
REPORT ON HARNESSING ML FOR ATMOSPHERIC RETRIEVAL OF EXOPLANETS

Swastik Dewan
Tasneem Basra Khan
National Institute of Scientific Education and Research
`swastik.dewan@niser.ac.in`
`tasneembasra.khan@niser.ac.in`

Abstract

The Report consists of the mid-way updates of the project on Harnessing ML for Atmospheric Retrieval of Exoplanets. Atmospheric Retrieval is a method to go deep inside the components present in the atmosphere of the exoplanets, to perform the atmospheric retrieval various traditional methods are used but lead to computational complexity so using machine Learning we can increase performance and generate transmission spectra in a much computationally efficient way. HELA which uses random forest is used here to train and test the data set and retrieval is performed.

1 Targets Achieved:

- Completing the literature review about the Atmospheric Retrieval and details about the Radiative Transfer equation.
- Training and testing the already existing Models (HELA (uses Random forest), POSEIDON(Non-ML-nested sampling method)). Both are Python-based models.

2 Literature Review

2.1 Introduction

Exoplanets are planets that orbit stars other than the Sun outside the solar system.(1) The detection of exoplanets and subsequently studying their atmospheric properties such as the chemical compositions, temperature profiles, clouds/hazes, and energy circulation make up a fascinating area of astronomy, in part because the search for worlds orbiting stars other than our Sun provides a unique opportunity to understand the formation of our solar system's planets and the possible end of our own (Madhusudan,2018).

An exoplanet's spectrum offers a glimpse of its atmosphere. The several interrelated physicochemical processes and characteristics of the atmosphere that are disclosed by their impact on the radiation that emerges from the atmosphere and reaches the observer are encoded in a spectrum. In turn, the retrieved attributes can shed light on the many physical and chemical processes that affect the atmosphere and their development history.

The process of characterizing planetary atmospheres involves determining which model parameters best fit the observed exoplanet spectra. Atmospheric retrieval is primarily concerned with connecting exoplanet spectra to the parameters of intricate forward models of atmospheric physicochemical processes.

Bayesian inference methods such as MCMC, nested sampling, and optimal estimation algorithms have been widely used for exoplanet retrieval. The problem with traditional sampling methods is the huge computational time associated with processing observational data, thus machine learning can speed up the process without compromising the accuracy. many such Machine learning models are available like HELA (random forest), INARA, ExoGan (Neural Networks), etc. Once trained the model gives a much faster result than the traditional methods.(Vasist,2023)

2.2 Radiative Transfer Equation

$$t_{\lambda,i} = \text{EXP} \left(- \sum_{j=1}^{N_{\text{lay}}} \alpha_{\lambda,j} \mathcal{P}_{i,j} \Delta h_j \right) = \text{EXP} \left(- \sum_{j=1}^{N_{\text{lay}}} \Delta \tau_{\lambda,j} \mathcal{P}_{i,j} \right).$$

Figure 1: Radiative Transfer Equation : Transit depth

$$\left(\frac{R_{p,\lambda}}{R_s} \right)^2 = \frac{1}{R_s^2} \left(R_p^2 + 2 \sum_{i=1}^{N_i} [1 - t_{\lambda,i}] b_i \Delta b_i \right),$$

Figure 2: Radiative Transfer Equation to be used.

This equation shows how a linear ray passes across the atmosphere of the exoplanet. log of Transmission is the summation of path distribution over the atmospheric layers. From the transmission, we can evaluate the radius of the exoplanet.

2.3 Nested Sampling with unsupervised machine learning

In this method, unsupervised machine learning is used along with the traditional nested sampling method. In the nested sampling, our main aim remains to find out the prior ranges, to begin with, we give a broad range of prior data across the wavelength, this results in higher computational complexity so we try to minimize the prior range to begin with. K means cluster which is an unsupervised ML algorithm, is used to separate the spectrums into classes each class has a corresponding range of wavelength. Thus k means is used to reduce the prior ranges to ease the method of nested sampling.

3 Methodology

3.1 HELA

HELA uses Random Forest to evaluate the value ranges of the parameter. HELA has used a Data set that contains about 1 lakh of WFC3 transmission spectra of which 80,000 would be used as Training Data sets and 20,000 would be used as testing data sets.

We have used the HELA model to run the training using 80,000 transmission spectra and 20,000 data for testing the data. The plot R^2 Score which checks the accuracy between the predicted and real value is given below, The $R^2 Score = 1$ is considered to be the most accurate.

