First dedicated center for Astrophysical AI

ASTROAI

Enabling Next Generation Astrophysics

CENTER FOR ASTROPHYSICS | HARVARD & SMITHSONIAN

AstroAI @ CfA wins the ARIEL data challenge on Exoplanets Atmospheres Retrieval

Mayeul AUBIN m.mayeul.aubin@gmail.com, article link: https://arxiv.org/abs/2309.09337

I. Context of the ARIEL data challenge

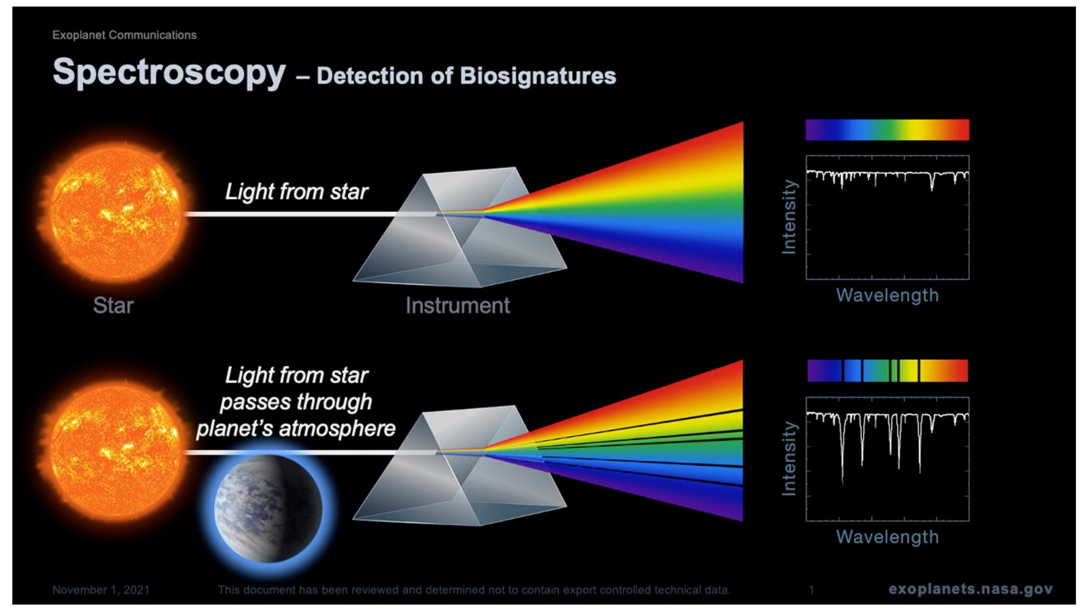

Fig 1: Principles of exoplanets atmopsheres spectroscopy

When an exoplanet transits in front of its host star, a fraction of the light passes through the exoplanet atmosphere. By analysing its spectrum, we can infer the atmospheric composition. Telescopes such as the JWST and the ARIEL will provide spectra to analyse.

