Exploration of exoplanet atmosphere with machine learning

Antonia Vojtekova[1] , Ingo Waldmann[1] , Kai Hou Yip1, Bruno Merín[2]
[1] University College London
[2] European Space Agency (ESA)

Feature Engineering

Exploring the statistical properties of the dataset aids in identifying appropriate data normalisation, detecting outliers, and assessing data complexity. Additionally, we analysed the behaviour of molecules in every layer and discovered patterns when coloured by properties such as C/O ratio, temperature or metallicity.

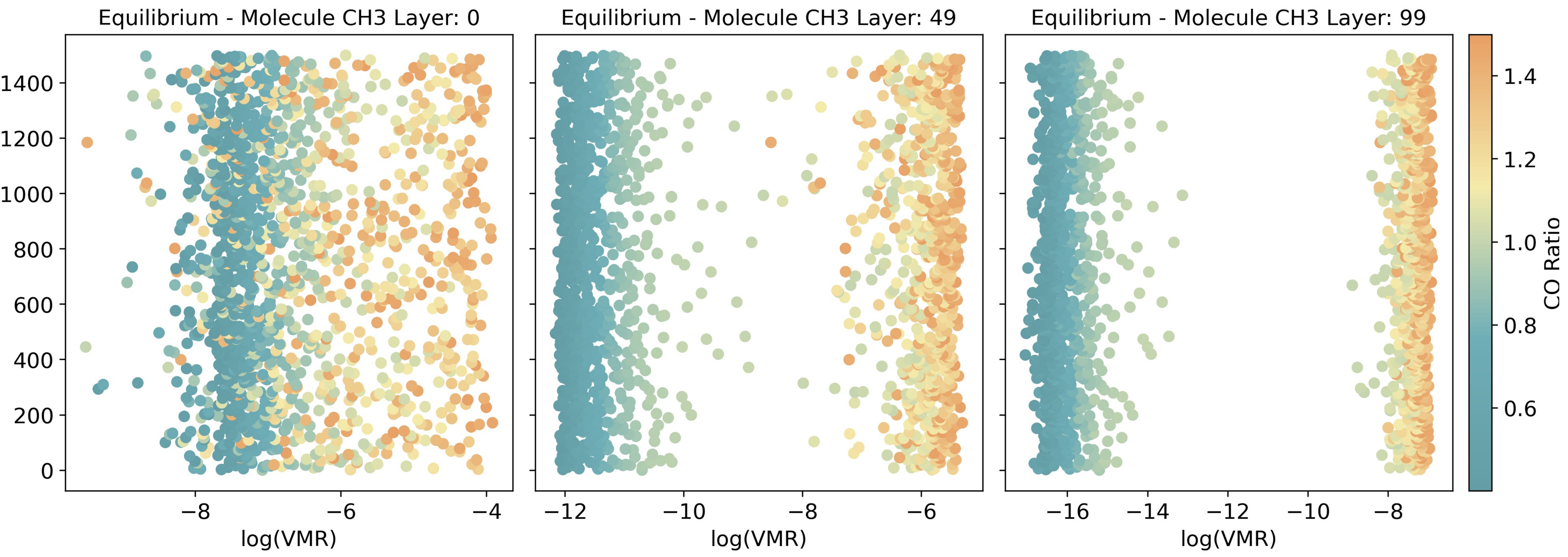
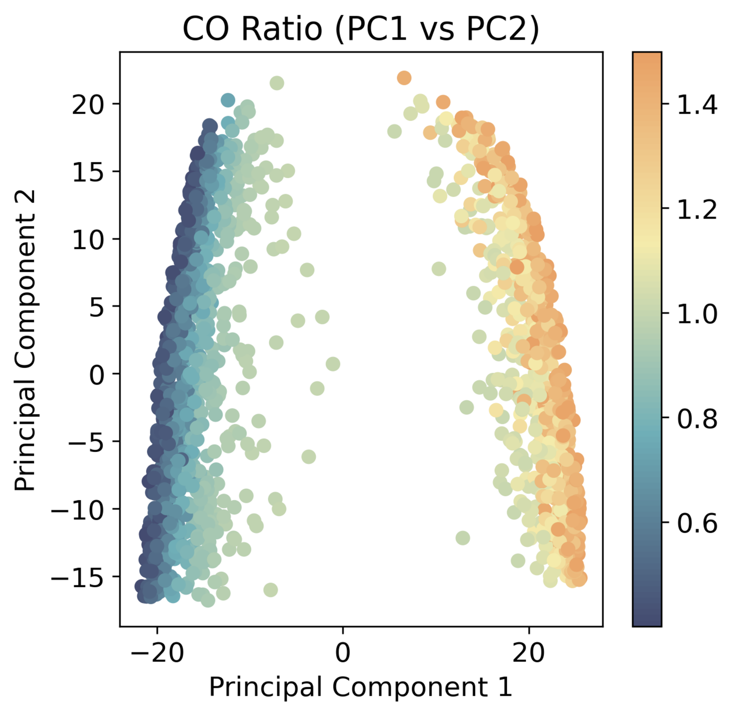


