guillou, L. L., Van Suu, A. L., Lee, J. H., Lesser, M., Levasseur, L. P., Li, T. S., Mann, J. L., Marshall, R., Martnez-Vzquez, C. E., Martini, P., du Mas des Bourboux, H., McManus, S., Meier, T. G., Mnard, B., Metcalfe, N., Muoz-Gutirrez, A., Najita, J., Napier, K., Narayan, G., Newman, J. A., Nie, J., Nord, B., Norman, D. J., Olsen, K. A. G., Paat, A., Palanque-Delabrouille, N., Peng, X., Poppett, C. L., Poremba, M. R., Prakash, A., Rabinowitz, D., Raichoor, A., Rezaie, M., Robertson, A. N., Roe, N. A., Ross, A. J., Ross, N. P., Rudnick, G., Safonova, S., Saha, A., Snchez, F. J., Savary, E., Schweiker, H., Scott, A., Seo, H.-J., Shan, H., Silva, D. R., Slepian, Z., Soto, C., Sprayberry, D., Staten, R., Stillman, C. M., Stupak, R. J., Summers, D. L., Tie, S. S., Tirado, H., Vargas-Magaa, M., Vivas, A. K., Wechsler, R. H., Williams, D., Yang, J., Yang, Q., Yapici, T., Zaritsky, D., Zenteno, A., Zhang, K., Zhang, T., Zhou, R., and Zhou, Z. (2019). Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459.

Fissler, T. and Pohle, M.-O. (2023). Generalised covariances and correlations.

Fraser-McKelvie, A., Cortese, L., Groves, B., Brough, S., Bryant, J., Catinella, B., Croom, S., DEugenio, F., Lpez-Snchez, . R., van de Sande, J., Sweet, S., Vaughan, S., Bland-Hawthorn, J., Lawrence, J., Lorente, N., and Owers, M. (2021). The sami galaxy survey: the drivers of gas and stellar metallicity differences in galaxies. *Monthly Notices of the Royal Astronomical Society*, 510(1):320333.

Frster Schreiber, N. M. and Wuyts, S. (2020). Star-forming galaxies at cosmic noon. *Annual Review of Astronomy and Astrophysics*, 58(1):661725.

Garn, T. and Best, P. N. (2010). Predicting dust extinction from the stellar mass of a galaxy: Dust extinction and stellar mass. *Monthly Notices of the Royal Astronomical Society*, 409(1):421432.

Girelli, G., Bolzonella, M., and Cimatti, A. (2019). Massive and old quiescent galaxies at high redshift. *Astronomy & Astrophysics*, 632:A80.

Glazebrook, K., Blake, C., Economou, F., Lilly, S., and Colless, M. (1999). Measurement of the star formation rate from ha in field galaxies at z=1. *Monthly Notices of the Royal Astronomical Society*, 306(4):843856.

Guezenoc, J., Gallet-Budynek, A., and Bousquet, B. (2019). Critical review and advices on spectral-based normalization methods for LIBS quantitative analysis. *Spectrochimica Acta*, 160:105688.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Hinton, G. E. and Zemel, R. S. (1993). Autoencoders, minimum description length and helmholtz free energy. In *Neural Information Processing Systems*.

Hou, X., Sun, K., Shen, L., and Qiu, G. (2019). Improving variational autoencoder with deep feature consistent and generative adversarial training. *Neurocomputing*, 341:183194.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., and Shao, L. (2020). Normalization techniques in training dnns: Methodology, analysis and application.

Iten, R., Metger, T., Wilming, H., del Rio, L., and Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1).

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

Kennicutt, Jr., R. C. (1998). The global schmidt law in starforming galaxies. *The Astrophysical Journal*, 498(2):541552.

Kettlety, T., Hesling, J., Phillipps, S., Bremer, M. N., Cluver, M. E., Taylor, E. N., Bland-Hawthorn, J., Brough, S., De Propris, R., Driver, S. P., Holwerda, B. W., Kelvin, L. S., Sutherland, W., and Wright, A. H. (2017). Galaxy and mass assembly (gama): the consistency of gama and wise derived mass-to-light ratios. *Monthly Notices of the Royal Astronomical Society*, 473(1):776783.

Kewley, L. J., Nicholls, D. C., and Sutherland, R. S. (2019). Understanding galaxy evolution through emission lines. *Annual Review of Astronomy and Astrophysics*, 57(1):511570.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307392.

