Autoencoders and Quasar Emission Lines: Using New Techniques to Solve an Old Problem

Collin McLeod*, Alexander Kerr, Karen Leighly

University of Oklahoma Homer L. Dodge Department of Physics and Astronomy
*collin.mcleod@ou.edu

Abstract

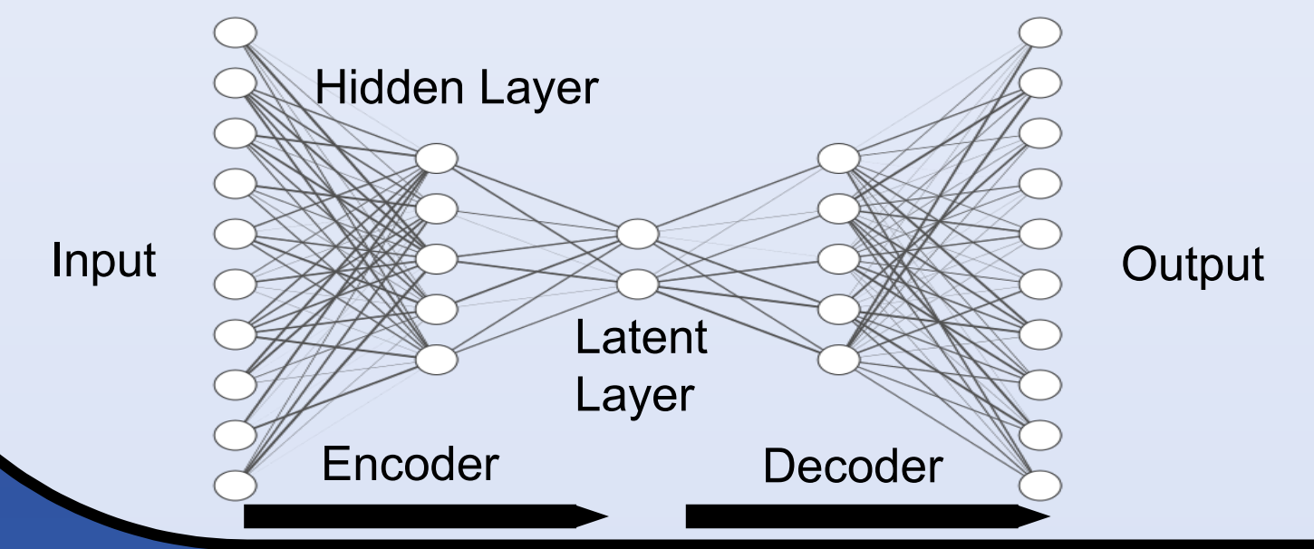
Despite over thirty years of study, no existing model can adequately explain the differences in emission line properties of quasars. In the rest-frame optical and UV, spectra of quasars show the same lines but with different properties. The changes in line width, ratios, and other factors follow certain patterns that are not fully understood, but may be related to physical properties such as accretion rate and black hole mass. Applying modern analysis techniques to characterize these patterns may help us learn more about the relations between quasars.

A widely used method for studying variance is principal component analysis (PCA). While PCA is a powerful technique, it is a linear analysis, and is therefore limited when used to investigate nonlinear variations (the widening of emission lines, for example). Autoencoders offer a potential alternative to PCA. An autoencoder is composed of two neural nets, trained through unsupervised learning to reduce data to a small set of latent parameters, then reconstruct it to mirror the original input. Much like PCA, autoencoders can reduce the dimensionality of a data set. Because autoencoders may be nonlinear, however, they should model nonlinear variations more effectively than PCA.

We use the Python package Tensorflow to create autoencoders to model quasar spectra, training them on quasar spectral data sets as well as on synthetic data sets constructed to emulate the variations in observed spectra. We compare its performance to PCA to determine whether the autoencoder models nonlinear variations more accurately. We investigate the correspondence of latent parameters to spectral features, and potentially their relationship to physical parameters. Finally, we test the use of the autoencoder as a generative model to create realistic synthetic spectra.

Introduction

- **Hypothesis:** Quasars show substantial variation in the shape and size of their emission lines. The lack of a unified physical model to describe these variations presents a problem in spectral fitting. In order to study absorption in broad absorption line quasars (e.g., Leighly et al. 2018), the emission properties must first be accounted for. We hypothesize that the variation in emission properties can be explained by a small number of parameters. We suggest that these consist of some linear variation, i.e., a linear combination of template spectra, and some nonlinear variation, namely the Doppler broadening of emission lines due to the gas in the broad line region revolving about the central black hole.
- **Principal component analysis** is a popular technique for reducing dimensionality. PCA extracts orthogonal vectors called principal components which account for the largest variance in a data set. PCA can be applied to the emission line problem, but it has limitations. Notably, PCA is a purely linear transformation, and will therefore fail to accurately recreate the nonlinear variation caused by line broadening.
- **Autoencoders** are a type of neural net commonly used to reduce dimensionality. An autoencoder maps each input vector to a lower dimensional latent space, using a fixed number of variables, then maps the latent representation of each vector back to the original vector space. In this way, an autoencoder learns a lower dimensional representation of the data. **Variational autoencoders (VAEs)** are similar to traditional autoencoders except they constrain the data's latent variables to follow normal distributions. Neural nets are not necessarily linear, and can approximate a variety of functions. Because of this, they may be a better choice for modeling quasar emission, particularly line broadening. We have built our autoencoder models using Tensorflow 2.0 and probability layers from Tensorflow Probability.
- **Goal:** Our goal is to create an autoencoder model which accurately describes the variance in quasar emission in a nonlinear way, and to compare its performance to PCA. Potential uses for such a model include modeling emission properties in spectral fitting algorithms (like the spectral synthesis code SimBAL [Leighly et al. 2018]), or identifying the intrinsic parameters which drive variation in quasar emission.