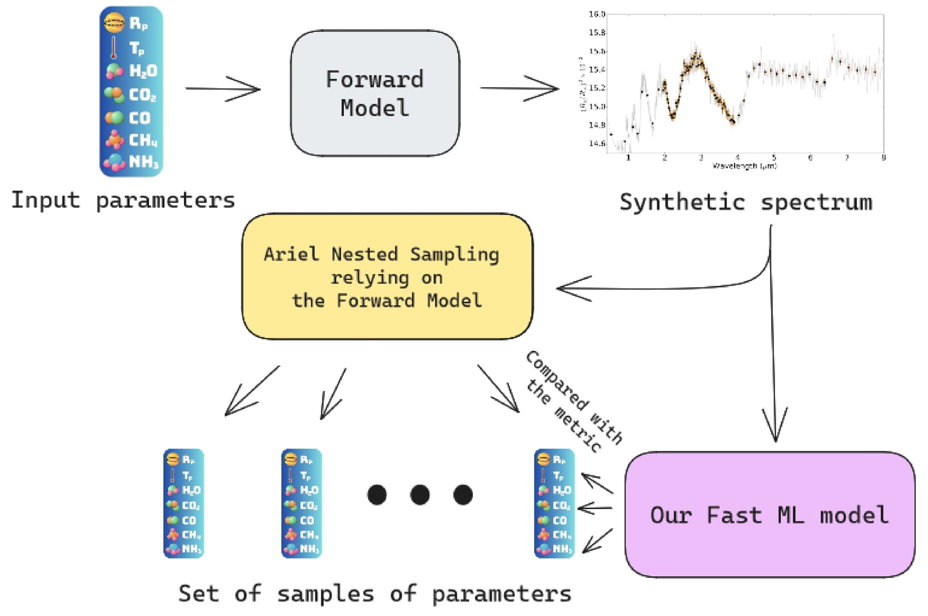

Fig 2: Dataset generation and definition of the task

Using the radiative transfer software TauREx3, the Ariel team generated synthetic spectra from input parameters. Then they ran a Nested Sampling to produce a set of samples compatible with the spectrum. The task was to replace the Nested Sampling by a fast machine learning model. They varied the atmospheric models across the training and testing sets.

II. Our solution: Normalising Flows with domain knowledge

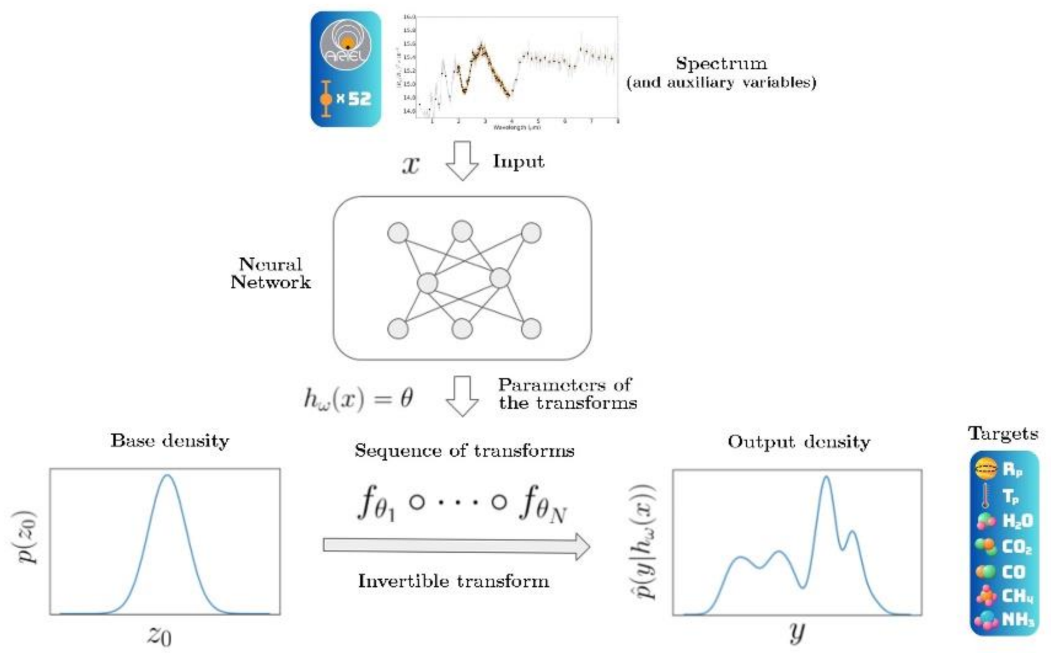

Fig 3: Concept of a Normalising Flow (source siboehm.com)

A Normalising Flow outputs a posterior probability distribution by transforming a base density (a gaussian) with a sequence of parametrised transformations, whose parameters are learned by a neural network. It can then be used to easily produce samples, or evaluate the probability density function.

