The Astronomical Journal, 160:45 (17pp), 2020 July https://doi.org/10.3847/1538-3881/ab9644

Dimensionality Reduction of SDSS Spectra with Variational Autoencoders

Stephen K. N. Portillo[1], John K. Parejko[1], Jorge R. Vergara[2,3], and Andrew J. Connolly[1]

[1] DIRAC Institute, Department of Astronomy, University of Washington, 3910 15th Ave. NE, Seattle, WA 98195, USA; sportill@uw.edu

[2] Department of Computing, Universidad Tecnológica Metropolitana, Santiago, Chile

[3] Millennium Institute of Astrophysics, Santiago, Chile

Received 2020 February 22; revised 2020 May 15; accepted 2020 May 21; published 2020 June 29

Abstract

High-resolution galaxy spectra contain much information about galactic physics, but the high dimensionality of these spectra makes it difficult to fully utilize the information they contain. We apply variational autoencoders (VAEs), a nonlinear dimensionality reduction technique, to a sample of spectra from the Sloan Digital Sky Survey (SDSS). In contrast to principal component analysis (PCA), a widely used technique, VAEs can capture nonlinear relationships between latent parameters and the data. We find that a VAE can reconstruct the SDSS spectra well with only six latent parameters, outperforming PCA with the same number of components. Different galaxy classes are naturally separated in this latent space, without class labels having been given to the VAE. The VAE latent space is interpretable because the VAE can be used to make synthetic spectra at any point in latent space. For example, making synthetic spectra along tracks in latent space yields sequences of realistic spectra that interpolate between two different types of galaxies. Using the latent space to find outliers may yield interesting spectra: in our small sample, we immediately find unusual data artifacts and stars misclassified as galaxies. In this exploratory work, we show that VAEs create compact, interpretable latent spaces that capture nonlinear features of the data. While a VAE takes substantial time to train (≈1 day for 48,000 spectra), once trained, VAEs can enable the fast exploration of large astronomical data sets.

Unified Astronomy Thesaurus concepts: Galaxies (573); Galaxy classification systems (582); Astroinformatics (78); Neural networks (1933); Dimensionality reduction (1943); Spectroscopy (1558)

1. Introduction

The galaxy spectra in the Sloan Digital Sky Survey (SDSS) contain a wealth of information about the physical processes occurring in galaxies, but to fully make use of these spectra, methods are needed that can handle both the complexity of galaxy spectra and the vast number of available spectra. Spectra consist of thousands of flux measurements that depend on galaxy properties in complex ways. To investigate physical trends in galaxies, scientists must extract statistics from each spectrum that are easily related to the physical properties of the galaxy. In other words, scientists reduce the dimensionality of the data in a way that enables them to probe the galaxies' properties.

Spectra can be fit with theoretical or semi-analytic models, yielding physically interpretable parameters for each spectrum. This approach has the advantage of directly connecting the observations to a physical model. Incomplete models can, however, introduce biases to the inferred parameters or miss interesting behavior not included in the model.

To supplement model-based interpretation of the data, empirical methods are desirable. One approach is to focus on parts of the spectrum that are known to be physically significant, like emission lines. Line ratio tests (Baldwin et al. 1981; Osterbrock & de Robertis 1985; Kewley et al. 2001, 2006, 2019) leverage the sensitivity of emission lines to ionization levels while minimizing the effect of foreground dust extinction. These tests are very useful in classifying star-forming galaxies and active galactic nuclei (AGNs), but ignore the information present in the continuum of the spectrum.

Another approach is to use dimensionality reduction or classification techniques that extract useful features from the data without any prior astrophysical knowledge. These methods include principal component analysis (PCA), independent component analysis (Lu et al. 2006), local linear embedding (LLE; Vanderplas & Connolly 2009), and neural net classification (Folkes et al. 1996; Ball et al. 2004). Of these methods, PCA is one of the most successful and widely used. PCA finds components that can be added in linear combinations to best approximate the input data. The coefficients of the linear combination that approximates a given spectrum can be thought of as a compression of the spectrum. Yip et al. (2004a) found that the coefficients of the first three PCA components of the SDSS spectra are sufficient to separate early-type galaxies, late-type galaxies, and extreme emission-line galaxies. They also found that the first eight PCA components are sufficient to reconstruct the continuum levels and spectral line ratios of all but the most extreme emission-line galaxies, representing a considerable reduction in dimensionality from the 3839 pixels in each spectrum. Because the PCA components are simply added together to yield spectra, each component can be interpreted. For example, the first PCA component removes blue continuum and nebular emission lines and adds continuum light, suggesting that this component is negatively correlated with star formation activity.

PCA's linearity, however, inhibits its ability to efficiently reduce the dimensionality of data when nonlinear features are present. Yip et al. (2004b) apply PCA to SDSS quasars out to $z = 5.41$ and find that 50 components are necessary to acceptably reconstruct a typical spectrum. By binning the quasars by redshift and luminosity and performing PCA separately in each bin, they can get a similar reconstruction with only 10 components. They also find that broad absorption lines cannot be captured by a single PCA component, but instead are reconstructed as the combination of many components.

There are many empirical dimensionality reduction and classification techniques that are nonlinear, giving them more flexiblity than PCA in dealing with nonlinear features. These techniques include diffusion maps (Richards et al. 2009), LLE (Vanderplas & Connolly 2009), *k*-means clustering (Almeida et al. 2010), locally-biased semi-supervised eigenvectors (Lawlor et al. 2016), self-organizing maps (Meusinger et al. 2012, 2017), and random forests (Baron & Poznanski 2017; Reis et al. 2018).

Autoencoders are a class of neural network that has been widely studied in the machine learning literature and that astronomers have started to adopt. For example, autoencoders have been used to estimate stellar atmospheric properties from spectra (Yang & Li 2015; Li et al. 2017), identify spatial structures in Tycho's supernova remnant using X-ray spectra (Iwasaki et al. 2019), morphologically classify radio AGNs (Ma et al. 2019), classify variable starlight curves (Naul et al. 2018; Tsang & Schultz 2019), and emulate cosmological simulations (Chardin et al. 2019; Tröster et al. 2019).

The goal of this work is to use variational autoencoders (VAEs) (a specific subclass of autoencoders) to address the limitations of PCA. A key feature of PCA is that it focuses on reconstruction: the PCA coefficients for an input are a low-dimensional representation that can be used to reconstruct an approximation to that input. We can then compare the reconstruction to the input to see what features the PCA components are capturing. Not only can we reconstruct observed spectra, any combination of PCA coefficients can be used to construct new, synthetic spectra. In the case of PCA, any synthetic spectra will simply be linear combinations of the PCA components. Similarly, VAEs have an encoder that returns a low-dimensional latent representation of a given input and a decoder that returns a reconstruction given that latent representation. We can use the decoder to construct a synthetic spectrum for any point in the latent space. Unlike PCA, the VAE's nonlinearity allows it to find nonlinear relationships between the latent representations and the spectra. We introduce autoencoders and VAEs in Section 2. Then in Section 3 we apply VAEs to a subset of the SDSS spectra and consider the quality of the VAE reconstructions and the intepretability of the latent representations. We discuss the advantages and challenges of using VAEs as a dimensionality reduction technique in Section 4 before concluding in Section 5.

2. Variational Autoencoders

Autoencoders are feedforward neural networks that learn efficient encodings of data in an unsupervised manner. An autoencoder consists of two parts: an encoder that takes data as input and compresses it to produce a latent representation, followed by a decoder that takes the latent representation and decompresses it to produce a reconstruction of the original data. In an undercomplete[4] autoencoder, the latent representation is lower-dimensional than the data, so the autoencoder cannot simply learn the identity function. An autoencoder that learns the identity function would not be useful; its reconstructions would simply be a blind copy of the input and it would not have learned a compact representation of the data. Instead, it must find a compression that allows the input data to be approximately reconstructed. This compression is tailored for the data set that the autoencoder is trained on, thus the latent representation often reflects meaningful properties of the data. The latent representations can be thought of as occupying a latent space; this space is the range of the encoder and domain of the decoder. PCA reconstruction can be thought of as a restricted autoencoder. In this case, the encoder projects the data onto the first n PCA components, giving a low-dimensional latent representation: the projection coefficients. The corresponding decoder then returns the linear combination of the first n PCA components with these coefficients as the reconstruction of the input. Training PCA reconstruction consists of finding the first n PCA components: of all the possible bases of n vectors, the first n PCA components are the basis that minimizes the mean-squared reconstruction error. In an autoencoder, the encoder and decoder can learn more complex operations than projection and linear combination, allowing the autoencoder to capture nonlinear features of the data. We will outline how autoencoders work in this section, but more details can be found in machine learning references such as Goodfellow et al. (2016; Chapter 14).

A feedforward neural network consists of neurons arranged in a sequence of layers, where the activation of a neuron depends only on the activations of the neurons in the previous layers. Figure 1 depicts a simple autoencoder that takes in a four-dimensional input $\boldsymbol{x}$, compresses it to a two-dimensional latent representation $\boldsymbol{z}$, and then outputs a four-dimensional reconstruction $\boldsymbol{x}'$. In our case, this autoencoder would take in spectra with four pixels and compress each to a two-dimensional representation before reconstructing the spectrum's four pixel values. This autoencoder has five layers (arranged from left to right in Figure 1) with 4, 3, 2, 3, and 4 neurons, respectively, so it has a 4-3-2-3-4 architecture. The first layer is the input layer: when one of the four-dimensional input data is fed into the autoencoder, the four input layer neurons' activations are set to the four values in the input data. The encoder consists of a single fully connected layer, meaning that each of the activations of its three neurons depends on the activations of all of the input neurons. Specifically, writing the input layer activations as a vector $\boldsymbol{x}$, the vector of encoder neuron activations $\boldsymbol{e}$ is

$$\boldsymbol{e} = f(W^{(e)}\boldsymbol{x} + \boldsymbol{b}^{(e)}) \tag{1}$$

where $W^{(e)}$ is a matrix of weights, $\boldsymbol{b}^{(e)}$ is a vector of biases, and f is the nonlinear activation function that is applied element-wise to its argument. A common choice is the rectified linear unit (ReLU)

$$\mathrm{ReLU}(y) = \begin{cases} y & y \geqslant 0 \\ 0 & y < 0 \end{cases} \tag{2}$$

because it is simple to take the derivative of, which is useful when training the network. While the activation function for a single neuron may be simple, having many neurons in a layer and many layers in the network gives the neural network considerable expressive power. The encoder neuron activations for a given input can be interpreted as a partially compressed representation of that input. The next layer, the latent layer, is fully connected to the encoder layer, with the latent neuron

[4] There are autoencoder architectures that do not require the latent space to be lower-dimensional than the data, but we focus on undercomplete autoencoders in this work.

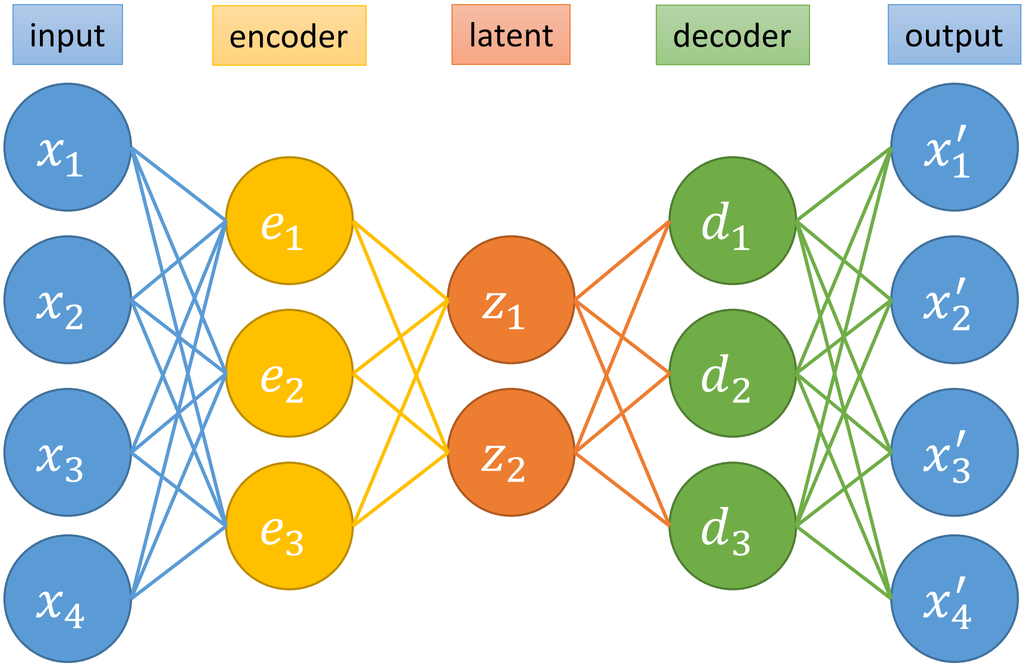

Figure 1. Structure of an autoencoder with a 4-3-2-3-4 architecture. It compresses four-dimensional inputs to a two-dimensional latent space and both the encoder and decoder are single layers with three neurons.

activations being given by

$$z = W^{(z)}\boldsymbol{e} + \boldsymbol{b}^{(z)} \tag{3}$$

where $W^{(z)}$ is a matrix of weights, $\boldsymbol{b}^{(z)}$ is a vector of biases, and there is no activation function. We note that the weights and biases can be different for each layer. The latent neuron activations for a given input comprise the autoencoder's most compact representation of that input. The latent neuron activations are analogous to the PCA coefficients, which are the compact representation that PCA produces.