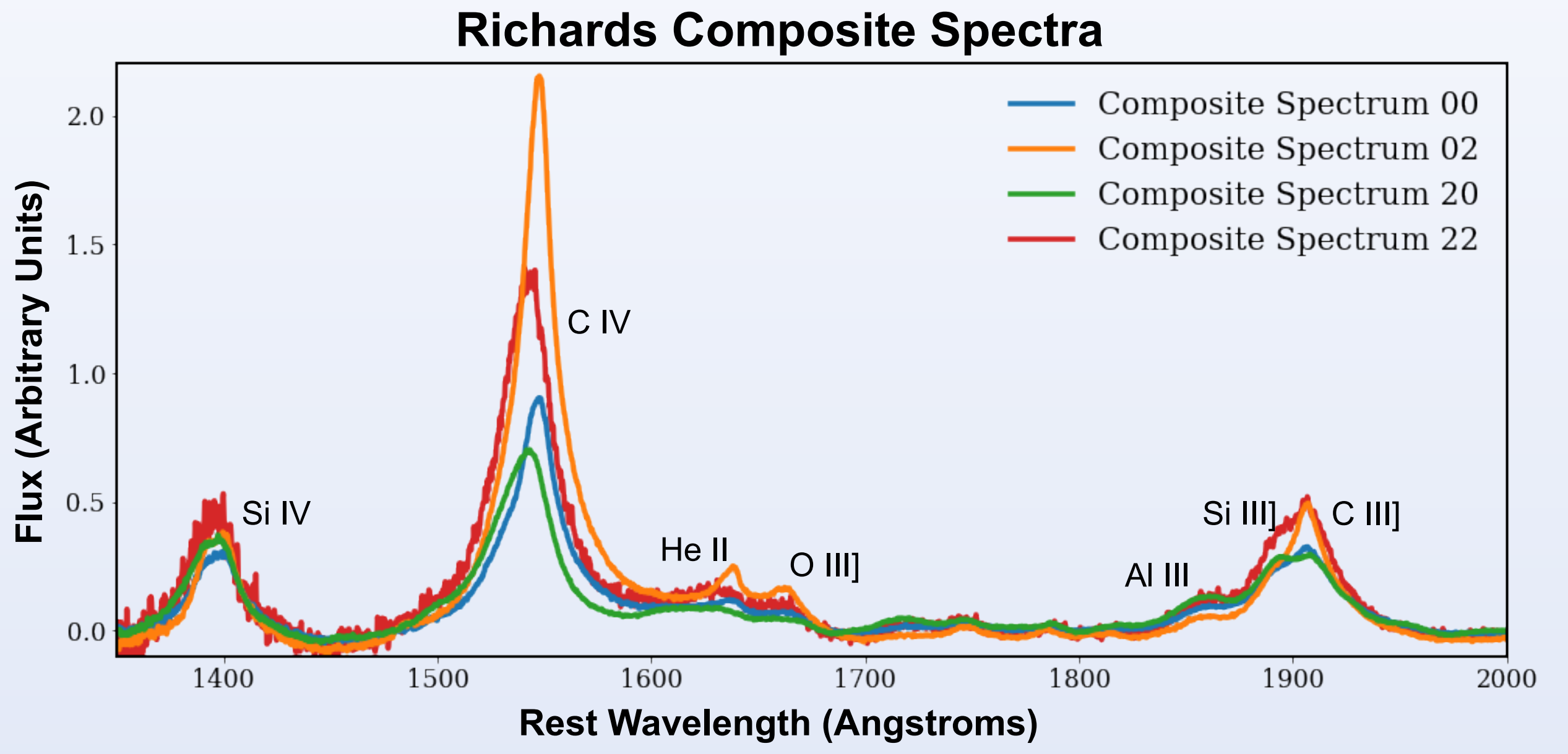
Left: A visual representation of the various fully-connected layers in an autoencoder.

Creating Data for Training

In order to test how an autoencoder will model linear and nonlinear variations, a synthetic dataset was constructed which fits our hypothesis. The synthetic data is defined by three parameters, two linear and one nonlinear.

- **Richards 1 and Richards 2:** a bilinear interpolation based on fits (calculated by Wagner et al. [poster 2015.15]) of composite spectra constructed by Richards et al. (2011), which primarily vary in C IV blueshift and equivalent width. These two parameters dictated the linear variance.
- **Broadening:** the third parameter was a broadening term. A velocity width was added in quadrature to the full width at half maximum of the spectrum's emission lines. This parameter defined the nonlinear variance.