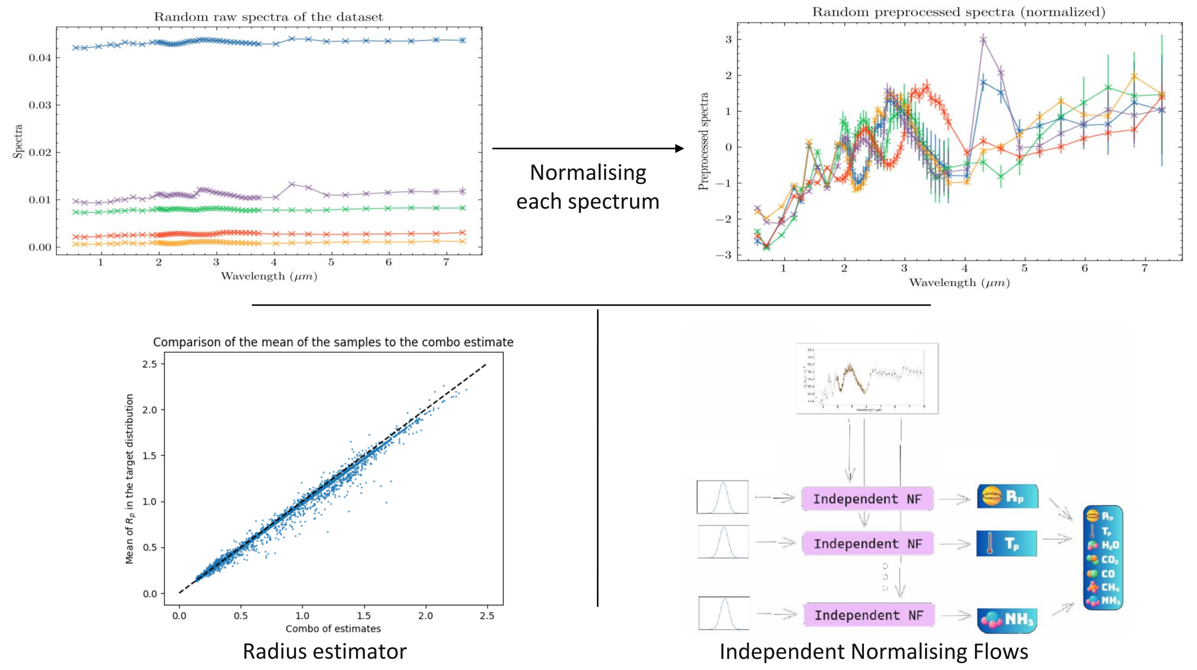

Fig 4: The three main tricks used to win the challenge

Expert knowledge for maximising the challenge score:

- Preprocess the spectra to highlight the features
- Compute radius estimators to help the model learn
- Use independent normalising flows (one for each parameter) because 80% of the metric is only on the marginals