The decoder is also a single layer of three neurons, fully connected to the latent layer, with their activations being given by

$$\boldsymbol{d} = f(W^{(d)}\boldsymbol{z} + \boldsymbol{b}^{(d)}) \tag{4}$$

where $W^{(d)}$ is a matrix of weights, $\boldsymbol{b}^{(d)}$ is a vector of biases, and f is the activation function. The decoder neuron activations can be interpreted as a partially decompressed version of the latent neuron activations. Finally, the output reconstruction is given by the activation of the last layer:

$$\boldsymbol{x}' = W^{(x')}\boldsymbol{d} + \boldsymbol{b}^{(x')} \tag{5}$$

where $W^{(x')}$ is a matrix of weights, $\boldsymbol{b}^{(x')}$ is a vector of biases, and there is no activation function. The output neuron activations are the autoencoder's reconstruction of the input data. Each neuron in this layer maps one-to-one with a neuron in the input layer, and thus with one of the dimensions of the data (one of the spectrum pixels, in our case).

The autoencoder is trained to minimize a defined reconstruction loss which decreases when the output is closer to the input, like the χ^2 test statistic between the output and input. That is, the weights $W^{(e,z,d,x')}$ and biases $\boldsymbol{b}^{(e,z,d,x')}$ are iteratively changed to decrease the average reconstruction loss on some training data set. Finding the optimal weights and biases is a high-dimensional and nonlinear optimization problem that requires specialized optimizers. Optimizers often make use of the gradient of the reconstruction loss with respect to the weights and biases. Simple activation functions like ReLU make these gradients fast to calculate. Often, the gradient of the average loss is not calculated on the entire training data set, but on random subsets of the training set called mini-batches. This subsampling adds some stochasticisty to the gradient, which can help the optimizer avoid local minima. Also, optimizers often allow step sizes to be adaptively changed for each weight and bias during training. One popular optimizer, Adam (Kingma & Ba 2015), keeps a moving average of the gradient and gradient squared to determine good step sizes as training progresses. In finding a combination of weights and biases that minimizes the reconstruction error, the autoencoder is implicitly learning a dimensionality reduction of the data. These weights and biases define the encoder that maps input data to a lower-dimensional latent representation as well as the decoder that produces a reconstruction from a latent representation.

In a VAE (Kingma & Welling 2013), inputs get mapped onto a distribution in latent space, rather than a single point. Without this change, the latent space may not be continuous, as inputs can simply be mapped to disjoint points in latent space. By mapping inputs to distributions, similar inputs are allowed to have overlapping distributions, allowing interpolation between them. Kingma & Welling (2013) map each input onto a multivariate Gaussian distribution with no correlations. The latent layer is replaced with two sets of neurons: one representing the means in each of the dimensions of latent space, the other representing the log variances (log variances are used to guarantee that the variances are positive). We will refer to these values as the latent means and latent log variances. The decoder samples a point in latent space from this Gaussian distribution and then decodes it as the reconstruction. Figure 2 shows a VAE with a similar architecture to the autoencoder in Figure 1. The latent means and latent variances

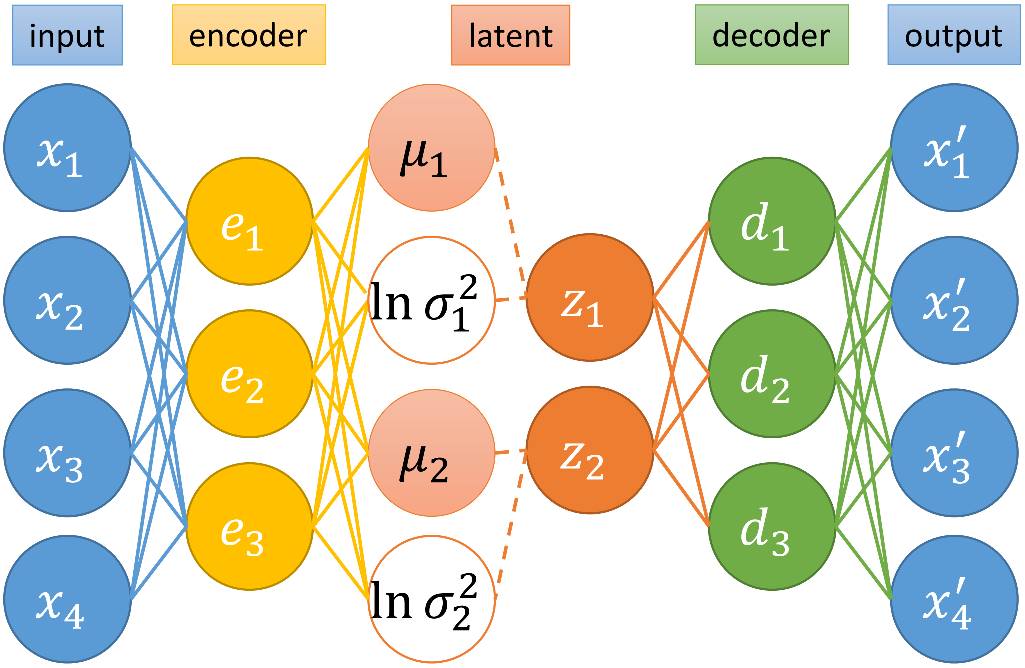

Figure 2. Structure of a variational autoencoder with a 4-3-2-3-4 architecture (similar structure to the autoencoder in Figure 1). The dashed lines indicate that each latent value $z_{1,2}$ is drawn from a Gaussian distribution defined by a latent mean $\mu_{1,2}$ and log variance $\ln \sigma_{1,2}^2$.

each have their own weights and biases

$$\boldsymbol{\mu} = W^{(\mu)}\boldsymbol{e} + \boldsymbol{b}^{(\mu)} \tag{6}$$

$$\ln \boldsymbol{\sigma}^2 = W^{(\ln \sigma^2)}\boldsymbol{e} + \boldsymbol{b}^{(\ln \sigma^2)} \tag{7}$$

and a point in latent space is sampled from $q(z|x)$, the Gaussian with these latent means and variances

$$\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x}) \equiv \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\exp(\ln \boldsymbol{\sigma}^2))). \tag{8}$$

This Gaussian can be thought of as a distribution of latent representations that are consistent with the input. Instead of propagating this distribution forward, the decoder samples a point from it and propagates it:

$$\boldsymbol{d} = f(W^{(d)}\boldsymbol{z} + \boldsymbol{b}^{(d)}). \tag{9}$$

Note that this sampling step is not deterministic but, as the VAE is trained, the same input will be seen many times, and many samples in latent space will be propagated through the decoder. A prior is placed on the latent space such that the latent distributions, averaged over the entire data set, comprise a standard multivariate Gaussian. The latent distributions for each input are already Gaussians, but this prior acts to regularize the means and variances of each latent distribution so that they roughly add up to a standard multivariate Gaussian for the entire data set. Different priors would give different regularizations, but a standard multivariate Gaussian is commonly used to make training the VAE faster. The objective function used is the evidence lower bound (ELBO), which is the sum of the reconstruction loss and the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951) between the latent distribution for the input $q(z|\boldsymbol{x})$ and the prior $p(z)$:

$$\mathrm{ELBO} = L(\boldsymbol{x}, \boldsymbol{x}') + D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})). \tag{10}$$

The KL divergence between two probability distributions p and q is a measure of the difference between the two distributions and is defined as

$$D_{\mathrm{KL}}(q||p) = \int q(z)\log\left(\frac{q(z)}{p(z)}\right)dz. \tag{11}$$

The reconstruction loss can be any function measuring the difference between the input and the reconstruction, like a χ^2 loss. If the reconstruction loss is a negative log likelihood (like the χ^2 loss for inputs with Gaussian errors) and the decoder is thought of as a generative model for the data, then the encoder is performing variational inference. That is, the Gaussian described by the encoder is the closest (by KL divergence) such Gaussian to the posterior that is implied by the combination of the prior, decoder (generative model), and reconstruction loss (negative log likelihood).

We use InfoVAE (Zhao et al. 2019), a VAE variant that addresses two issues with the ELBO objective function. First, the KL divergence term is not strong enough to discourage the VAE from mapping different inputs to disjoint distributions. This problem is worst when the dimensionality of the input is much greater than that of the latent space. Second, the KL divergence term is minimized when the latent distribution for each input matches the prior. In this case, the latent distribution does not depend on the input at all, as it always matches the prior. If this term is strengthened by multiplying it by a large number, as in β-VAE (Higgins et al. 2017), then this behavior will cause the VAE to under-utilize the latent space as it is encouraged to match the latent distribution to the prior. InfoVAE uses an objective function that alleviates these problems:

$$L_{\mathrm{InfoVAE}} = L(\boldsymbol{x}, \boldsymbol{x}') + (1-\alpha)D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + (\alpha + \lambda - 1)D_{\mathrm{MMD}}(q(\boldsymbol{z})||p(\boldsymbol{z})). \tag{12}$$

With $\alpha = 0$, the first two terms are the same as the ELBO objective. The last term compares $q(z)$, the latent distribution averaged over inputs, with the prior. The latent distribution averaged over inputs is a sum of Gaussians, but its KL divergence with the prior would be expensive to calculate. Instead, samples from the latent distributions of a minibatch of inputs are drawn, alongside samples from the prior. The maximum mean discrepancy (MMD; Gretton et al. 2007) is then calculated between the samples from the latent distributions and the prior, given by

$$D_{\mathrm{MMD}} = \frac{1}{m^2}\sum_{i,j=1}^{m} \kappa(u_i, u_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} \kappa(u_i, v_j) + \frac{1}{n^2}\sum_{j=1}^{n} \kappa(v_i, v_j) \quad (13)$$

where $u_{1\ldots m}$ are samples from the latent distributions, $v_{1\ldots n}$ are samples from the prior, and κ is a positive-definite kernel function such as a squared exponential. The MMD term encourages the latent distribution averaged over all inputs (rather than just the distribution for a single input) to match the prior, and can be strengthened by choosing $\lambda > 1$.

3. Application

As an initial demonstration of VAEs, we train a VAE using a subset of ≈64,000 spectra taken from the SDSS DR7 (York et al. 2000; Strauss et al. 2002; Gunn et al. 2006; Abazajian et al. 2009; Smee et al. 2013). Following the approach of Vanderplas & Connolly (2009), we select spectra from the main galaxy (≈59,000 spectra) and quasar (≈4500 spectra) samples with redshifts $z < 0.36$ using the AstroML (VanderPlas et al. 2012) SDSS query generator.[5] We then use AstroML to pre-process the selected spectra. First, we shift all spectra to their rest frames, resample them to 1000 logarithmically spaced wavelength pixels between 3388 and 8318 Å, and normalize by total flux in this wavelength range. This resampling is much coarser than the SDSS spectral resolution, and so small-scale information is lost. Since this work is meant to be a first demonstration of VAEs on SDSS spectra, we accept this loss of information in order to reduce the computational cost of training our VAEs. Then, we infill bad pixels (due to, e.g., bad sky line subtraction) using an iterative PCA procedure, as done in Yip et al. (2004a). We then only keep spectra that are classified by SDSS's spectroscopic pipeline `spectro1d` (Stoughton et al. 2002, Section 4.10) as galaxies (`SPEC_CLN=2`) or quasars (`SPEC_CLN=3`), excluding 2% of the spectra. A validation set of ≈16,000 spectra of the remaining spectra is set aside, leaving ≈47,000 spectra in the training set.

