Determining the Composition of Exoplanet Atmospheres with Machine Learning

Team Members: Abhishek Amol Mishra, Aman Verma, Dylan Nguyen
Research Lead: Divya Gollapalli | Faculty Advisor: Professor Rishabh Iyer

Abstract

Hundreds of thousands of exoplanets have been discovered, yet our knowledge of the structure and composition of these exoplanets is nowhere near as expansive. Recent research has determined that machine learning would be far more efficient than traditional methods for atmospheric retrieval, but creating a quick, robust, and independent model is an ongoing challenge. We present **ExoNeural** - a new machine learning approach to determining the composition of exoplanet atmospheres that could potentially be more efficient and robust than current models.

Introduction

The traditional way to analyze the atmosphere of an exoplanet is through a look-up table combined with a physics-based forward model. Forward models use the laws of molecular physics to determine a planet's emitted radiation from the atmospheric makeup of the planet. Recently, neural networks have been applied to directly predict atmospheric gas abundances from emitted radiation. One of the first examples of this was Waldmann in 2016 [3].

Data Collection

We collected our data from the NASA Astrobiology II Team that collaborated with Google Cloud to work on a project similar to ours. We were able to access our dataset through the cloud, accessing the Google Cloud public bucket to get **10,000** planets with their spectra in **4,378** columns along with the atmospheric composition of **12** key gases. We chose five of those gases to have our model predict: CO_2, H_2O, N_2, CH_4, and O_2.

While the data set was clean, there was one significant obstacle. 10,000 rows was not enough data to learn nuanced patterns and, as a result, we used a **Gaussian Copula Model** to be able to synthetically create more data.

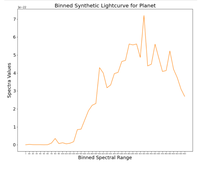

Figure 1. Binned Synthetic Light Curve

Our synthetic data generation was very successful as our **two-sample Kolmogorov–Smirnov test** gave a high score of **0.88**. After generating synthetic data, we binned the features and left the real data set for testing and used the synthetic data set for training and validation.

Model

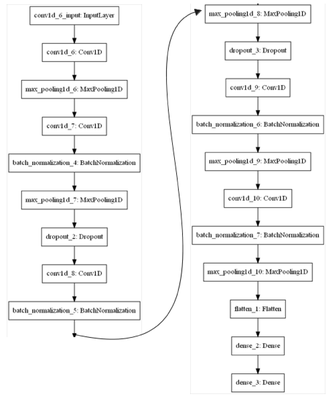

Figure 2. Model Architecture

In our modelling efforts, we used a 1D CNN. We chose this neural network because our input was composed of **45** wavelength bins of measured radiation. To implement this, we used the *keras* and *tensorflow* libraries.

In our model, we utilized **pooling** layers after our convolutional layers to reduce dimensionality. We also applied **batch normalization** layers to have normalized data along and **dropout** layers to prevent neurons from arriving to the same conclusion—essentially making it harder for them to learn.

After a series of such layers, we flattened the model with a **flatten** layer and put it through a **dense** layer to predict our five compositions.

Results

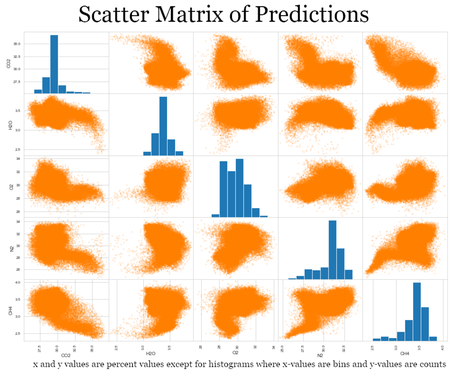

Figure 3. Scatter Matrix of Predictions

Drawing a scatter matrix of the five elements, we can begin to see our modelling gave us our predictions that had a **unimodal distribution**. The predictions centered around a value had small deviations away from that as seen in the histogram.

Our mean absolute error was **1.004** however, in our training, the MAE dropoff was very small.

Analysis and Comparison

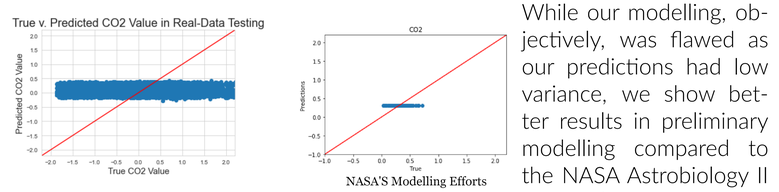

Figure 4. Scatter Matrix of Predictions

While our modelling, objectively, was flawed as our predictions had low variance, we show better results in preliminary modelling compared to the NASA Astrobiology II Team who attempted to do something similar.

Comparing our modelling efforts and their modelling efforts, we see that we arrived at the same obstacle as them in terms of predictions. In their paper, they were able to overcome this obstacle through the variance in their input features. That being said, we got a healthier variance which means with a better dataset, our model can perform better automatically.

Future Endeavors

The quality of our dataset was a big thing that we could work on in the future. In our preprocessing, we applied a quality assessment with comparing the lightcurves' mean distribution among certain bins of an element. When the mean distributions were compared, the differences were negligible. Thus, a lack of variance in our dataset meant that the quality of our dataset was poor, but in the future, that's something we can focus on.

Seeing that this was the only available dataset that gave us exoplanet spectra **and** atmospheric elemental composition, future endeavors will involve gaining access to a more complete dataset. This will allow our model's functionality to be elevated automatically since our current model's flaws are tied down to dataset quality.

References

[1] Matthew C Nixon and Nikku Madhusudhan. Assessment of supervised machine learning for atmospheric retrieval of exoplanets. *Monthly Notices of the Royal Astronomical Society*, 496(1):269–281, 06 2020.

[2] Frank Soboczenski, Michael D. Himes, Molly D. O'Beirne, Simone Zorzan, Atılım Güneş Baydin, Adam D. Cobb, Yarin Gal, Daniel Angerhausen, Massimo Mascaro, Giada N. Arney, and Shawn D. Domagal-Goldman. Bayesian deep learning for exoplanet atmospheric retrieval. In *Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montreal, Canada*, 2018.

[3] I. P. Waldmann. Dreaming of atmospheres. *The Astrophysical Journal*, 820(2):107, mar 2016.

We would like to thank Natalie Hinkel at the Southwest Research Institute and Brianna Lacy at the University of Texas at Austin for their advice and help.