Figure 1: Illustration of Data Patterns for Molecule CH3 in Varied Atmospheric Layers - Demonstrating the Inverse Relationship between Volume Mixing Ratio and Atmospheric C/O Ratio. In this representation, atmospheres with lower C/O ratios exhibit reduced volume mixing ratios in upper layers, while those with higher C/O ratios display higher values, highlighting the distinctive pattern.

Principal component analysis (PCA)

PCA (Hotelling, H., 1933) is a dimensionality reduction technique that transforms data into a new coordinate system to highlight its underlying patterns by capturing the most significant variance in the data.

PCA analysis reveals that for equilibrium data, four principal components effectively capture the interrelationships among C/O ratio, Temperature, and Metallicity. In contrast, disequilibrium data require six components to capture the same interrelationships, indicating higher complexity. These findings highlight PCA's ability to uncover nuanced patterns in chemical data.



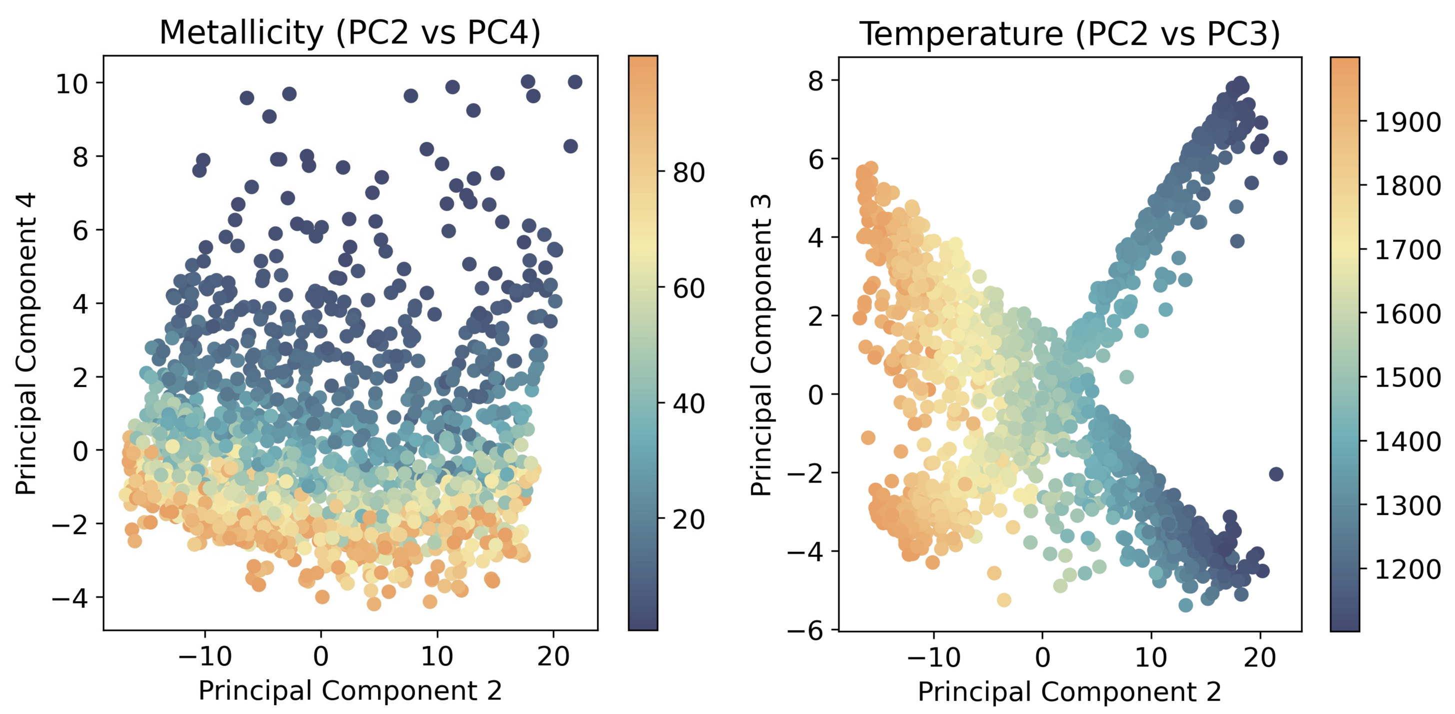


Figure 2: Example of different principal components capturing the relationship for C/O ratio, temperature and metallicity at the equilibrium state

In today's exoplanetary research, we're facing a data revolution, moving towards big data science. To tackle the complexities of exoplanet atmospheres, computational demands are growing. Machine learning is the key to speeding up these calculations without replacing existing models.