For the reconstruction loss, we use a Gaussian log likelihood using the reported uncertainties for each spectrum. Bad pixels have their weight in the loss set to zero (which is equivalent to giving them infinite uncertainty). To prevent a small number of very high signal-to-noise ratio (S/N) pixels from dominating the total loss, an uncertainty floor is added to each good pixel that caps the maximum pixel S/N to 50. This cap affects 1% of pixels across all spectra and only 0.5% of spectra have a median pixel S/N greater than 50.

Table 1
Best Architectures and MMD Coefficients λ Found by Random Search for VAEs with Two, Four, Six, and 10 Latent Parameters

Latent Parameters	Architecture	λ
2	1000-1663-42-2-42-1663-1000	11.2
4	1000-1134-64-4-64-1134-1000	21.2
6	1000-703-94-6-94-703-1000	3.02
10	1000-549-110-10-110-549-1000	7.72

We implement InfoVAE using `pytorch` (Paszke et al. 2017) and train our VAEs using `pytorch`'s built-in Adam optimizer (Kingma & Ba 2015), with a batch size of 64 and a starting learning rate of 10^{-3}. When the objective function does not improve for five epochs on the validation set, we reduce the learning rate by a factor of 10. We stop training if the objective function does not improve for 10 epochs on the validation set (i.e., early stopping). For the InfoVAE objective function hyperparameters, we set $\alpha = 0$ (as recommended by Zhao et al. 2019 for a simple decoding distribution) and use random search for λ. We use two hidden layers each for the encoder and decoder and use random search for the sizes of the hidden layers. All activations are ReLU, except for the nodes in the code layer and reconstruction layer which have linear activations. We make VAEs with two, four, six, and 10 latent parameters, stopping at 10 as this is the number of PCA components required to reconstruct non-quasar galaxies in Yip et al. (2004a). The architectures and values for λ found by random search are listed in Table 1.

We extract the emission-line equivalent widths and `spectro1d` spectral classifications from the FITS headers. The objects found by `spectro1d` to be quasars (`SPEC_CLN = 3`) are denoted "broad-line AGNs" (1.4%); 83.6% of these broad-line AGNs were targeted as quasars. We subdivide the remaining galaxies (`SPEC_CLN = 2`) based on their emission lines. The galaxies with Balmer emission strengths less than 3σ are denoted "quiescent galaxies" (60.1%), while galaxies with strong Balmer emission are denoted "emission-line galaxies" (31.9%) or "narrow-line AGNs" (4.0%) using the AstroML implementation of the Kewley et al. (2001) N II/Hα line-ratio diagnostic. The code we used to download and process the spectra, train the VAEs, and make the plots in this paper is available at https://github.com/stephenportillo/SDSS-VAE under an MIT license; version 1 has been archived in Zenodo:10.5281/zenodo.3840643. Training the VAE, including the hyperparameter search, takes about a day on two cores of an AMD EPYC 7401 processor.

3.1. Reconstruction Accuracy

To demonstrate VAEs' ability to compress and reconstruct spectra, we train VAEs with two, four, six, and 10 latent parameters. We then measure the reconstruction error of these VAEs applied to the validation set of spectra. As a comparison, we create PCA and non-negative matrix factorization (NMF) reconstructions with the same numbers of parameters. Both PCA and NMF reconstruct data as a linear combination of templates. We also train (non-variational) autoencoders (AEs) with the same numbers of latent parameters. Figure 3 shows the mean-squared reconstruction error of all reconstructions as a function of the number of parameters/components. Generally, at a given number of parameters, the AE and VAE perform better than the two linear methods. NMF gives somewhat better

[5] The query we used was `SELECT TOP 64000 plate, mjd, fiberid FROM specObj WHERE ((PrimTarget & (TARGET_GALAXY + TARGET_QSO_CAP + TARGET_QSO_SKIRT)) > 0) AND (z > 0) AND (z<= 0.36).`

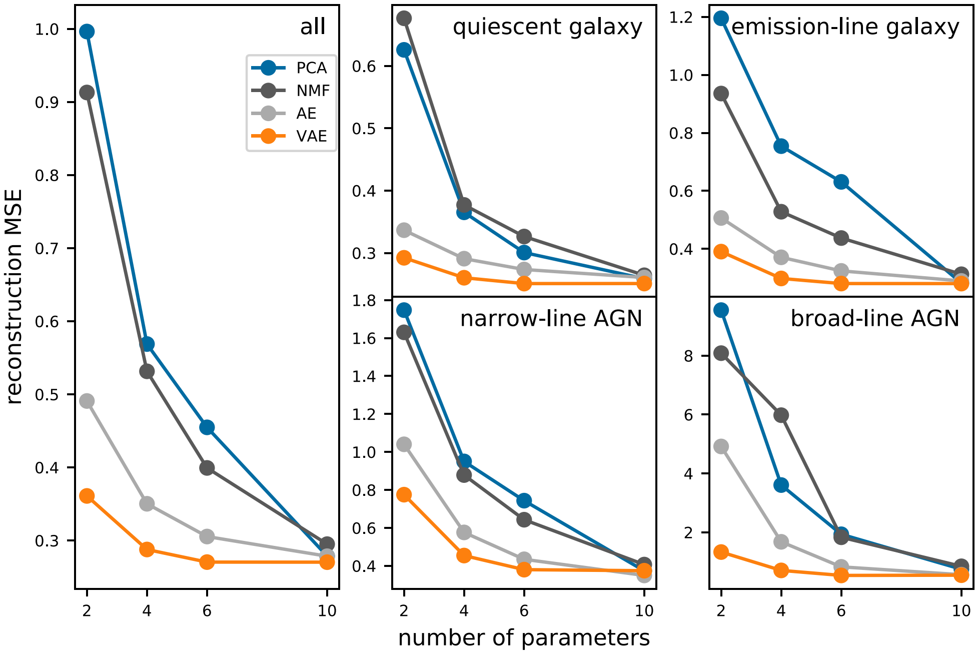

Figure 3. Mean-squared reconstruction error for principal component analysis, non-negative matrix factorization, non-variational autoencoder, and variational autoencoder (VAE). The VAE outperforms the other methods, with the greatest advantage being at small numbers of parameters. The VAE has the greatest relative advantage for broad-line active galaxies, which also have the highest absolute reconstruction error.

reconstructions than PCA, while the AE gives even better reconstructions than NMF, and the VAE gives the best reconstructions. The performance gaps are most significant when fewer parameters are used. The VAE performs especially well on broad-line AGNs, which have a wide range of emission-line widths. The PCA and NMF reconstructions, being linear combinations of templates, capture this variation using templates with deficits at the peaks of spectral lines and excesses in the wings. Therefore, these methods need more components to reconstruct broader spectral lines. By contrast, because the VAE is nonlinear, it can control the width of spectral lines with a single parameter. This nonlinearity allows a VAE with only two parameters to reconstruct spectra at least as well as a PCA reconstruction with six components. The performance gaps between the methods narrow as more parameters are used, but there are still significant differences in the details of the reconstructions.

Increasing the number of VAE parameters from six to 10 does not improve the VAE reconstructions much, suggesting that the first six parameters capture more information than the last four. In the 10-parameter VAE, we also find that the latent variances of four of the parameters are typically wider than the prior, meaning that the VAE is not constraining these parameters for each spectrum. This finding indicates that specifying these four parameters is not important for reconstructing spectra, again suggesting that the other six parameters are the most important. We find that the 10-parameter VAE sometimes gives slightly worse reconstructions than the six-parameter VAE: the four unconstrained parameters seem to be adding noise in these cases.

We now compare the VAE and PCA reconstructions more qualitatively by zooming in on interesting regions of selected spectra in the validation set. We show the reconstruction that corresponds to the latent mean for each spectrum; that is, we do not sample from the latent distribution defined by the latent mean and latent variance, as is done by the VAE during training. We find that variance in the reconstruction that arises from sampling from the latent distribution is much smaller than the spectrum measurement uncertainties. In Figure 4, we show the Na 5896, N II 6550/6585, and Hα lines of a high-S/N quiescent galaxy (median pixel S/N 82). With two parameters, both reconstructions show the Na absorption and N II emission, but the VAE better reproduces the continuum. The PCA reconstruction at two parameters shows hints of Hα emission that are absent from the VAE reconstruction. Both reconstructions better approximate the amplitude of the lines as more parameters are added, except the VAE reconstruction of the Na absorption lines appears to worsen between six and 10 parameters. Still, with four or six parameters, the VAE reconstruction better fits the continuum and the lines.

In Figure 5, we show the Na 5896, N II 6550/6585, and Hα lines of a low-S/N quiescent galaxy (median pixel S/N 11). Even with the lower S/N, both reconstructions capture the Na absorption, N II emission, and Hα emission with only two components. The differences between the two reconstructions are less pronounced at this lower S/N. At four and six components, the PCA reconstruction overestimates the flux in the Hα line and N II 6565 line, respectively, but these residuals do not exceed 2σ.

In Figure 6, we show the O II 3727/3730, Hα, and N II 6550/6585 lines of an emission-line galaxy. The lines have equivalent widths of 22, 25, 62, 2, and 7 Å, respectively. While all the lines are present in both reconstructions with only two parameters, both reconstructions struggle to match the amplitudes of the lines. With four parameters or more, both reconstructions underestimate the O II and N II emission while overestimating the Hα emission. The VAE reconstruction does not improve after four parameters, while the PCA reconstruction does improve with 10 parameters. This mismatch may be due to our choice to cap the pixel S/N to 50 (see the beginning of Section 3), which puts less emphasis on reconstructing high-S/N pixels like extreme emission lines.

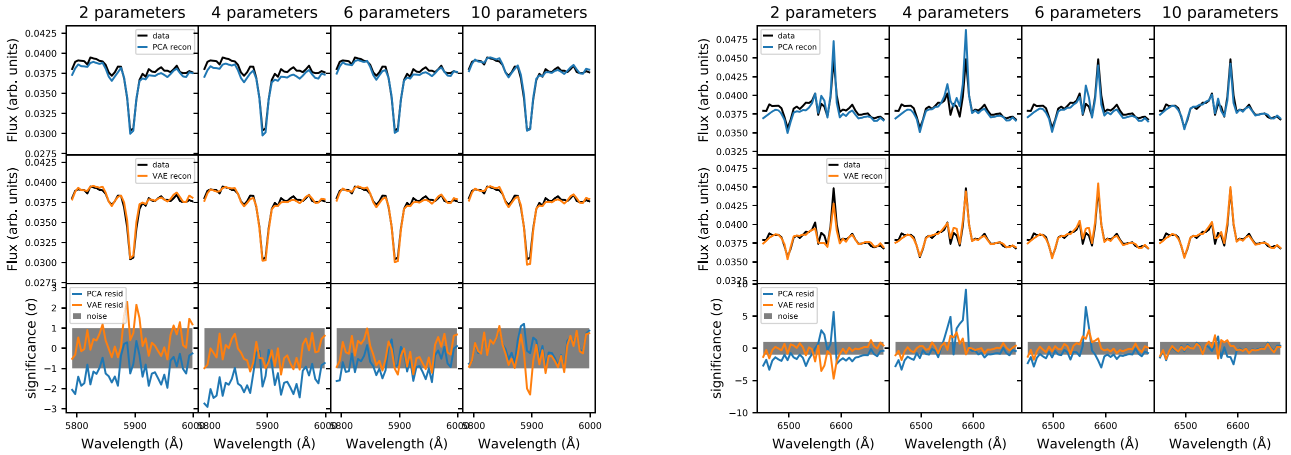

Figure 4. High-S/N quiescent galaxy (Plate-MJD-Fiber 394-51913-561): principal component analysis (PCA) and VAE reconstructions of the Na 5896 absorption line (left) and N II 6550/6585 and Hα emission lines (right). The noise band represents the SDSS stated pixel uncertainties. The median S/N of the spectrum is 78 in the left panel and 102 in the right panel. Both reconstructions find the significant Na absorption and N II emission with only two components, but the VAE provides a better reconstruction overall than PCA with the same number of components.