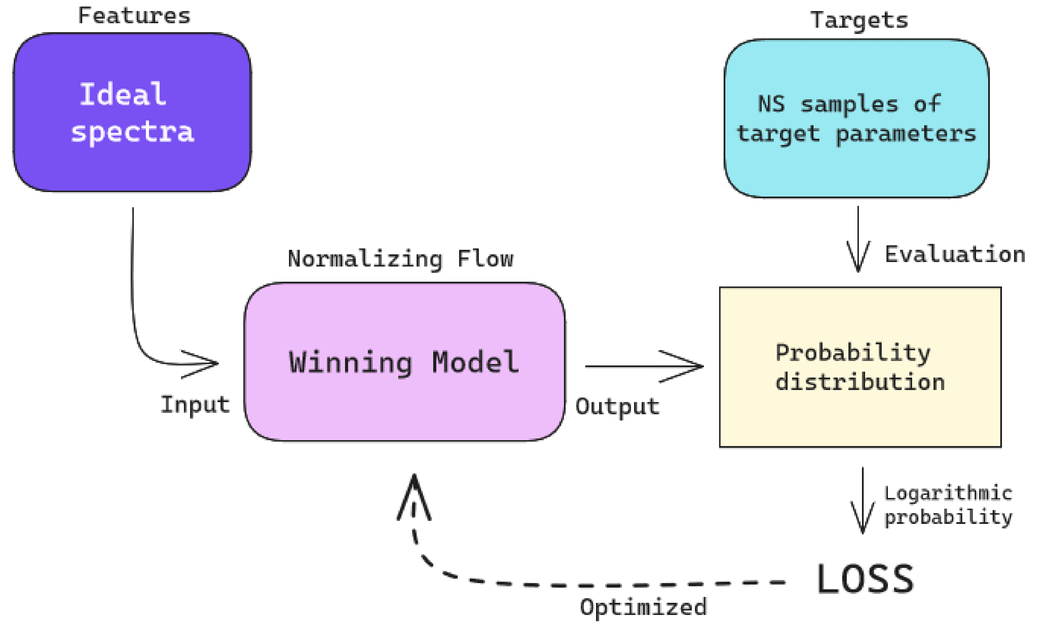

Fig 5: Training scheme of the winning model

The winning model was trained with ideal spectra (no noised added) as input and Nested Sampling samples as targets to match the challenge setup. While our alternative model was trained on noised spectra (more realistic) as input and the original parameters as targets (to avoid any approximation of the Nested Sampling).

III. Results: Winning and alternative models

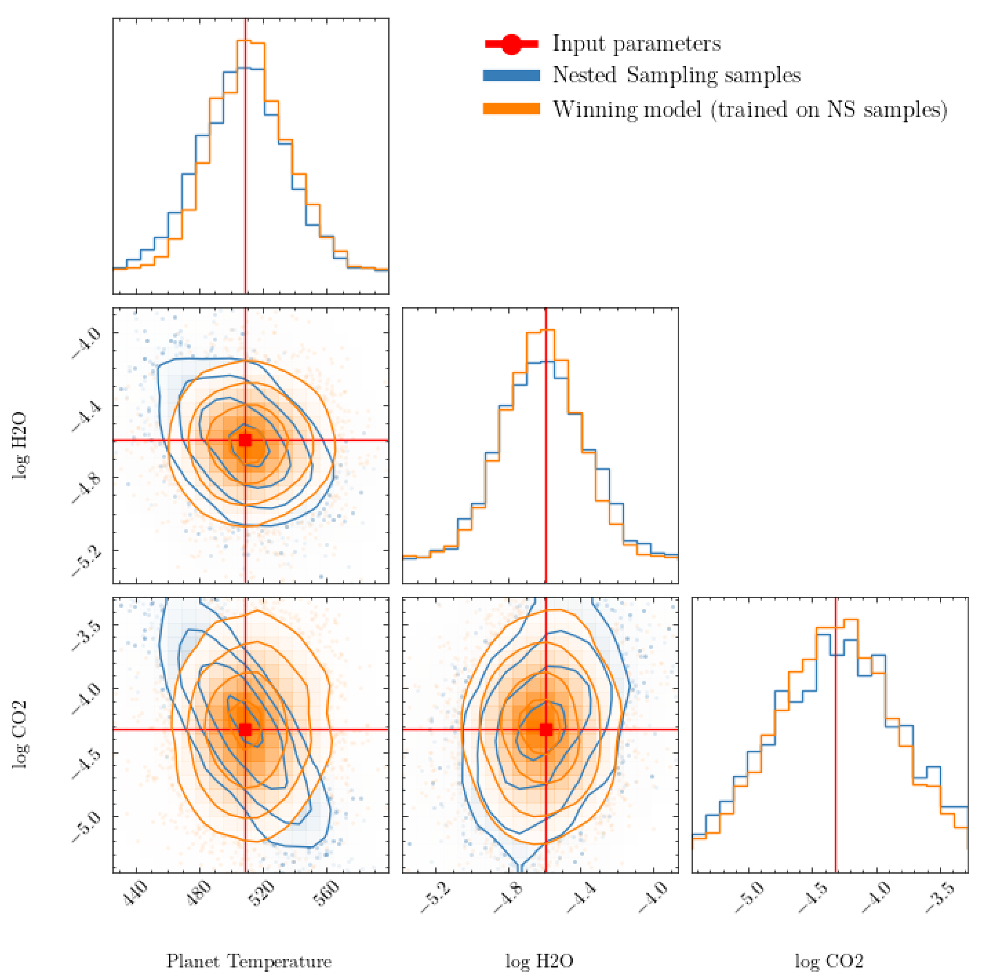

Fig 6: Extract of posteriors of the Nested Sampling target (blue) and our winning model (orange) on a validation ideal spectrum, with the true original parameters (red)

Metrics	Winning model	Alternative model
Logprob (ideal spectra)	3.16	2.86
Logprob (noised spectra)	1.03	2.48
Challenge score	688.13	577.32

Fig 7: Table of scores and average logarithmic probability on validation set

We won the challenge, ranking **1st out of 293 teams**. Our winning model is excellent at **imitating the behaviour** of the Nested Sampling on the marginals for ideal spectra. However, it was **missing any correlation** (joint probability), as shown on the T-CO_2 plot (lower-left corner).