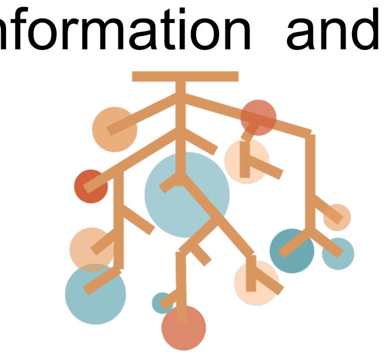
We're presenting a project focused on improving the efficiency of modelling disequilibrium chemistry in exoplanet atmospheres. We aim to develop a machine learning algorithm to accelerate species abundance calculations, shifting from equilibrium (ACE, Agundez et al. 2012) to disequilibrium chemistry (FRECKLL, Al-Refaie et al. 2022). Before diving into deep learning, we've thoroughly analysed our dataset to uncover valuable patterns.

This poster presents our initial exploration of the chemistry dataset.

Random forest (RF) / feature importance

Utilising Random Forest (Breiman, L., 2001) analysis, including the assessment of feature importance, allows us to gain in-depth insights into the relative significance of individual variables in our predictive models. This knowledge guides us in making informed decisions regarding feature selection and model optimisation.

Our objective was to investigate the impact of normalisation and molecule removal on RF's ability to identify key molecules for predicting C/O ratio, temperature, and metallicity. This analysis revealed that the chosen normalisation and dataset reduction methods do not significantly affect the underlying information and patterns within the data.