Kriek, M., van Dokkum, P. G., Franx, M., Forster Schreiber, N. M., Gawiser, E., Illingworth, G. D., Labbe, I., Marchesini, D., Quadri, R., Rix, H., Rudnick, G., Toft, S., van der Werf, P., and Wuyts, S. (2006). Direct measurements of the stellar continua and balmer/4000 a breaks of red z¿2 galaxies: Redshifts and improved constraints on stellar populations. *The Astrophysical Journal*, 645(1):4454.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lanusse, F., Parker, L., Golkar, S., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Ohana, R., Pettee, M., Blancard, B. R.-S., Tesileanu, T., Cho, K., and Ho, S. (2023). Astroclip: Cross-modal pre-training for astronomical foundation models.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Leja, J., Speagle , J. S. ., Ting , Y.-S. ., Johnson, B. D., Conroy, C., Whitaker, K. E., Nelson, E. J., Dokkum, P. v., and Franx, M. (2022). A new census of the 0.2 ¡ z ¡ 3.0 universe. ii. the star-forming sequence. *The Astrophysical Journal*, 936(2):165.

Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., and Jiang, Z. M. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. *ArXiv*, abs/2401.10359.

Liang, Y., Melchior, P., Hahn, C., Shen, J., Goulding, A., and Ward, C. (2023a). Outlier detection in the desi bright galaxy survey.

Liang, Y., Melchior, P., Lu, S., Goulding, A., and Ward, C. (2023b). Autoencoding galaxy spectra. ii. redshift invariance and outlier detection. *The Astronomical Journal*, 166(2):75.

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., and Wei, W. (2024). Machine learning for synthetic data generation: A review.

Lucie-Smith, L., Peiris, H. V., and Pontzen, A. (2024). Explaining dark matter halo density profiles with neural networks. *Physical Review Letters*, 132(3).

Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., Thiyagalingam, J., and Piras, D. (2022). Discovering the building blocks of dark matter halo density profiles with neural networks. *Physical Review D*, 105(10).

Madau, P. and Dickinson, M. (2014). Cosmic star-formation history. *Annual Review of Astronomy and Astrophysics*, 52(1):415486.

Masters, D. and Luschi, C. (2018). Revisiting small batch training for deep neural networks.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders.

Melchior, P., Liang, Y., Hahn, C., and Goulding, A. (2023). Autoencoding galaxy spectra. i. architecture. *The Astronomical Journal*, 166(2):74.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comput.*, 15(6):11911253.

Panter, B., Jimenez, R., Heavens, A. F., and Charlot, S. (2008). The cosmic evolution of metallicity from the sdss fossil record. *Monthly Notices of the Royal Astronomical Society*, 391(3):11171126.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

Piras, D., Peiris, H. V., Pontzen, A., Lucie-Smith, L., Guo, N., and Nord, B. (2023). A robust estimator of mutual information for deep learning interpretability. *Machine Learning: Science and Technology*, 4(2):025006.

Plaut, D. C., Nowlan, S. J., and Hinton, G. E. (1986). Experiments on learning by back propagation.

Poggianti, B. M. and Barbaro, G. (1997). Indicators of star formation: 4000 angstrom break and balmer lines.

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.

Proux, C. and Howk, J. C. (2020). The cosmic baryon and metal cycles. *Annual Review of Astronomy and Astrophysics*, 58(1):363406.

Qin, J., Zheng, X. Z., Wuyts, S., Lyu, Z., Qiao, M., Huang, J.-S., Liu, F. S., Katsianis, A., Gonzalez, V., Bian, F., Xu, H., Pan, Z., Liu, W., Tan, Q.-H., An, F. X., Shi, D. D., Zhang, Y., Wen, R., Liu, S., and Yang, C. (2023). Understanding the universal dust attenuation scaling relation of star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 528(1):658675.

Renard, P., Siudek, M., Eriksen, M. B., Cabayol, L., Cai, Z., Carretero, J., Casas, R., Castander, F. J., Fernandez, E., Garca-Bellido, J., Gaztanaga, E., Hoekstra, H., Joachimi, B., Miquel, R., Navarro-Girones, D., Padilla, C., Sanchez, E., Serrano, S., Tallada-Cresp, P., De Vicente, J., Wittje, A., and Wright, A. H. (2022). The pau survey: measurements of the 4000 spectral break with narrow-band photometry. *Monthly Notices of the Royal Astronomical Society*, 515(1):146166.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Salim, S. and Narayanan, D. (2020). The dust attenuation law in galaxies. *Annual Review of Astronomy and Astrophysics*, 58(1):529575.

Salim, S., Rich, R. M., Charlot, S., Brinchmann, J., Johnson, B. D., Schiminovich, D., Seibert, M., Mallery, R., Heckman, T. M., Forster, K., Friedman, P. G., Martin, D. C., Morrissey, P., Neff, S. G., Small, T., Wyder, T. K., Bianchi, L., Donas, J., Lee, Y., Madore, B. F., Milliard, B., Szalay, A. S., Welsh, B. Y., and Yi, S. K. (2007). Uv star formation rates in the local universe. *The Astrophysical Journal Supplement Series*, 173(2):267292.