Richards Composite Spectra



Composite spectra from Richards et al. 2011. In the constructed synthetic data set, Richards 1 interpolates from spectrum 00 to 02, while Richards 2 interpolates from 20 to 22.

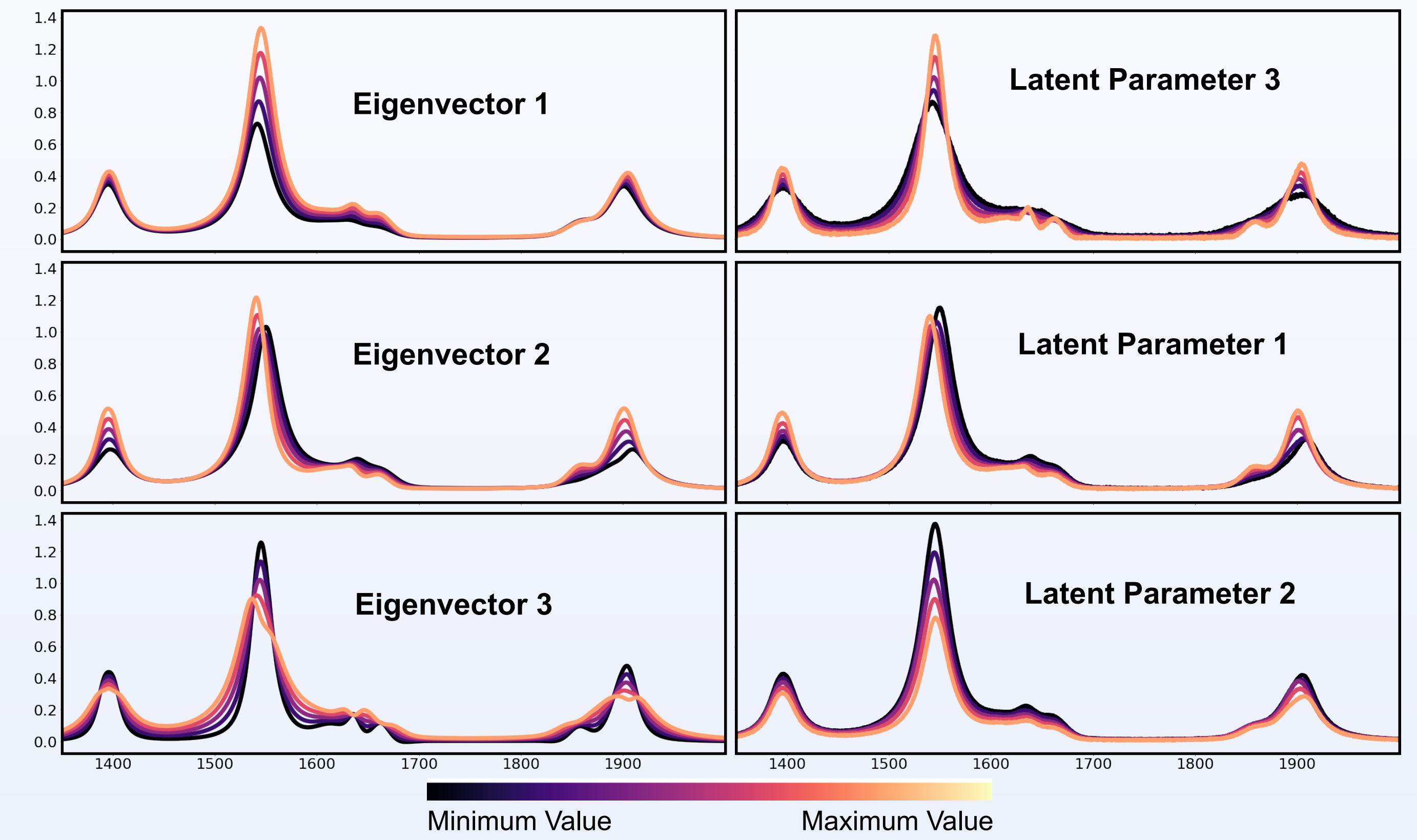
The final dataset was made up of 5000 synthetic spectra, each of which had 3000 wavelength bins, constructed from random values of the three parameters. The autoencoder used to model these data was a variational model with 3 latent parameters. It was composed of a sequence of fully-connected layers, with a single hidden layer of size 1000.