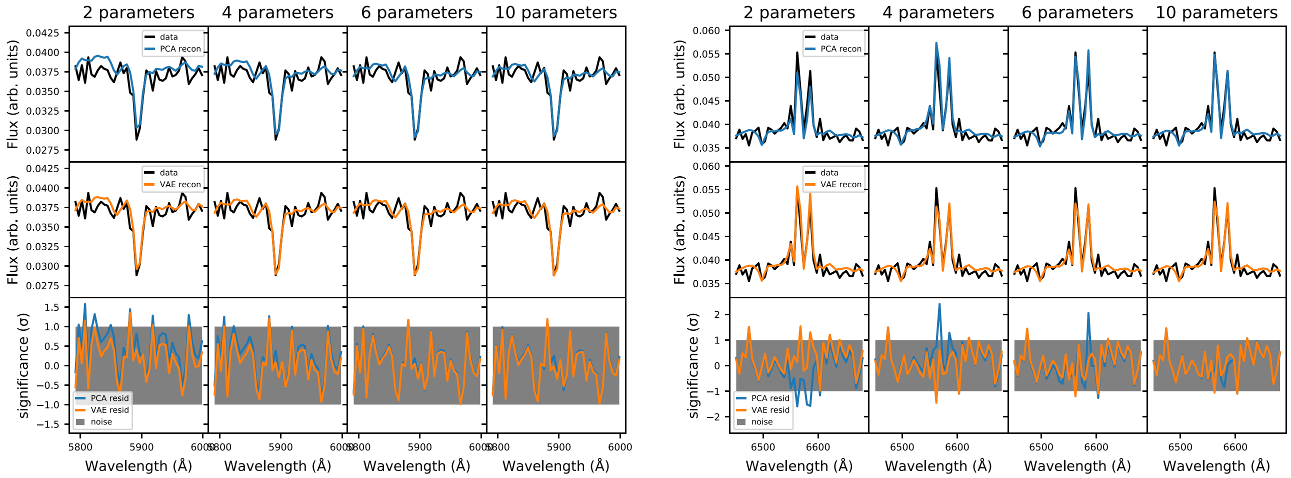

Figure 5. Low-S/N quiescent galaxy (Plate-MJD-Fiber 497-51989-573): PCA and VAE reconstructions of the Na 5896 absorption line (left) and N II 6550/6585 and Hα emission lines (right). The median S/N of the spectrum is 12 in the left panel and 18 in the right panel. Even at low S/N, both reconstructions find the Na absorption, Hα, and N II emission with only two components. The VAE still provides a better reconstruction compared to PCA with the same number of parameters, although the differences are less significant at low S/N.

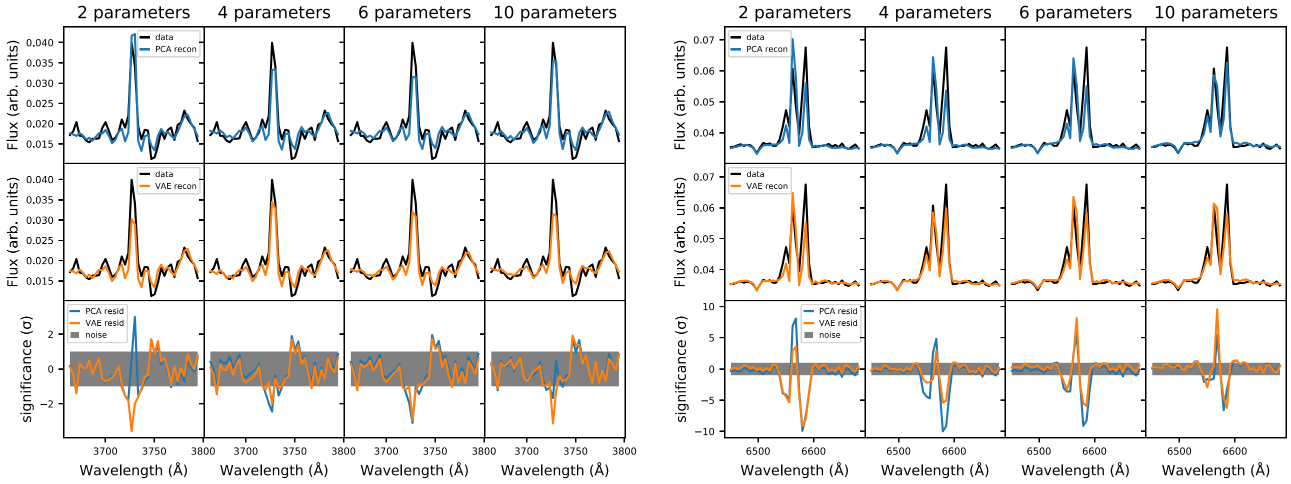

Figure 6. Extreme emission-line galaxy (Plate-MJD-Fiber 391-51782-353): PCA and VAE reconstructions of the O II 3726/3729 (left) and Hα and S II emission lines (right). The median S/N of the spectrum is 9 in the left panel and 13 in the right panel. While the PCA reconstruction of the strong O II 3727 line gets better with more components, the VAE reconstruction does not improve after four parameters.

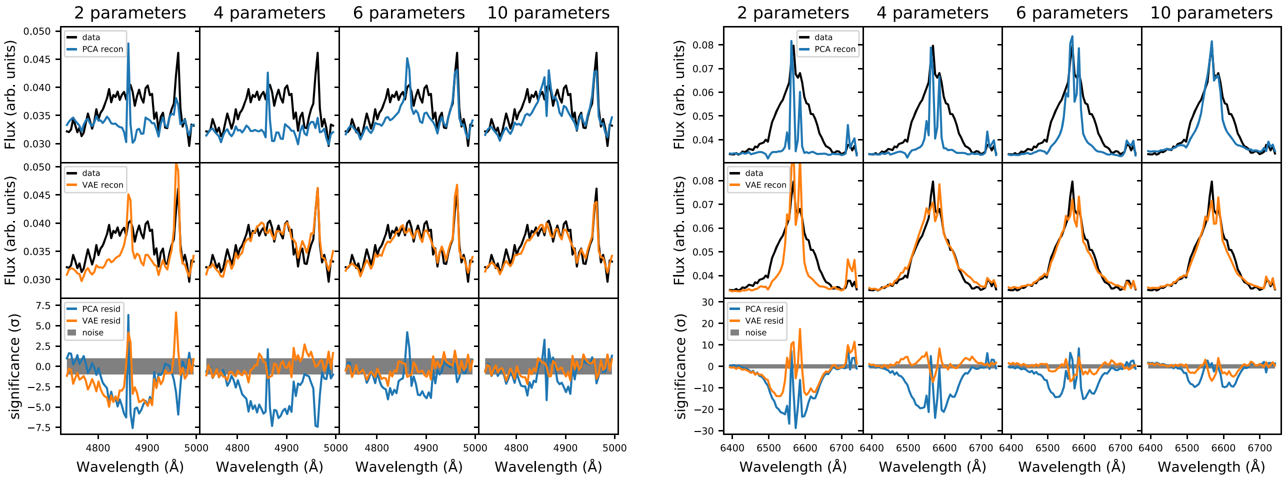

Figure 7. Broad-line active galaxy (Plate-MJD-Fiber 541-51959-519): PCA and VAE reconstructions of the Hβ and O III 4959 emission lines (left) and Hα, N II 6550/6585, and S II 6718/6733 emission lines (right). The median S/N of the spectrum is 30 in the left panel and 47 in the right panel. The VAE reconstructs the amplitude and width of the lines with four parameters, while the lines in the PCA reconstruction are too narrow, even with 10 parameters.

In Figure 7, we show the Hβ, O III 4959, Hα, N II 6550/6585, and S II 6718/6733 lines of a broad-line active galaxy. The SDSS pipeline reports that the Balmer lines have a width of ≈3000 km s^{-1} while the other lines have a width of ≈1800 km s^{-1}. The four-parameter VAE captures the width of all of these spectral lines well. The four-component PCA reconstruction has these emission lines, but they are far too narrow. As more components are added, the PCA reconstruction improves, but even with 10 components the emission lines are not wide enough.

3.2. Interpretation of VAE Tracks

In Figure 8, we show a corner plot of the latent mean of each galaxy in VAE latent space. The axes of the VAE latent space are arbitrary and may not be the most human-interpretable directions in latent space. We construct a basis in latent space by using PCA on the latent means to find vectors that describe the variance of the spectra *in latent space*. This procedure differs from using PCA directly on the spectra, which finds eigenspectra that describe the variance of the spectra *in spectrum space*. The last four components explain very little of the variance in latent space, which is expected since the six-dimensional VAE can reconstruct spectra nearly as well as the 10-dimensional VAE, as discussed in Section 3.1. Thus, we drop the last four components and only plot projections onto the first six components. For brevity, we will refer to these PCA components of the VAE latent space simply as VAE 1 through VAE 6, in descending order of explained variance.

The VAE latent space separates quiescent galaxies, emission-line galaxies, narrow-line AGNs, and broad-line AGN. This separation can most clearly be seen in the scatter plot of VAE 1 and VAE 5 in Figure 8. This projection shows two distinct arms: one that goes from quiescent galaxies to emission-line galaxies, and another that splits off for the narrow-line and broad-line AGNs. The hook at the end of the emission-line track corresponds to the extended linear feature that can be seen in many of the other projections.

To interpret the VAE latent space, we create tracks that follow the distribution of galaxies and then use the VAE to generate synthetic spectra along these tracks. To create a track, we first choose some spectral classes to follow and a VAE component (or linear combination of components) to use to bin the spectra. In each bin, we find the centroid of all included spectra that fall into this bin and then make this centroid a point in the track. For example, to create the star formation track, we first select the quiescent and emission-line spectra and then bin these spectra by their VAE 1 values. Then for each bin in VAE 1, we find the centroid in latent space and add that point to the track. We show these tracks in two projections (VAE 1 versus VAE 2 and VAE 1 versus VAE 5) in Figure 9. These sequences of spectra show how galaxy spectra change as we move in the VAE latent space.

3.2.1. Star Formation Track

The first track we create goes from quiescent galaxies to emission-line galaxies (labeled S1-5 in Figure 9), and its spectra are plotted in Figure 10. This track mostly follows VAE 1, and we stop it before it goes into the emission-line galaxy hook that is seen in the projection of the VAE 1 and VAE 5. This track mostly changes in VAE 1 without moving in the other components. At S1, the spectrum is of a quiescent galaxy: the continuum is red, indicating an old stellar population, with absorption lines like Ca H, Ca K, G band, Mg, and Na. As we proceed toward S5, the continuum gets bluer, indicating younger stellar populations, and nebular emission lines start to appear. First at S2, hints of O II 3726/3729 and N II 6583 appear as VAE 1 increases. Then these emission lines strengthen and are joined by lines like S II, O III 5007, and the Balmer series. This track acts similarly to the second PCA component in Yip et al. (2004a): red galaxies have large, positive second PCA coefficients while blue galaxies have small, or even negative, second PCA coefficients.

3.2.2. Extreme Line Emitters

Our second track (labeled E1-4 in Figure 9) picks up at the end of the previous track and continues into the emission-line galaxy hook seen in VAE 1 and VAE 5. This track's spectra are plotted in Figure 11. These galaxies all have extremely strong line emission, and as the track continues up the hook, the emission lines get stronger and the continuum blueward of 4000 Å increases. From E1 to E4, log(O III/Hβ) stays almost constant at ≈0.3, while log(S II/Hα) decreases from −0.5 to −0.7 and log(N II/Hα) decreases from −0.7 to −1.6.