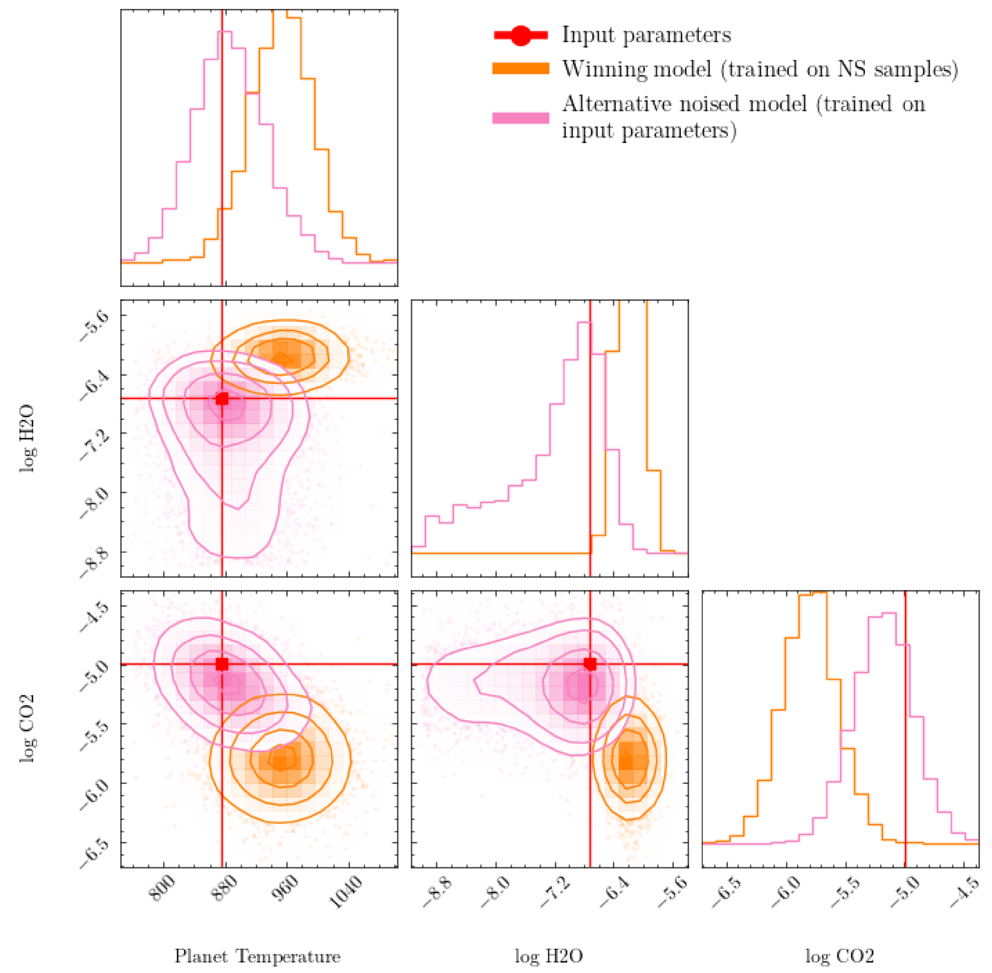

Fig 8: Extract of posteriors of our winning model (orange) and alternative model (pink) on a random validation noised spectrum, for the parameters T, H_2O, and CO_2

We also proposed an **alternative model**, less performant at the challenge score, but that seems to be **more precise** at constraining the original input parameters on **noised spectra,** reflecting more what's expected for an inference tool. Therefore, we recommend to **evaluate the models on noised spectra**, with a noise model as close as possible from what we will observe (include all forms of uncertainties and noise).

Thanks for reading !