Training the Autoencoder with Synthetic Data

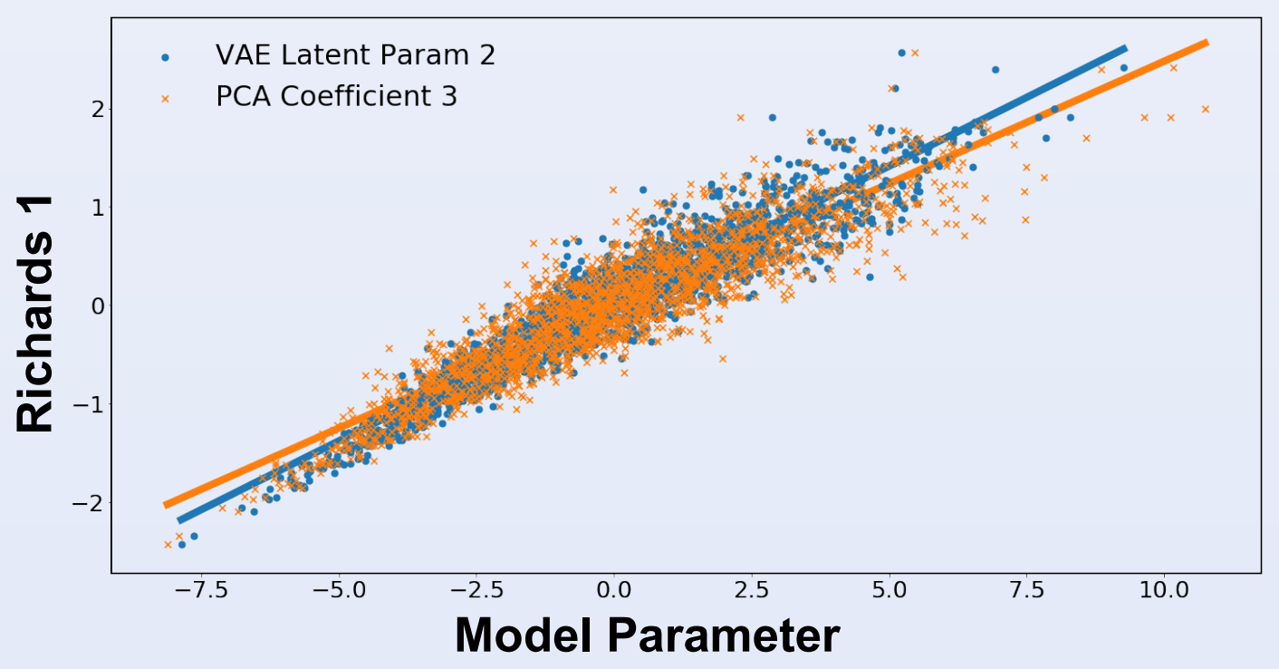
The synthetic data set was randomly split into training and testing sets for training the autoencoder. It was trained for 60 epochs on the training set. Principal component analysis was also calculated on the training set. While PCA's principal components are ordered so that component 1 accounts for more variance than any other, the autoencoder's latent parameters are unordered.

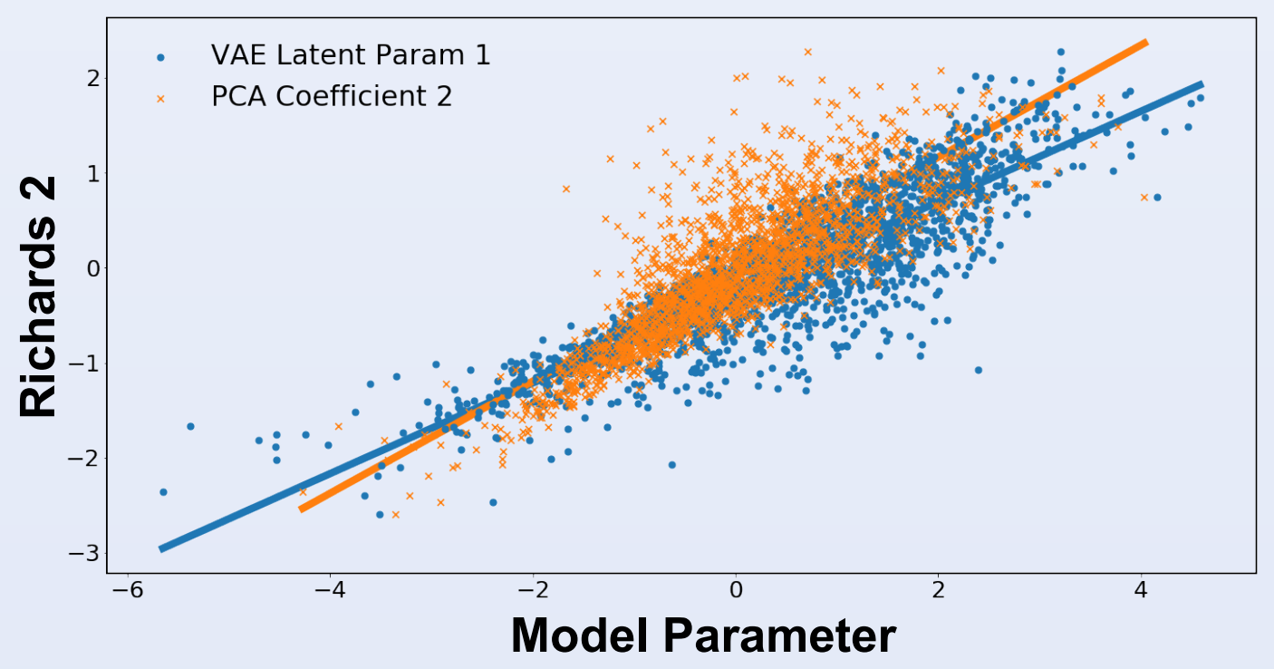
Also unlike PCA, an autoencoder does not generate easily visualized eigenvectors. Instead, we show the effect of varying one latent parameter while the others are held fixed (shown in the right panels below). Similarly, the left panels show the effect of adding successively larger multiples of each principal component. In this plot, it is clear where the autoencoder and PCA are modeling the same variance. The x axis shows wavelength in angstroms, and the y axis shows flux in arbitrary units.