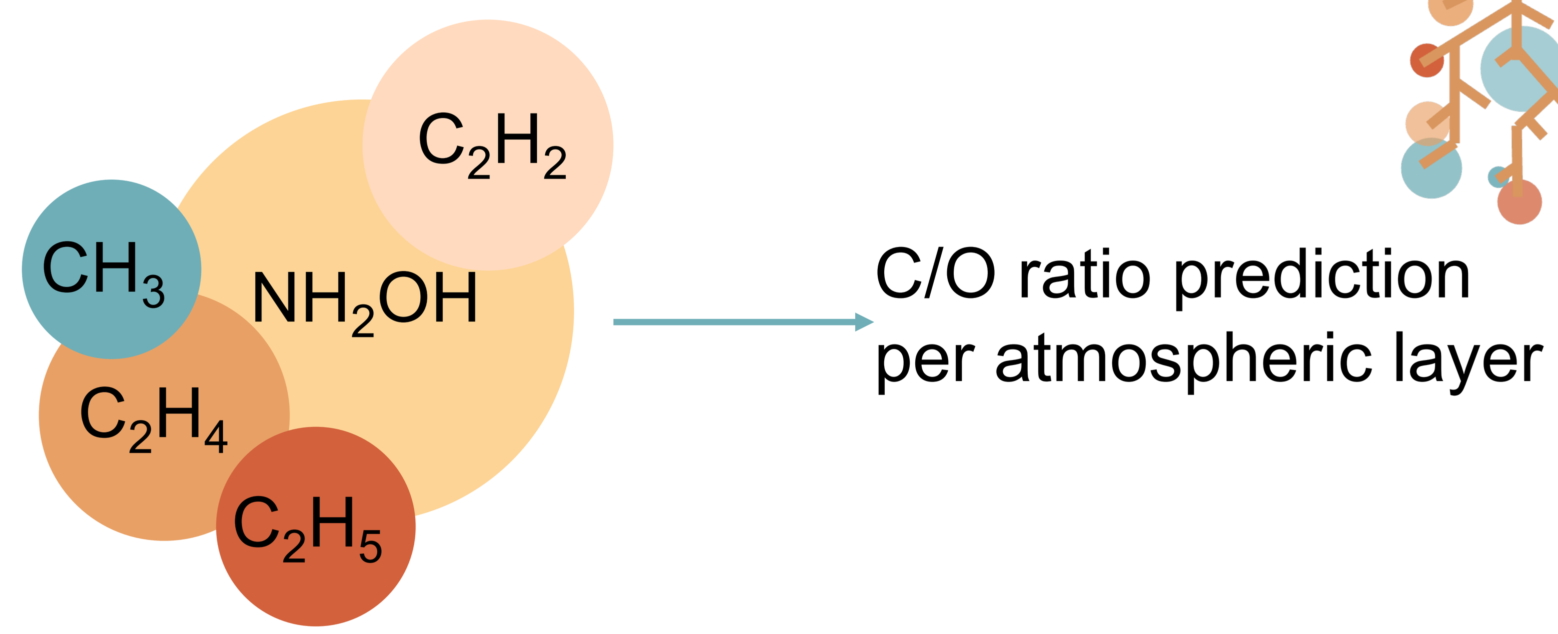


Figure 3: The five most influential molecules to predict the C/O ratio for every layer

Neural networks

Currently, we are exploring convolutional neural networks like U-net to assess their suitability in capturing the equilibrium-to-disequilibrium chemistry transition. Future work involves comparing these networks with more advanced algorithms, including graph neural networks.

Conclusion

Our analysis (statistical techniques, PCA, and RF) improves understanding of key factors in predicting C/O ratio, temperature, and metallicity in exoplanetary atmospheres. Robust data handling methods enhance reliability for future machine learning applications, advancing data-driven exoplanetary research and predictions of crucial atmospheric properties.

References:
- Hotelling, Harold. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24 (1933): 498-520.
- Breiman, L. (2001). Random Forests. Machine Learning. 45. 5-32. 10.1023/A:1010950718922.
- Agúndez, Marcelino et al. "The impact of atmospheric circulation on the chemistry of the hot Jupiter HD 209458b." *Astronomy and Astrophysics* 548 (2012): 1-8.
- Al-Refaie, Ahmed F. et al. "FRECKLL: Full and Reduced Exoplanet Chemical Kinetics distiLLed." (2022).

Contact:
antonia.vojtekova.22@ucl.ac.uk
https://www.linkedin.com/in/antonia-vojtekova/

Disclaimer: This poster was created with the assistance of ChatGPT due to time constraints in pursuing our research endeavours.