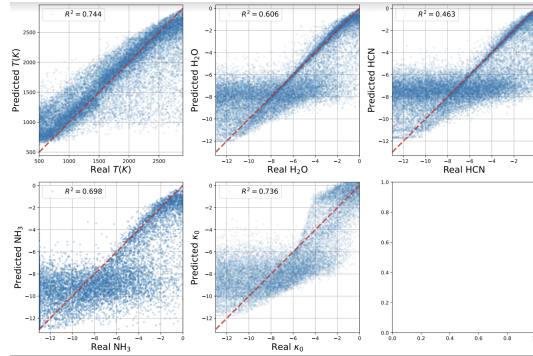


Figure 3: R^2 score after performing the testing.

We also ran the retrieval, in which 1000 regression trees were formed, each regression tree gives a set value range of the parameter. Here 5 parameters were used. The Parameters used include temperature, H2O, HCN, NH3, and specific gravity. The parameter space radius of the exoplanet can be found by using the radiative transfer equation. In Helia, the closest transit generated is considered for calculation and the respective parameter value ranges are the output of the final transmission spectrum.

The output is generated in the form of corner plots as shown below:

Each corner plot gives the range value of the parameters used in this case the given 5.

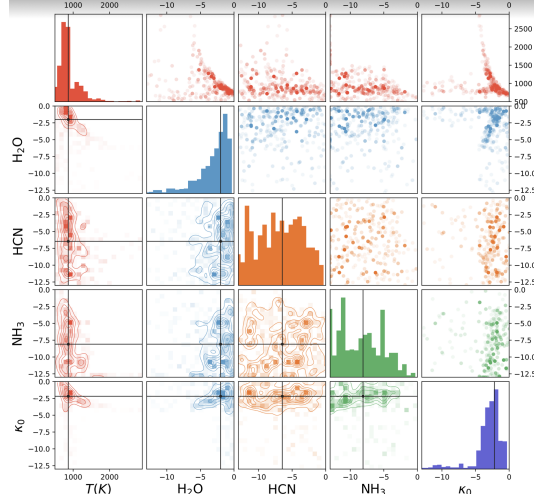


Figure 4: Corner plot of retrieval run in HELA.

3.2 POSEIDON:

POSEIDON is the non-ML Atmospheric Retrieval method that uses the Nested sampling method in Bayesian inference.

In Nested sampling, in the given dataset range some random live points are taken the likelihood of each point is found and the point with the least likelihood is discarded and the contour forms give the value range of the parameters thus the forward models are generated, iterations are performed to minimise the error and thus a final fitting gives the desired transmission spectrum with uncertainties.

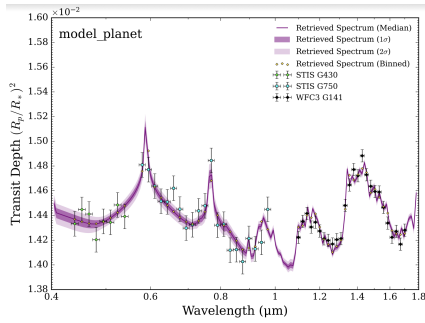


Figure 5: Snapshot of the forward model

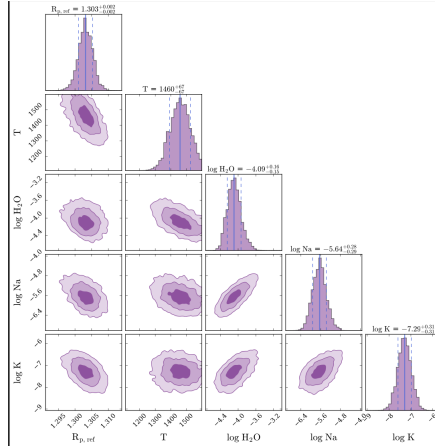


Figure 6: Corner plot of the POSEIDON run retrieval.

4 Further plans

- Data Set would be taken from HELA DATA SET (80,000 2. WFC3 transmission spectra for training and 20,000 datasets for testing.).
- Implementation of Neural networks and Ensemble neural networks with the data set.

5 References:

1. Robinson, T. D. (2017, February 20). A Theory of Exoplanet Transits with Light Scattering. *The Astrophysical Journal*, 836(2), 236.
2. Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K. (2018, June 25). Supervised machine learning for analyzing spectra of exoplanetary atmospheres. *Nature Astronomy*, 2(9), 719–724.
3. MacDonald, R. J. (2023, January 13). POSEIDON: A Multidimensional Atmospheric Retrieval Code for Exoplanet Spectra. *Journal of Open Source Software*, 8(81), 4873.
4. Hayes, J. J. C., Kerins, E., Awiphan, S., McDonald, I., Morgan, J. S., Chuanraksasat, P., Komonjinda, S., Sanguansak, N., Kittara, P. (2020, April 14). Optimizing exoplanet atmosphere retrieval using unsupervised machine-learning classification. *Monthly Notices of the Royal Astronomical Society*, 494(3), 4492–4508. <https://doi.org/10.1093/mnras/staa978>
5. All the codes used for training and testing (in HELA) can be found here.