Varying PCA Eigenvectors / Varying Autoencoder Latent Parameters



Both methods appear to model the underlying parameters which define the data set. The below plots show the identification between model parameters (projections onto principal components for PCA, and encoded [latent] representations for the autoencoder) and intrinsic parameters. The autoencoder's third latent parameter and PCA's first principal component are both correlated with the broadening term used in the construction of the synthetic spectra.

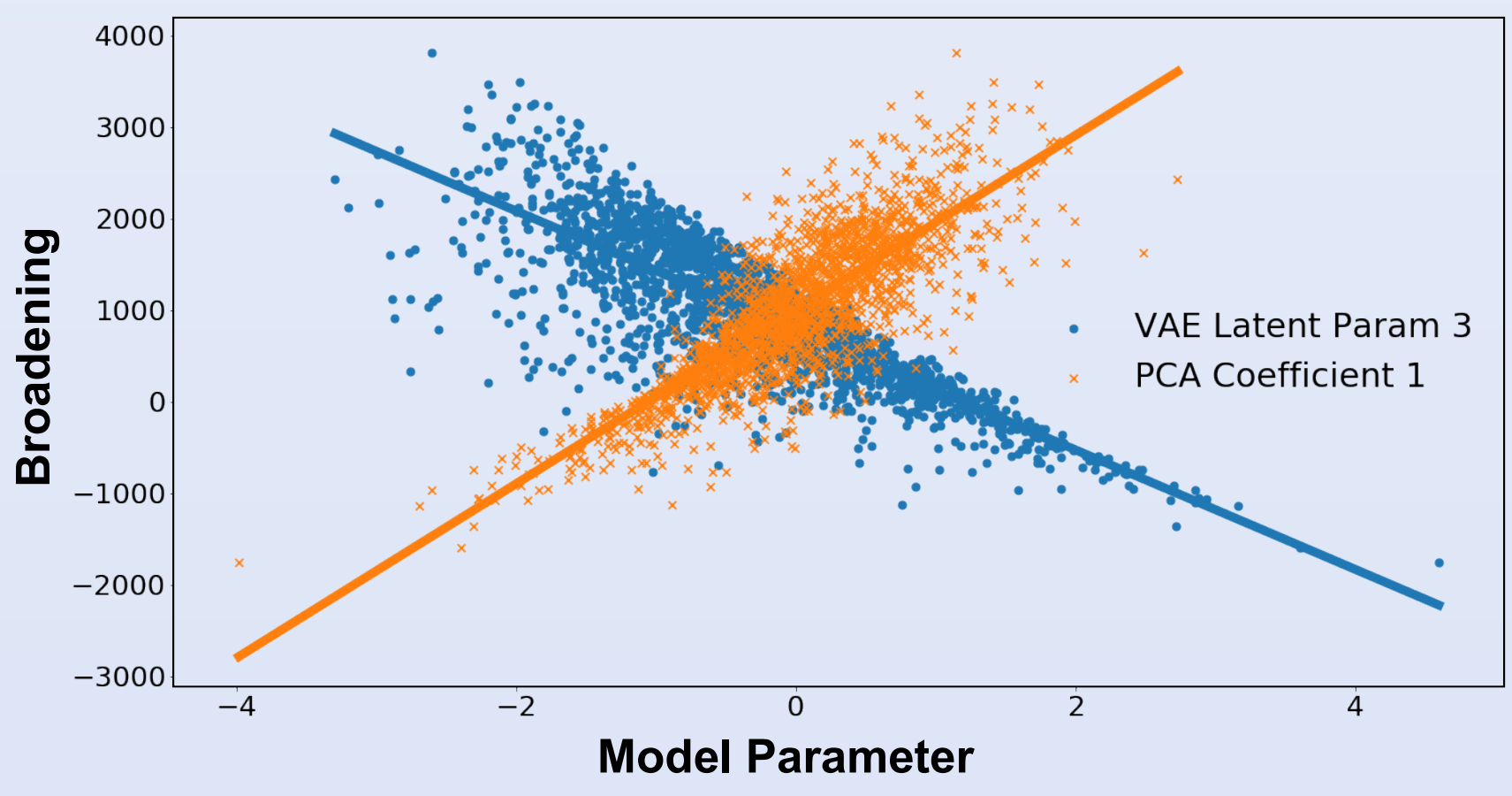




The two models do not recreate broadening in quite the same way, which can be seen in the accuracy of their reconstructions of spectra.