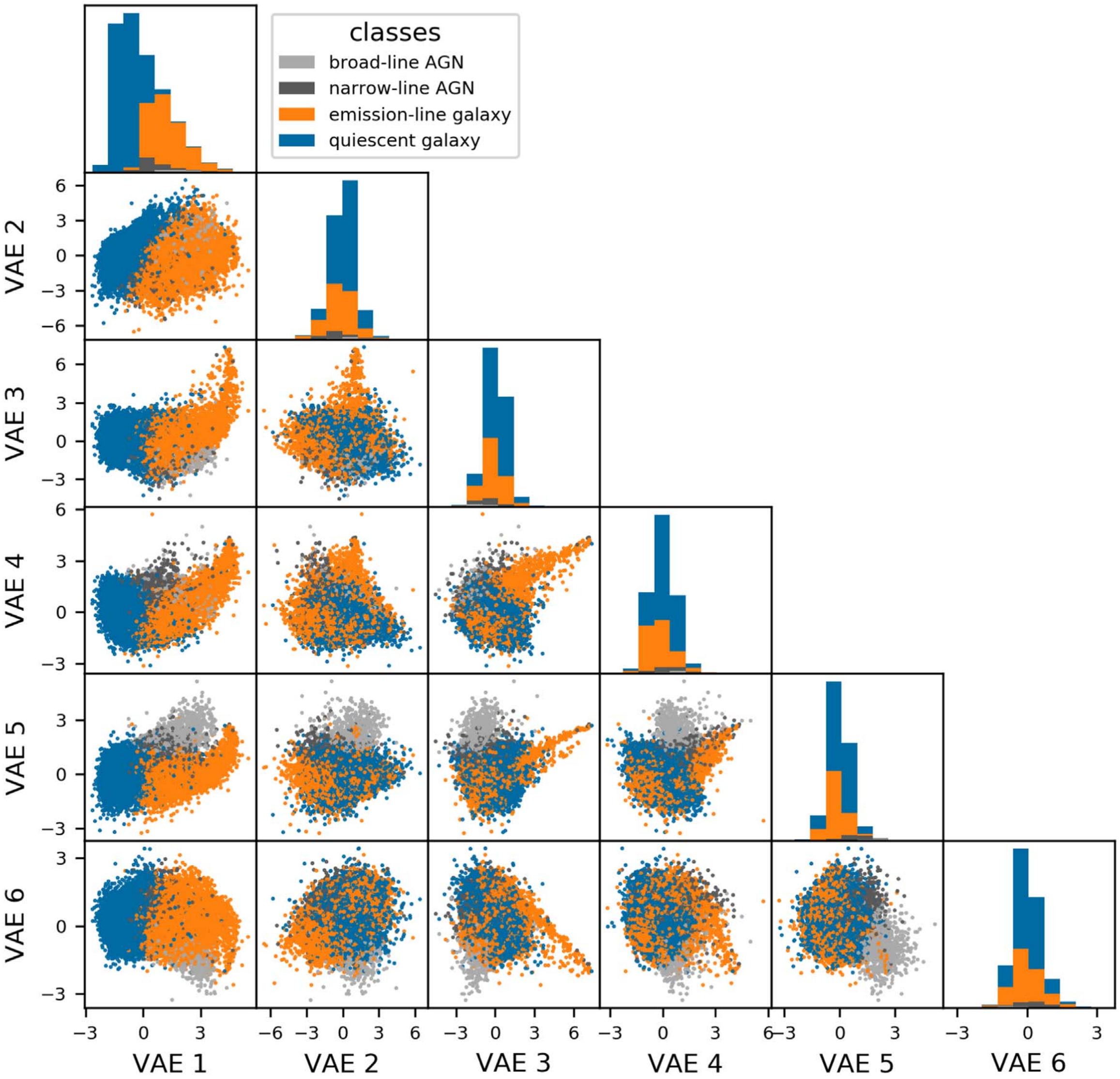

Figure 8. Corner plot of the first six VAE components of all training and validation spectra. The spectra are color-coded based on the classification outlined in Section 3.

Yip et al. (2004a) find a similar track in their parameters (ϕ_{KL}, θ_{KL}), which are transformations of their first three PCA coefficients. They find a track in this space that proceeds from red to blue to extreme emission-line galaxies, showing a similar progression as our S and E tracks put together.

3.2.3. Post-starburst Track

The third track we create follows the quiescent galaxies as VAE 2 changes (labeled P1-5 in Figure 9), with the spectra plotted in Figure 12. The middle of this track (P3) intersects the star formation track near S2. At P1, the spectrum has a red continuum, indicating older stellar populations, but also has nebular emission lines. Proceeding along the track, the continuum gets bluer and the nebular emission lines weaken. P4 and P5 are post-starburst spectra, with a blue continuum from young stars and weak nebular emission lines. Similarly, Yip et al. (2004a) find that galaxies with negative second and third PCA coefficients are post-starburst galaxies.

3.2.4. Active Galaxy Track

This track (labeled A1-5 in Figure 9) starts near S3 on the star formation track and diverges upward in VAE 5 as VAE 1 increases. The spectra along this track are plotted in Figure 13. At A1, the spectrum has narrow emission lines, while at A5, the spectrum is a broad-line active galaxy. Proceeding along the track, the nebular emission lines get stronger and the Balmer lines get stronger and broaden. The continuum also gets bluer, especially blueward of 4000 Å.

3.2.5. Individual VAE Components

We also create tracks that change only one VAE component at a time, with the middle of each track intersecting with the centroid of all galaxies' latent means. The spectra from these tracks are shown in Figure 14. These spectra show the effect that a VAE component has in isolation, but do not show effects that occur when multiple components are changed from their mean values.

Increasing VAE 1 changes spectra in a similar way as going along the star formation track: strengthening the nebular

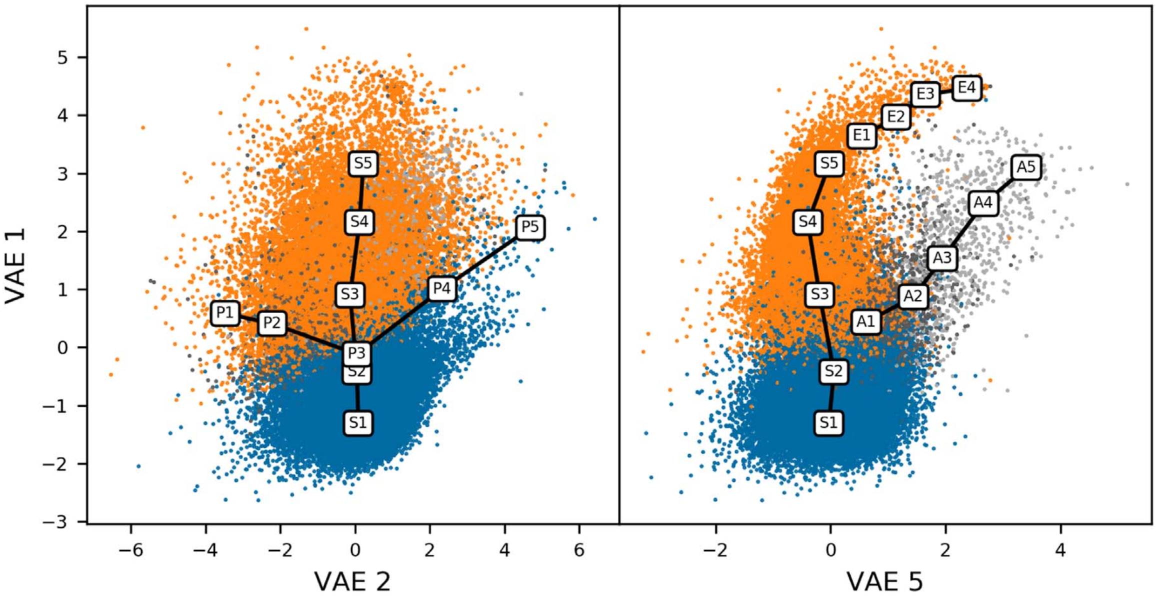

Figure 9. Scatter plot of the first, second, and fifth VAE components, with the four tracks discussed in Section 3.2 overlaid. The tracks are the (S) star formation track, the (E) extreme line emitters, the (P) post-starburst track, and the (A) active galaxy track.

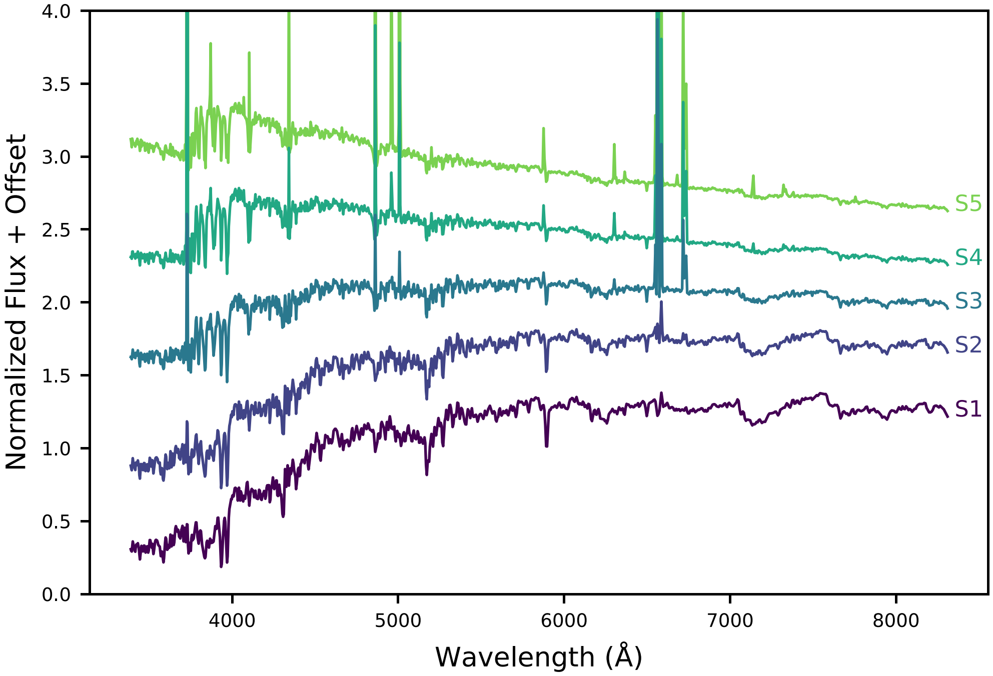

Figure 10. Synthetic spectra along the star formation track in Figure 8, plotted with offsets added to separate the spectra. Proceeding through the sequence, the continuum changes from that of an old to a young stellar population, and nebular emission lines appear and strengthen.

emission lines and making the continuum bluer. This behavior is expected, as the main component that changes along the star formation track is VAE 1. By contrast, decreasing VAE 2 strengthens the nebular emission lines while making the continuum slightly redder. O II 3727 and O III are also more prominent as VAE 2 decreases. These two parameters give the VAE the ability to reconstruct different combinations of stellar population age and star formation activity. Increasing VAE 1 corresponds to a younger stellar population and higher star formation activity, while increasing VAE 2 corresponds to a younger stellar population and lower star formation activity. Given this combination of effects, it is not suprising that the post-starburst galaxies (P4 and P5) have high VAE 1 and VAE 2.

Increasing VAE 3 weakens Hα and N II while retaining O II 3727. The continuum does not change much, in contrast to changing VAE 1 and VAE 2. Increasing VAE 4 weakens Hα and N II while also strengthening O II 3727 and O III 5007 and making the spectrum slightly bluer.

Increasing VAE 5 broadens the Hα line and also makes the continuum redder. Increasing VAE 6 weakens Hα and N II while strengthening O II 3727 and O III 5007, similarly to VAE

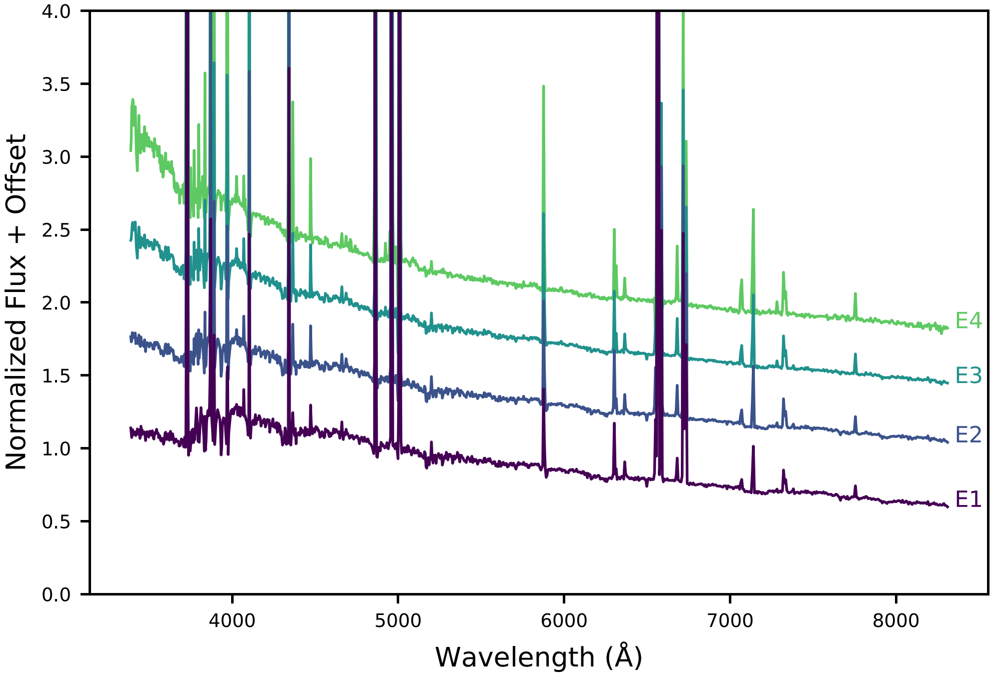

Figure 11. Synthetic spectra along the starburst sequence in Figure 8, plotted with offsets added to separate the spectra. Proceeding through the sequence, the nebular emission lines strengthen and the blue end of the continuum rises.