Santini, P., Fontana, A., Castellano, M., Criscienzo, M. D., Merlin, E., Amorin, R., Cullen, F., Daddi, E., Dickinson, M., Dunlop, J. S., Grazian, A., Lamastra, A., McLure, R. J., Michaowski, M. J., Pentericci, L., and Shu, X. (2017). The star formation main sequence in the hubble space telescope frontier fields. *The Astrophysical Journal*, 847(1):76.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. (2020). Fairness by learning orthogonal disentangled representations.

Schady, P., Yates, R., Christensen, L., Cia, A. D., DElia, A. R. V., Heintz, K., Jakobsson, P., Laskar, T., Levan, A., Salvaterra, R., Starling, R., Tanvir, N., Thne, C., Vergani, S., Wiersema, K., Arabsalmani, M., Chen, H.-W., Pasquale, M. D., Fruchter, A., Fynbo, J., Garca-Benito, R., Gompertz, B., Hartmann, D., Kouveliotou, C., Milvang-Jensen, B., Palazzi, E., Perley, D., Piranomonte, S., Pugliese, G., Savaglio, S., Sbarufatti, B., Schulze, S., Tagliaferri, G., de Ugarte Postigo, A., Watson, D., and Wiseman, P. (2024). Comparing emission- and absorption-based gas-phase metallicities in GRB host galaxies at z = 2 4 using JWST.

Schiminovich, D., Wyder, T. K., Martin, D. C., Johnson, B. D., Salim, S., Seibert, M., Treyer, M., Budavári, T., Hoopes, C. G., Zamojski, M., Barlow, T. A., Forster, K., Friedman, P. G., Morrissey, P., Neff, S. G., Small, T. A., Bianchi, L., Donas, J., Heckman, T. M., Lee, Y.-W., Madore, B. F., Milliard, B., Rich, R. M., Szalay, A. S., Welsh, B. Y., and Yi, S. K. (2007). The uv-optical color magnitude diagram. ii. physical properties and morphological evolution on and off of a star-forming sequence. *The Astrophysical Journal Supplement Series*, 173:315 – 341.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126:17631768.

Sedaghat, N., Romaniello, M., Carrick, J. E., and Pineau, F.-X. (2021). Machines learn to infer stellar parameters just by looking at a large number of spectra. *Monthly Notices of the Royal Astronomical Society*, 501(4):60266041.

Song, Y., Keller, T. A., Sebe, N., and Welling, M. (2023). Latent traversals in generative models as potential flows.

Steinhardt, C. L., Weaver, J. R., Maxfield, J., Davidzon, I., Faisst, A. L., Masters, D., Schemel, M., and Toft, S. (2020). A method to distinguish quiescent and dusty star-forming galaxies with machine learning. *The Astrophysical Journal*, 891(2):136.

Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., and Bluck, A. F. L. (2022). Mapping the diversity of galaxy spectra with deep unsupervised machine learning. *The Astronomical Journal*, 163(2):71.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method.

Tremonti, C. A., Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., White, S. D. M., Seibert, M., Peng, E. W., Schlegel, D. J., Uomoto, A., Fukugita, M., and Brinkmann, J. (2004). The origin of the massmetallicity relation: Insights from 53,000 starforming galaxies in the sloan digital sky survey. *The Astrophysical Journal*, 613(2):898913.

Tschannen, M., Bachem, O., and Lucic, M. (2018). Recent advances in autoencoder-based representation learning.

Vergara, J. R. and Estvez, P. A. (2013). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175186.

Walmsley, M., Ferguson, A. M. N., Mann, R. G., and Lintott, C. J. (2018). Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 483(3):29682982.

Wang, D., Wang, Y., Evans, L., and Tiwary, P. (2024). From latent dynamics to meaningful representations.

Worthey, G. (1994). Comprehensive Stellar Population Models and the Disentanglement of Age and Metallicity Effects. *Astrophysical Journal Supplement*, 95:107.

Worthey, G. (1999). The Age-Metallicity Degeneracy. In Hubeny, I., Heap, S., and Cornett, R., editors, *Spectrophotometric Dating of Stars and Galaxies*, volume 192 of *Astronomical Society of the Pacific Conference Series*, page 283.

Wu, Y., Liu, L., Bae, J., Chow, K.-H., Iyengar, A., Pu, C., Wei, W., Yu, L., and Zhang, Q. (2019). Demystifying learning rate policies for high accuracy training of deep neural networks.

Yang, R. (2022). Unsupervised machine learning for physical concepts.

Zahid, H. J., Yates, R. M., Kewley, L. J., and Kudritzki, R. P. (2013). The observed relation between stellar mass, dust extinction, and star formation rate in local galaxies. *The Astrophysical Journal*, 763(2):92.

Zhang, M., Xiao, T. Z., Paige, B., and Barber, D. (2022). Improving vae-based representation learning.

Zuo, Y., Avraham, G., and Drummond, T. (2018). Traversing latent space using decision ferns.