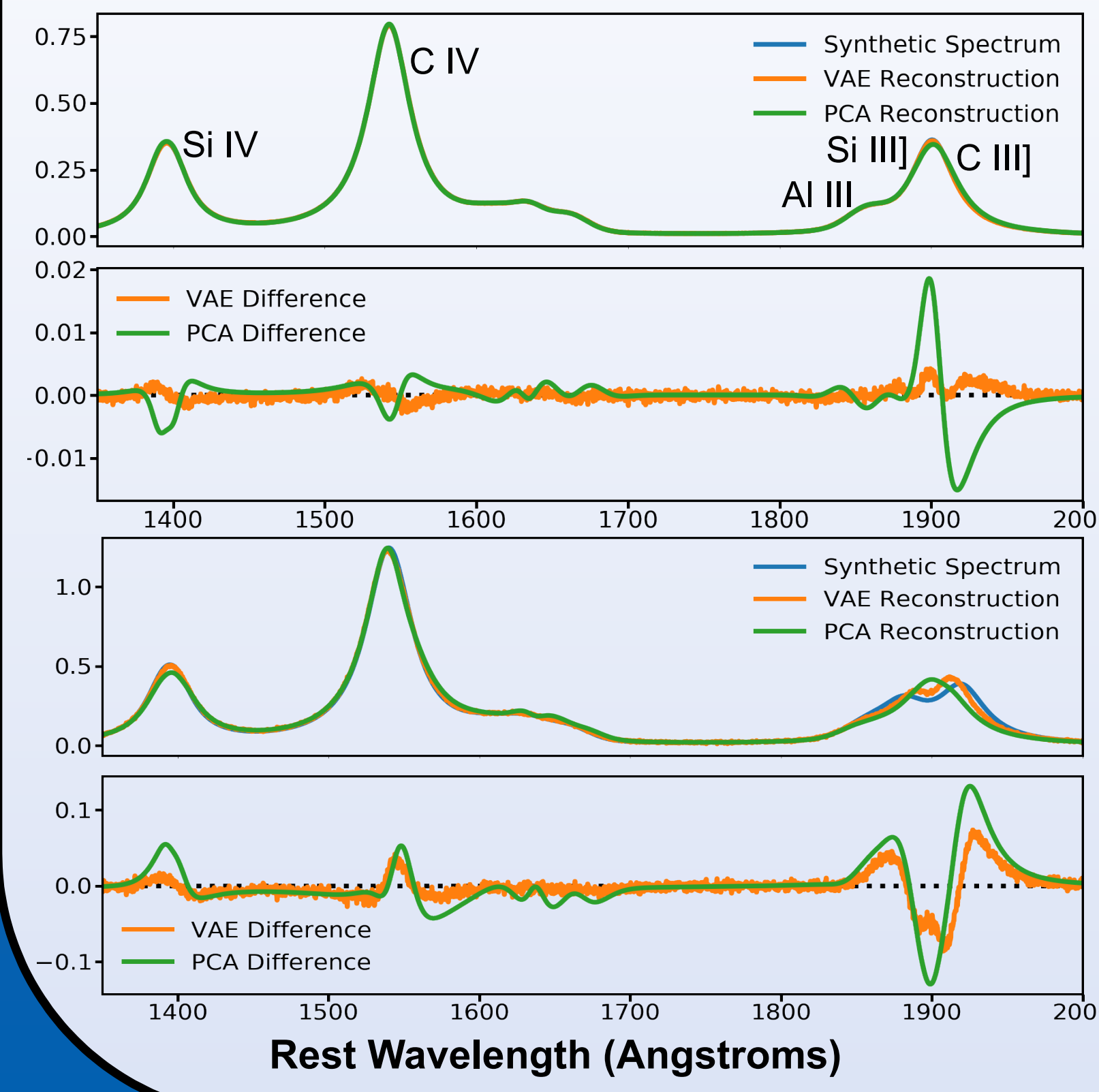
Autoencoder Reconstruction Error (Reduced Chi-Squared): 0.0373

PCA Reconstruction Error (Reduced Chi-Squared): 0.12015



Results from Synthetic Data

While both methods identify the underlying parameters, the autoencoder recreated the data more accurately. The autoencoder routinely achieved reconstruction errors (chi-squared error) 2-5 times lower than PCA. Below is a typical example of a synthetic spectrum which demonstrates where the autoencoder fit the data more accurately than PCA. Many of these arise from the failure of PCA to model the broadening accurately, where characteristic "w" shapes indicate that PCA is approximating the broadening with a linear component.