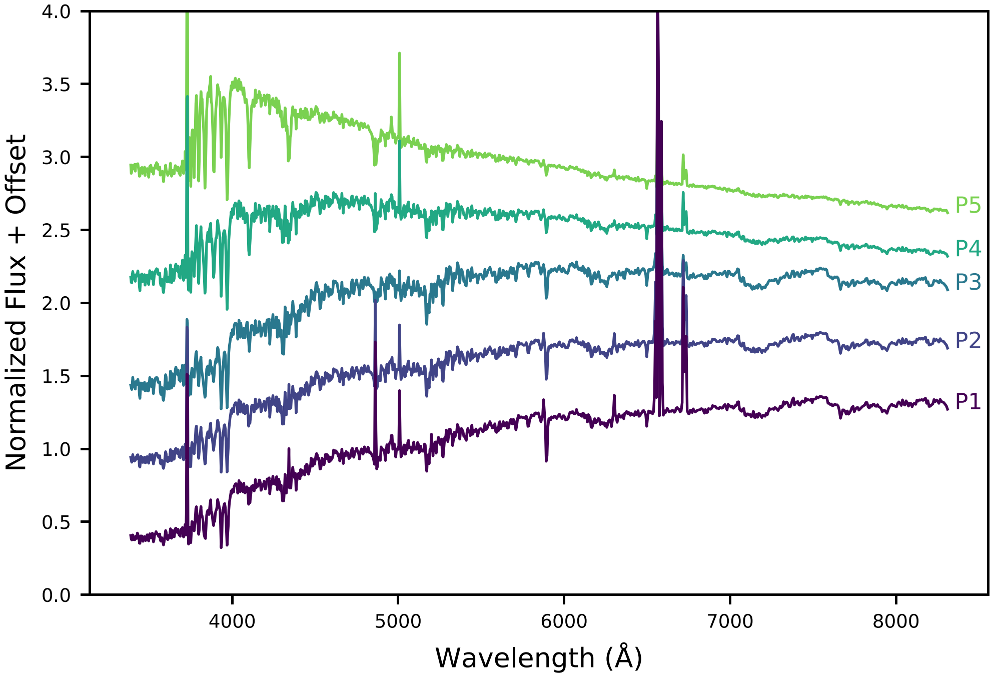

Figure 12. Synthetic spectra along the post-starburst sequence in Figure 8, plotted with offsets added to separate the spectra. The sequence starts with an old stellar population with nebular lines and ends with post-starburst spectrum with a young stellar population but weak nebular emission lines.

4. Unlike increasing VAE 4, increasing VAE 6 does not make the continuum bluer.

The six VAE components give the VAE the flexibility to fit different combinations of continuum color and emission-line strengths. While these sequences for each VAE component can be interpreted, we must emphasize that these sequences were all chosen to intersect at the centroid of latent space. Because the VAE is nonlinear, the effect of each component can vary as a function of position within latent space.

3.3. Outlier Identification

We consider using the VAE latent space to define outliers that are in atypical locations of latent space. These outliers could be members of rare classes or spectra that cannot be well reconstructed by the VAE. We use the local outlier factor (LOF) algorithm (Breunig et al. 2000) to identify outliers. The algorithm estimates the local density of each point by using *k* nearest neighbors and then identifies points with densities much lower than their neighbors' as outliers. The LOF algorithm is

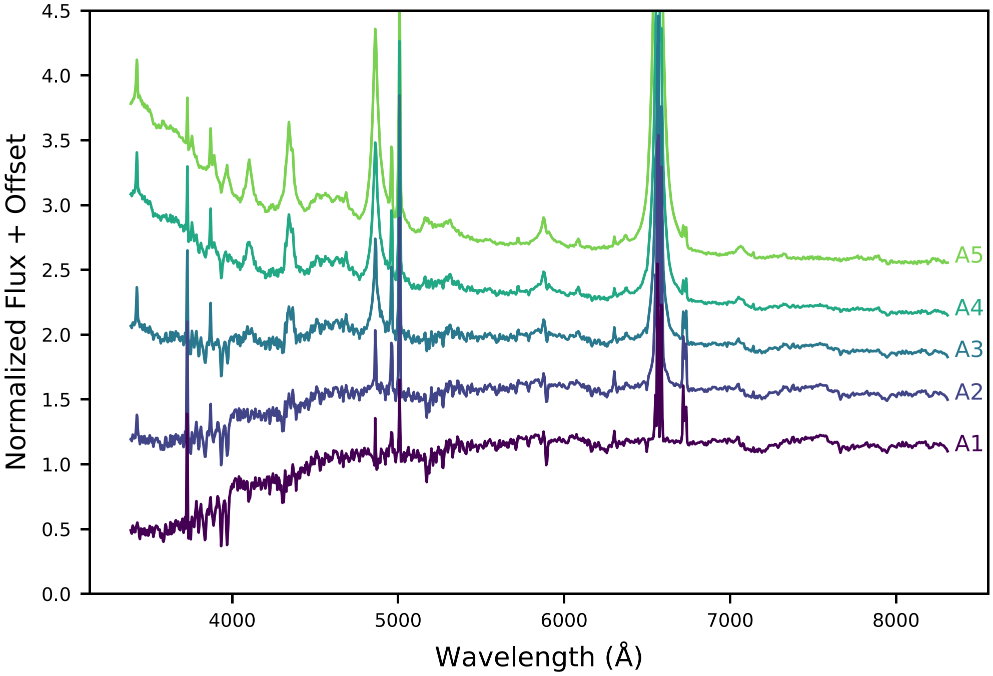

Figure 13. Synthetic spectra along the active galaxy sequence in Figure 8, plotted with offsets added to separate the spectra. The sequence progresses from a Type 2 Seyfert spectrum with only narrow emission lines to a Type 1 Seyfert spectrum with broad Balmer emission lines.

unsupervised (i.e., it is not given training labels for which spectra are outliers), so we can search for outliers in the training set as well as the validation set. We use LOF with $k = 20$ nearest neighbors and present the top 10 outliers (labeled O1–O10 with O1 being the top outlier) in Table 2 and Figure 15.

Two of the outliers, spectra O1 and O8, have very low S/N, suggesting that the VAE is unable to map these spectra to their true latent parameters because of the severe noise. O1 is particularly severe, with most of the pixels being identified as bad pixels.

Spectrum O10 looks like a quiescent galaxy, but with a large range of data missing. The iterative PCA procedure we use to infill missing data (similar to the one used in Yip et al. 2004a) has placed strong absorption lines in the missing wavelength range. These absorption lines are unlike those seen in quiescent galaxies. Although these infilled pixels are ignored in the reconstruction loss, they are still fed into the VAE and can affect the latent parameters of the spectrum. Apparently the erroneous absorption lines had a large enough effect to make O10 an outlier in latent space.

Spectra O2 and O4 have unusual continuum shapes, with a discontinuity at around 6000 Å in both cases. Taking into account the stated redshift for each spectrum, the discontinuity occurs at the wavelength separating the red and blue channels of the SDSS spectrograph. Thus, the discontinuity is likely an artifact arising from an error in calibration between the red and blue channel.

Spectrum O3 does not look like a typical galaxy spectrum, and it was taken within 10″ of a bright ($r = 10$) star. It is likely that this spectrum is being contaminated by stellar light, making it an outlier.

Four of the outliers are stars which were erroneously classified as galaxies, suggesting that the VAE can identify these spectra as not belonging to the classes that make up the vast majority of the training set. Spectrum O5 is an A star. Spectra O6, O7, and O9 are all M stars, with O7 and O9 being spatially coincident with a galaxy. In Figure 16, we plot the positions of a sample of SDSS stellar spectra in the VAE latent space. Some of the stellar spectra are easily identified as outliers in VAE latent space, while others overlap with the distribution of galactic spectra.

While the top 10 outliers do not contain any new classes of objects, they are all objects that are atypical in some way, suggesting that the VAE latent space can be used to find unusual spectra. Many of the spectra are outliers because of data artifacts, but four of them are stellar spectra that somehow were classified as galaxies. Stellar spectra are rare in our set of spectra because we selected for galaxies; the fact that the VAE identifies them as outliers means that it may be able to find other rare classes, given a larger data set, or may be used to reject incorrect targeting or failed spectral extraction.

3.4. Sensitivity to Noise

To test the VAE's sensitivity to noise, we take the highest-S/N validation spectrum (median pixel $S/N \approx 82$) and generate many realizations of white noise. In our training set, the spectra have median pixel S/Ns ranging from 5 to 88, with a median of 14. We add different levels of noise to the original spectrum to obtain simulated noisy spectra at varying S/N. Then, we encode this set of noisy spectra using the VAE, obtaining latent means and variances for each spectrum in the set. We focus on the six parameters with latent variances narrower than the prior in latent space, as these are the most important parameters for reconstruction (see Section 3.1). For each S/N, we consider the distribution of latent means in each latent parameter obtained by encoding the noisy spectra of that S/N. These distributions look roughly Gaussian, even as strong noise is added.

The variances of these distributions of latent means is a Monte Carlo measurement of the uncertainty of a spectrum's position in latent space caused by the addition of white noise. Ideally, the latent variance returned by the VAE for a spectrum would also be an estimate of the uncertainty of that spectrum's

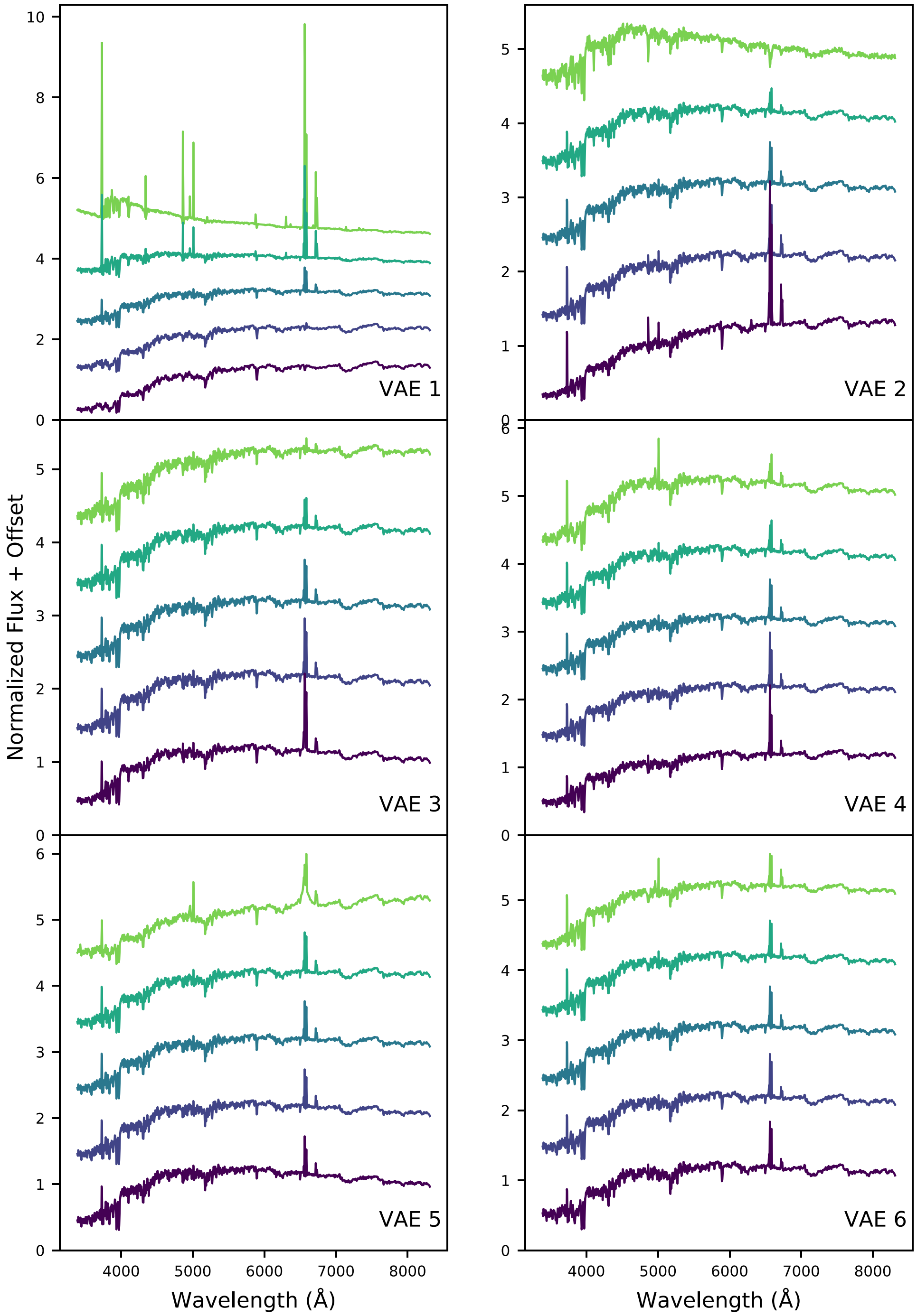

Figure 14. Synthetic spectra along sequences where one VAE parameter is changed at a time, plotted with offsets added to separate the spectra. The middle spectrum in each sequence corresponds to the centroid of the VAE latent means of all spectra. These sequences are discussed in Section 3.2.5.

position in latent space. We plot the ratio of these two estimates of the uncertainty, as a function of S/N, in the left panel of Figure 17. The variance of the distribution of latent means increases as stronger noise is added, showing that adding noise increases the uncertainty in estimating the latent parameters of the original spectrum. By contrast, we note that the latent variances do not change drastically as noise is added, showing that the VAE does not increase its reported uncertainty in response to added noise. Thus, the variance of latent means is larger than the latent variance at low S/N and smaller at high S/N. The VAE only sees the spectrum uncertainties in training, through the definition of the reconstruction loss. That is, the VAE is not given the spectrum uncertainties at test time and thus cannot directly propagate those uncertainties to inform the latent variance. The VAE apparently does not try to estimate the level of noise from the spectrum itself to set the latent variance. Thus, the latent variance overestimates the uncertainty at high S/N and underestimates the uncertainty at low S/N. The crossing point is at $S/N = 25$, which is higher than 90% of the spectra, so the VAE is underestimating the uncertainty in latent space for most spectra. Adding the spectrum uncertainties as input features that the VAE sees during training could allow the VAE to better estimate the uncertainty in latent space.