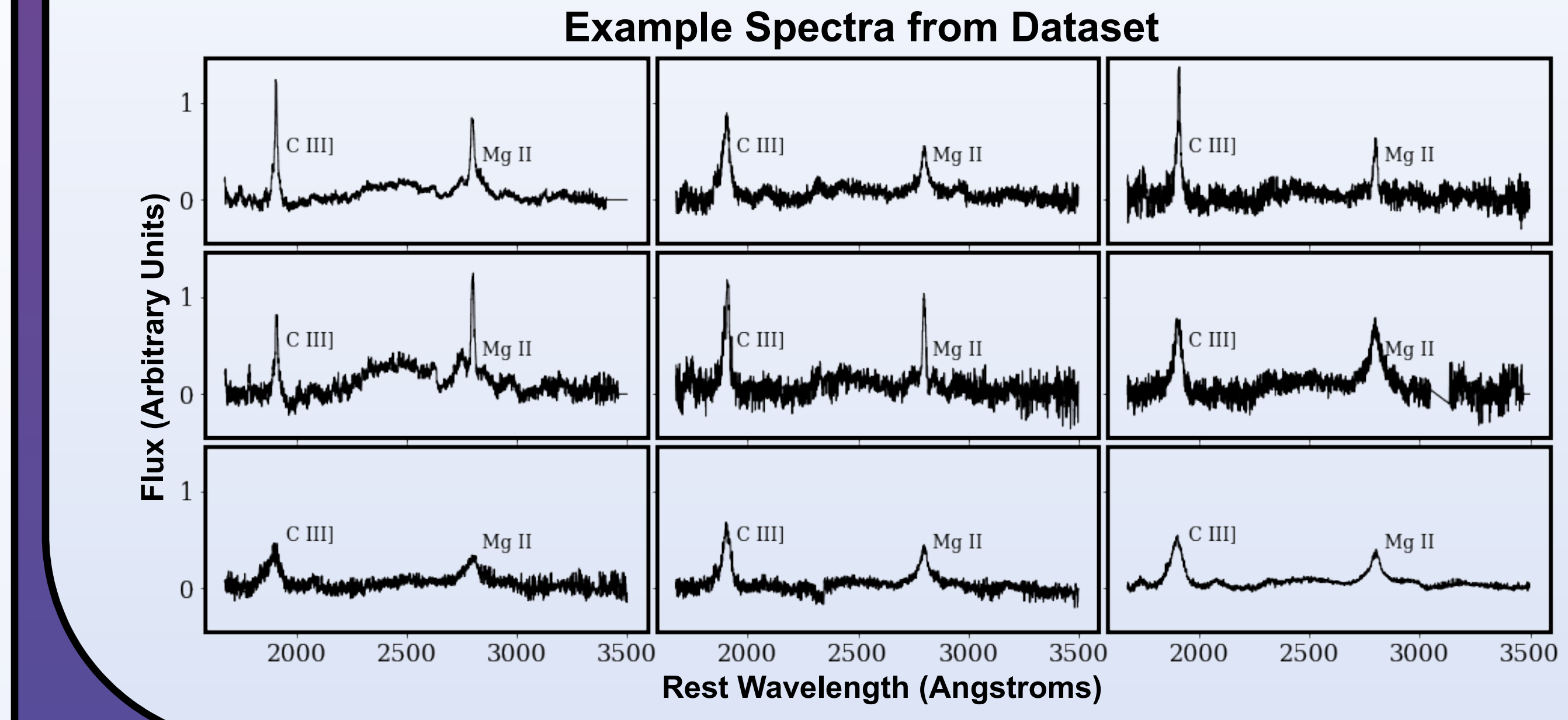


Left: spectra which are reconstructed well by the autoencoder, but poorly by PCA. Each plot shows the synthetic spectrum and both reconstructions above the difference between the real spectrum and each of the two reconstructions. PCA is particularly bad at correctly recreating the height of the Si IV line. In both graphs, PCA fails to model the blended line correctly. In the bottom graph, both PCA and the autoencoder struggle to recreate the complicated shape of the blended line, but the autoencoder is closer to the true spectrum.

Collecting Data for Training

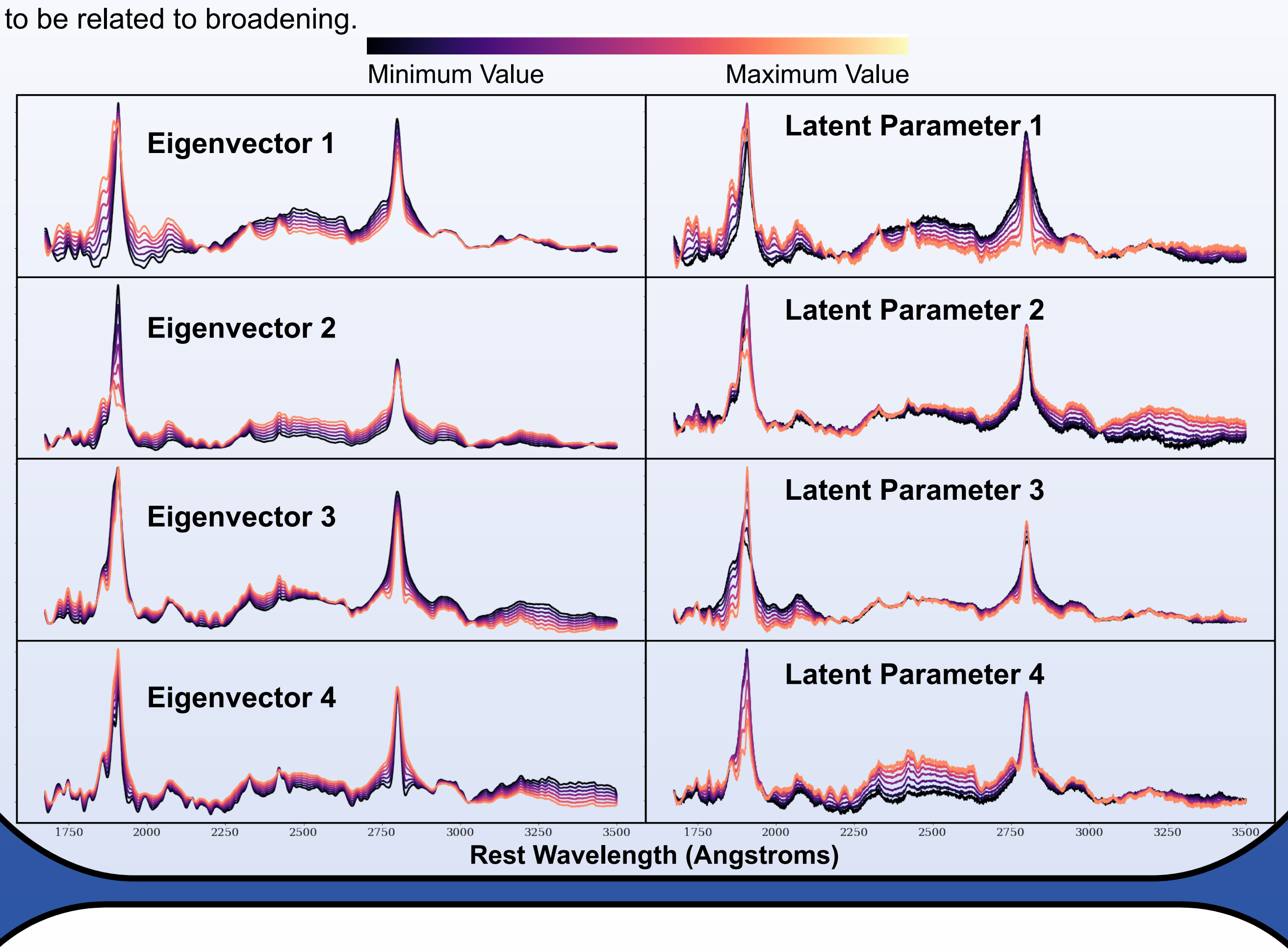
To test an autoencoder model with real data, we used an available sample of processed spectra. The parent sample was drawn from SDSS DR4 quasars with redshifts $1.2 < z < 1.8$ and identified by eye as having relatively narrow MgII lines and strong FeII emission. The spectra were fit between 2200-3050 Angstroms with a power law, MgII emission lines, an FeII pseudo-continuum template, and several weaker lines following the method of Leighly & Moore 2006. The final samples have signal-to-noise ratios between 2200-2600 Angstroms greater than the sample median, and are characterized by MgII FWHM < 4000 km/s (5557 spectra). A number of objects are plotted below, demonstrating a range of line widths.

Example Spectra from Dataset



Training on the Real Data Set

The autoencoder trained on the real data had two hidden layers, with sizes 1024 and 512. A latent dimension of 4 was chosen to match the number of eigenvectors used in the spectral synthesis code simBAL (Leighly et al. 2018). It was trained for a total of 80 epochs on the full set of spectra. PCA was also calculated on the full data set. Plotted below are the effects of varying each parameter, as before. PCA eigenvectors 1, 3, and 4, as well as latent parameters 1 and 3, appear to be related to broadening.



Preliminary Results and Continuing Work

- **Modeled Parameters:** Eigenvector 1 and latent parameter 1 show extremely similar effects: broadening Mg II, narrowing C III], and attenuating the continuum between them. Other similarities between the latent parameters and eigenvectors are also apparent: parameter 2 and eigenvector 4, parameter 4 and eigenvector 2, and parts of others. Parameter 3 (above) indicates that broader Mg II emission is correlated with shorter, broader C III] emission, which fits our hypothesis that Doppler motion broadens all the emission lines.
- **Accuracy:** On this dataset, the autoencoder is not noticeably more accurate than PCA, achieving a final reduced chi-squared reconstruction error of 0.0356, compared to 0.0365 for PCA, unlike our result on the synthetic data. We have possible explanations for this:

1) The real data are generally low to medium signal-to-noise objects. The substantial noise across the data set may prevent efficient training of the autoencoder.
2) Our hypothesis may be incorrect with regard to these data: it is possible that the variance is primarily linear, and the autoencoder is approximating PCA.

- **Issues and Future Work:**

1) Our current work involves models which do not account for the error in the data. Our upcoming work will involve using models which weight data points based on their error, so that high-error points are less weighted in the fit.
2) The autoencoder sometimes converges to local minima, where its performance is poor, but training is unable to further minimize the loss function. The variational model, which is more robust to this issue, may improve our results.

References

- Abadi, Martin, et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- Leighly, Karen M., & Moore, J. R., 2006, ApJ, 644, 748
- Leighly, Karen M. et al., 2018, ApJ, 866, 7
- Richards, Gordon T. et al., 2011, AJ, 141, 167
- Wagner et al., 2017, 229th AAS Meeting, id.250.15