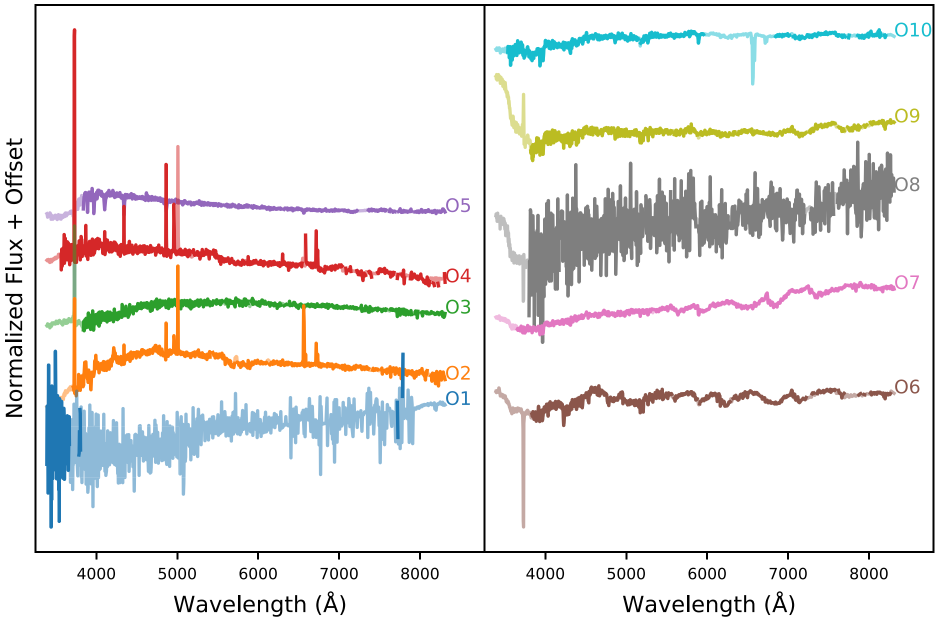

Figure 15. Top 10 spectra outliers in VAE latent space using the local outlier factor algorithm, plotted with offsets added to separate the spectra. The faded parts of the spectra are bad pixels or outside the wavelength range of the original spectrum and are infilled with iterative PCA (see Section 3).

Table 2
Plate-MJD-fiber IDs of the Top 10 Outlier Spectra Identified in the VAE Latent Space, with Likely Explanations for Why They Are Outlier Spectra

Spectrum	Plate	MJD	Fiber	Explanation
O1	445	51873	68	low S/N
O2	334	51993	203	bad calibration
O3	480	51989	77	close to bright star
O4	454	51908	607	bad calibration
O5	424	51893	587	A star
O6	305	51613	299	M star
O7	352	51694	340	M star
O8	529	52025	200	low S/N
O9	276	51909	2	M star
O10	414	51869	296	missing data

The distribution of latent means would ideally be centered on the latent mean of the original spectrum, showing that the addition of noise does not bias a spectrum's position in latent space. In the right panel of Figure 17 we show the bias of the latent means as a function of noise level. To make the biases in different parameters comparable, we divide the bias in each parameter by the square root of the latent variance of that parameter in the original spectrum. If the biases are comparable to the latent variance, then they are problematic as they are comparable to the VAE's stated uncertainty in latent space. The biases do grow at low S/N, but remain smaller than the latent variance for $S/N > 8$.

4. Discussion

VAEs are able to learn an effective compression of galaxy spectra. As shown in Section 3.1, a VAE with two parameters can reconstruct spectra as well as PCA with six components, and a VAE with six parameters can reconstruct spectra as well as PCA with 10 components. Because VAEs are nonlinear, they are better able to reconstruct broad spectral lines. VAEs also outperform NMF and AEs with the same number of parameters.

The VAE latent space is not only a good compression of spectra, but is also interpretable. The VAE can generate a synthetic spectrum for any point in latent space. In Section 3.2, we show that traversing latent space yields continuous changes in continuum shape, line amplitude, and line width. We find tracks in the latent space that correspond to changes in star formation rate, post-starburst activity, and a transition from narrow-line to broad-line spectra. We stress that synthetic spectra are only interpretable in parts of latent space that are supported by the training set: in parts of latent space far from any training examples, the VAE is forced to extrapolate. Some of these tracks in VAE latent space have analogs in the PCA coefficient space constructed by Yip et al. (2004a). Just as we find a track from quiescent to star-forming galaxies in VAE 1, they find a similar sequence by increasing their second PCA coefficient. We also identify parts of VAE latent space occupied by extreme line-emitting galaxies and post-starburst galaxies, just as they find parts of PCA coefficient parameter space occupied by the same classes of galaxies. Because our VAE is able to handle nonlinear features better than PCA, we are able to map broad-line AGNs onto the same latent space as the other classes of galaxies, unlike Yip et al. (2004a).

The latent space also separates quiescent galaxies, star-forming galaxies, narrow-line AGNs, and broad-line AGNs, without the VAE having ever seen categorically labeled spectra. Indeed, a classification system could be made based on subdivisions of latent space. Such a system would combine the continuum sensitivity of PCA with the line amplitude sensitivity of line-ratio tests, along with line widths. Furthermore, sequences of synthetic spectra could be made that show prototypical members of each class along with transitional objects that are near class boundaries. These synthetic spectra would allow the latent space classification to be linked to other classification schemes such as PCA or line-ratio tests.

Most of the computational cost of a VAE is in training it. After training, the main cost in encoding a spectrum is de-redshifting it; encoding the de-redshifted spectrum is

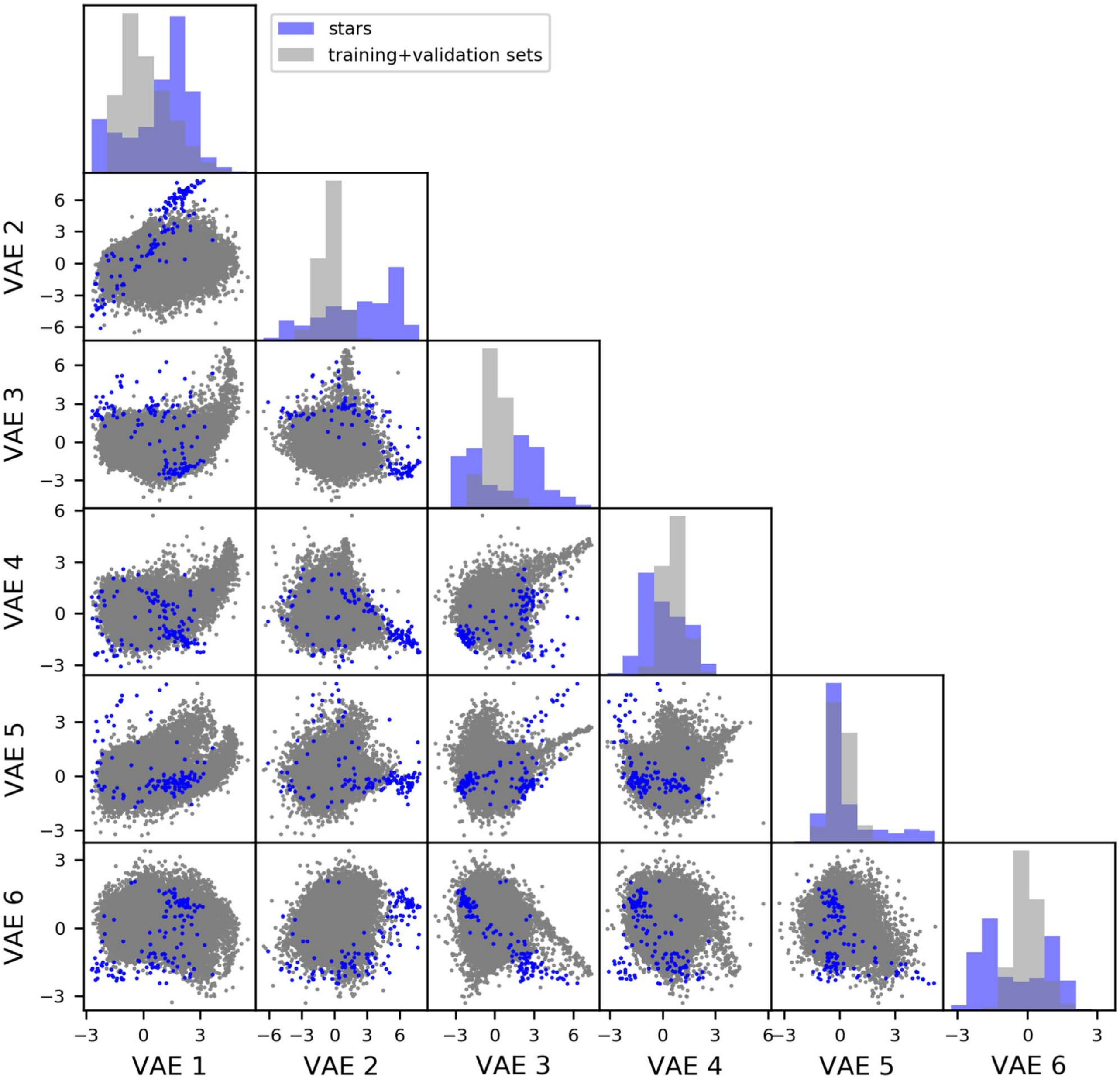

Figure 16. Corner plot of the first six VAE components of all training and validation spectra compared with a sample of stellar spectra.

inexpensive. A VAE might even be useful in determining the redshift of a spectrum: many trial redshifts could be tested to see which one produces a de-redshifted spectrum that is best reconstructed by the VAE and thus looks similar to the training de-redshifted spectra. Generating synthetic spectra for points in latent space is also inexpensive. We downsample the spectra to 1000 pixels and train on only ≈47,000 spectra in order to lessen the computational cost of training the VAE. Higher-resolution spectra would require a larger neural net to adequately reconstruct them, which would take more epochs to train. Expanding the training set could improve VAE reconstruction or reveal interesting interpretations of latent space, but would also make each epoch of training longer.

The VAE is successful in reducing the dimensionality of spectra from 1000 pixels to the six most important VAE components; however, even this smaller, six-dimensional space is difficult to fully explore. While highly informative, the set of scatter plots in Figure 8 are 2D projections that do not capture the full structure of latent space. Just as we use the `spectro1d` and line-ratio classifications to interpret the latent space, other classifications could be used to create further interpretations. Simulated spectra could also be given to the VAE to decode to see how the simulated galaxies' properties map onto latent space. Clustering techniques like *k*-means or DBSCAN (Ester et al. 1996) could be used to find meaningful divisions of latent space in an unsupervised manner. These clustering techniques should be faster to run and perhaps more effective in the smaller latent space, rather than the original space of spectra. Interactive visualizations could also be useful in allowing astronomers to explore latent space. With an interactive visualization, different latent parameters could be scatterplotted, and the user could quickly look at the spectra that correspond to different points in latent space or color the scatterplot by some measured property of the spectra. Reis et al. (2019) reduce the dimensionality of the SDSS spectra with

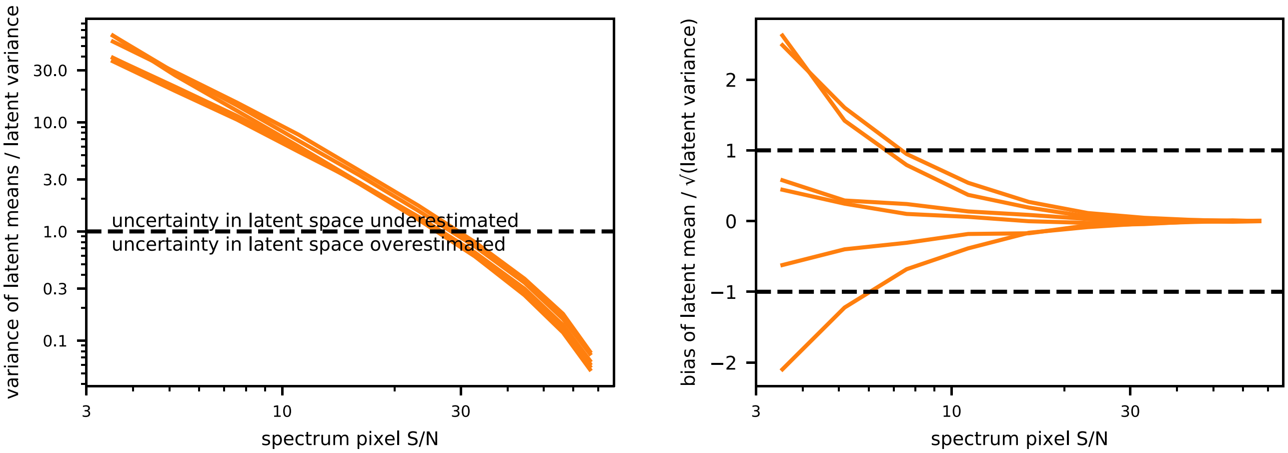

Figure 17. Variance (left) and bias (right) of the latent mean of each latent parameter (one line per parameter) when white noise is added to a high-S/N spectrum, in units of the latent variance. The latent variance underestimates the uncertainty in latent space for spectra with $S/N < 25$, which are 90% of the spectra. The latent mean is not significantly biased if the spectrum has $S/N \gtrsim 8$.

Uniform Manifold Approximation and Projection and present an interactive visualization.[6]

We made a first effort to find outlier spectra in Section 3.3, but a search for rare spectra would need to involve a much larger data set. The VAE could be trained on the entire SDSS main galaxy sample, for example, to then find the most unusual spectra. Techniques like DEMUD (Wagstaff et al. 2013) would be useful in finding interesting spectra. Again, the reduced dimensionality of the latent space should make outlier searches easier.

The VAE is able to encode spectra with missing data and levels of noise, but uncertainties in the data could be handled better. In our VAE, both missing data and measurement uncertainty are handled through weighting the data pixels in the reconstruction loss: good pixels are inverse-variance weighted and bad pixels are given a weight of zero. We find that setting a noise floor (i.e., a maximum weight) improves the stability of training the VAE. Excluding the bad pixels from the reconstruction loss means that the VAE is unconcerned with reconstructing the fluxes in these pixels. However, the fluxes in these bad pixels are still propagated through the encoder, possibly affecting the latent representation. We find that some sort of imputation is necessary for the VAE to handle bad pixels well, and we use the iterative PCA procedure presented in Yip et al. (2004a). One could imagine an iterative VAE procedure where bad pixels are imputed with this VAE in order to train a second VAE. In Section 3.4, we find that degrading spectra down to a pixel S/N of 8 does not drastically affect the latent means. However, we find that the latent variance does not grow when the S/N is degraded, meaning that the VAE is reporting the same variance in latent space regardless of S/N. This behavior is undesired because the VAE is underpredicting the uncertainty in latent space for low-S/N spectra. The pixel-level noise only appears in the reconstruction loss and is not included as a feature that the VAE can learn the variance from. Including the pixel-level noise as a feature may allow the VAE to learn appropriate latent variances for low-S/N spectra.

Just as can be done in PCA and LLE, the VAE can be extended to include measurements beyond spectra, e.g., fluxes or shape parameters. Some procedure to normalize the data is necessary for any dimensionality reduction technique. Treating the reconstruction loss as a negative log likelihood offers a natural prescription: additional measurements should be added to the loss function with inverse-variance weights.

Augmenting the VAE using multi-task learning (Caruana 1993) could improve the quality of the VAE latent representations as well as the accuracy of classification or regression tasks on the input spectra. In multi-task learning, one neural network is trained to perform multiple related tasks, which often yields improvements compared to training a separate network for each task. Since a set of neurons is being used for all tasks, this set of neurons tends to learn a representation that generalizes well across tasks, avoiding overfitting on any one task. Our VAE could be augmented into a semi-supervised AE (Haiyan et al. 2015; Zhuang et al. 2015) that also performs an additional task like classification or regression (e.g., on star formation rate). The network that performs the additional task could take the encoder's output as its input, sharing no weights with the decoder and producing its own output. We note that training the VAE and then training a network for the additional task that uses the VAE latent means as features is not multi-task learning. Instead, training the encoder, decoder, and the additional task simultaneously (e.g., alternating tasks every epoch) would constitute multi-task learning.

5. Conclusion

Efficiently exploring and classifying large astronomical data sets is an important problem that can be addressed with dimensionality reduction techniques. In this work, we have demonstrated that a VAE can be used to reduce the dimensionality of SDSS spectra to six latent parameters while retaining enough information to accurately reconstruct individual spectra. Due to its nonlinear behavior, the VAE can capture nonlinear features, such as line widths, with fewer parameters than PCA. Unlike line-ratio diagnostics, the VAE also uses the continuum information in the spectrum. The VAE latent space is interpretable and shows clear separations between different classes of galaxies, even though the VAE was never given these classifications in training. Tracks in latent space yield sequences of spectra whose physical properties vary smoothly, and unusual objects can be identified as outliers within the latent space. While even a six-parameter latent space is difficult to fully visualize and interpret, reducing

[6] Available at https://galaxyportal.space/.

the dimensionality of the spectra makes them more amenable to both computational techniques and human scrutiny. This latent space can be more fully explored using, for example, clustering techniques, outlier searches, and interactive visualizations. VAEs are a fast dimensionality reduction technique that yields compact, interpretable latent spaces and has the potential to enable rapid exploration and classification of large astronomical data sets.

S.K.N.P. acknowledges support from the DIRAC Institute in the Department of Astronomy at the University of Washington. The DIRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences, and the Washington Research Foundation. J.R.V. acknowledges support from FONDECYT postdoctoral grants 3160772 and the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS. The authors also acknowledge partial support from NSF grants AST-1715122 and OAC-1739419.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max Planck Institute for Astronomy (MPIA), the Max Planck Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Facility: Sloan.

Software: astropy (The Astropy Collaboration et al. 2013, 2018), scikit-learn (Pedregosa et al. 2011), AstroML (VanderPlas et al. 2012), pytorch (Paszke et al. 2017).

ORCID iDs

Stephen K. N. Portillo https://orcid.org/0000-0001-8132-8056
Andrew J. Connolly https://orcid.org/0000-0001-5576-8189

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Almeida, J. S., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. 2010, ApJ, 714, 487

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5

Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038

Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, ACM SIGMOD Record, 29, 93

Caruana, R. A. 1993, in Machine Learning Proc. 1993, ed. P. Utgoff (San Mateo, CA: Morgan Kaufmann), 41

Chardin, J., Uhlrich, G., Aubert, D., et al. 2019, MNRAS, 490, 1055

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, ed. E. Simoudis, J. Han, & U. Fayyad (Palo Alto, CA: AAAI Press), 226, https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Folkes, S. R., Lahav, O., & Maddox, S. J. 1996, MNRAS, 283, 651

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA: MIT Press)

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. 2007, in Advances in Neural Information Processing Systems 19, ed. B. Schölkopf, J. C. Platt, & T. Hoffman (Cambridge, MA: MIT Press), 513, http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332

Haiyan, W., Haomin, Y., Xueming, L., & Haijun, R. 2015, in 2015 Int. Conf. on Computational Intelligence and Communication Networks (CICN), ed. J. Agrawal et al. (Los Alamitos, CA: CPS), 1424

Higgins, I., Matthey, L., Pal, A., et al. 2017, in Int. Conf. on Learning Representations, ed. Y. Bengio & Y. LeCun (ICLR), https://openreview.net/forum?id=Sy2fzU9gl

Iwasaki, H., Ichinohe, Y., & Uchiyama, Y. 2019, MNRAS, 488, 4106

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121

Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961

Kewley, L. J., Nicholls, D. C., & Sutherland, R. S. 2019, ARA&A, 57, 511

Kingma, D. P., & Ba, J. 2015, in Int. Conf. on Learning Representations, ed. Y. Bengio & Y. LeCun arXiv:1412.6980

Kingma, D. P., & Welling, M. 2013, in Int. Conf. on Learning Representations, ed. Y. Bengio & Y. LeCun arXiv:1312.6114

Kullback, S., & Leibler, R. A. 1951, Ann. Math. Statist., 22, 79

Lawlor, D., Budavári, T., & Mahoney, M. W. 2016, ApJ, 833, 26

Li, X.-R., Pan, R.-Y., & Duan, F.-Q. 2017, RAA, 17, 036

Lu, H., Zhou, H., Wang, J., et al. 2006, AJ, 131, 790

Ma, Z., Xu, H., Zhu, J., et al. 2019, ApJS, 240, 34

Meusinger, H., Brünecke, J., Schalldach, P., & in der Au, A. 2017, A&A, 597, A134

Meusinger, H., Schalldach, P., Scholz, R.-D., et al. 2012, A&A, 541, A77

Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, NatAs, 2, 151

Osterbrock, D. E., & de Robertis, M. M. 1985, PASP, 97, 1129

Paszke, A., Gross, S., Chintala, S., et al. 2017, in 31st Conf. on Neural Information Processing Systems, ed. I. Guyon & U. Luxburg (Red Hook, NY: Curran Associates), 4

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825, http://www.jmlr.org/papers/v12/pedregosa11a

Reis, I., Poznanski, D., Baron, D., Zasowski, G., & Shahaf, S. 2018, MNRAS, 476, 2117

Reis, I., Rotman, M., Poznanski, D., Prochaska, J. X., & Wolf, L. 2019, arXiv:1911.06823

Richards, J. W., Freeman, P. E., Lee, A. B., & Schafer, C. M. 2009, ApJ, 691, 32

Smee, S. A., Gunn, J. E., Uomoto, A., et al. 2013, AJ, 146, 32

Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810

The Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123

The Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Tröster, T., Ferguson, C., Harnois-Déraps, J., & McCarthy, I. G. 2019, MNRAS: Lett., 487, L24

Tsang, B. T.-H., & Schultz, W. C. 2019, ApJL, 877, L14

Vanderplas, J., & Connolly, A. 2009, AJ, 138, 1365

VanderPlas, J., Connolly, A. J., Ivezic, Z., & Gray, A. 2012, in 2012 Conf. on Intelligent Data Understanding, ed. K. Das, N. V. Chawla, & A. N. Srivastava (Piscataway, NJ: IEEE), 47

Wagstaff, K. L., Lanza, N. L., Thompson, D. R., Dietterich, T. G., & Gilmore, M. S. 2013, in Proc. AAAI Conf. Artificial Intelligence, Vol. 27, ed. S. Kambhampati (Palo Alto, CA: AAAI Press), 905

Yang, T., & Li, X. 2015, MNRAS, 452, 158

Yip, C. W., Connolly, A. J., Szalay, A. S., et al. 2004a, AJ, 128, 585

Yip, C. W., Connolly, A. J., Vanden Berk, D. E., et al. 2004b, AJ, 128, 2603

York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579

Zhao, S., Song, J., & Ermon, S. 2019, in Proc.AAAI Conf. Artificial Intelligence, Vol 33, ed. P. Stone (Palo Alto, CA: AAAI Press), 5885

Zhuang, F., Luo, D., Jin, X., et al. 2015, in 2015 IEEE Int. Conf. on Data Mining, ed. C. Aggarwal et al. (Piscataway, NJ: IEEE), 1141