PERTURB: Machine Learning Interpretability in High Correlation Environments for Exoplanet Atmospheric Retrieval

Jools D. Clarke[1], Gordon Yip[2], Nikolaos Nikolaou[1]

[1]University College London, UK
[2]King's College London, UK

SEE MORE

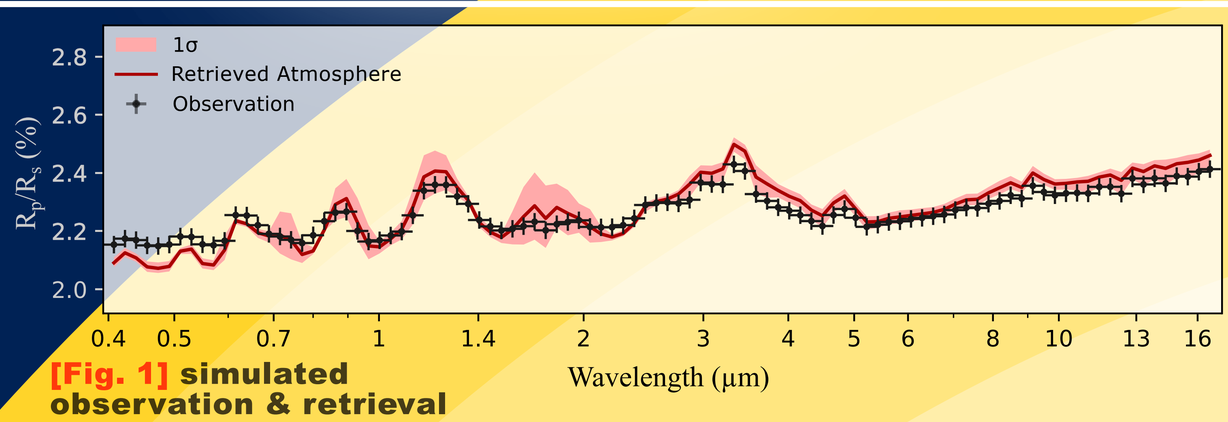
TRANSMISSION SPECTROSCOPY:

It is **crucial to study small exoplanets** to expand our knowledge of these distant worlds and our place in the cosmos, and to **drive innovation** in atmospheric modelling that improves our understanding of **Earth's climate**.

Observing super-Earth/sub-Neptune objects is challenging due to their **proximity** to, and **magnitude disparity** with, their host star.

For exoplanets that transit, they can be observed in a **transit observation**:
1) Exoplanet's orbit takes it between Earth and its host star (**transit**)
2) Light from the star passes through the atmosphere of the planet
3) Spectral signature of the planetary atmosphere can be isolated [Fig. 1]

Inferring the chemical composition of the atmosphere from this transmission spectrum (**retrieval**) is a:
- Computationally intensive process
- Relies on iterative generation of complex atmospheric models
- Can take hours to days depending on the model



[Fig. 1] simulated observation & retrieval

MACHINE LEARNING & SPECTRAL RETRIEVALS:

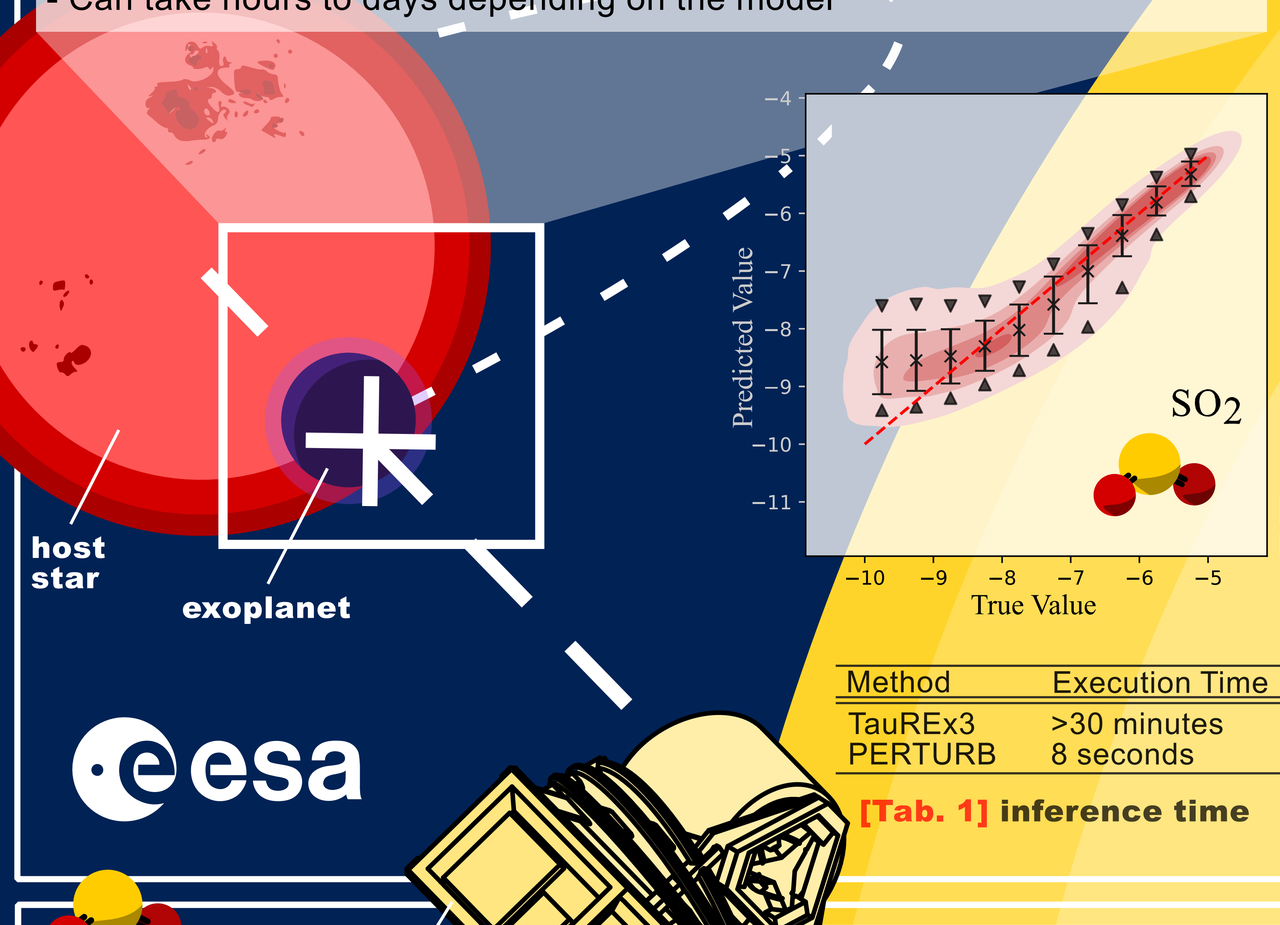
The transition toward machine learning (ML) retrievals to **minimise computational demand** is both **logical**, but also **inevitable** given the rate of increase in data acquisition with upcoming dedicated transit spectroscopy missions.

The computational expense is offset into an **amortised**† training expense, so the trained pipeline can be comparatively lightweight [Tab. 1].

Raw predictive accuracy of ML retrievals is comparable to that of Bayesian retrievals when tested on simulated spectra [Fig. 1-2], but they provide **less informative posteriors** than Monte Carlo methods.

This can be mitigated by **sampling** from priors, **uncertainty characterisation** of model performance in a test dataset, and **prediction limit characterisation** to infer reasonable posteriors [Fig. 3].

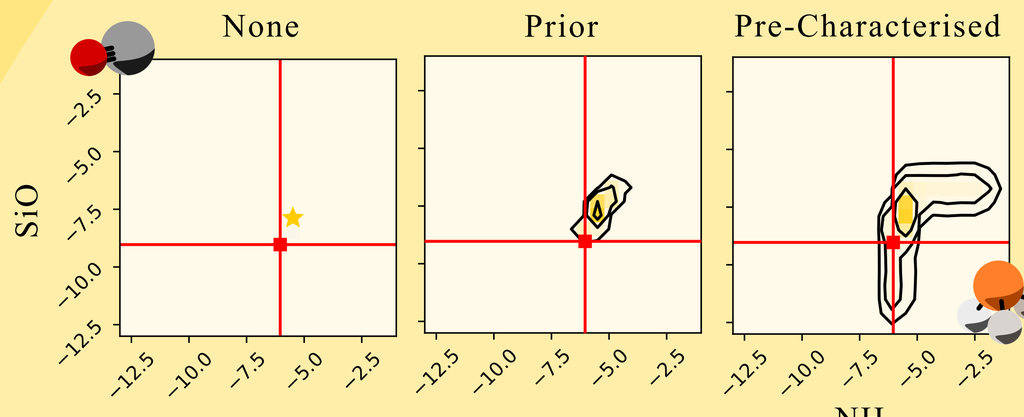
†*amortisation: front-loading an ongoing operating cost into a high one-time initial cost*



[Fig. 2] model performance

Method	Execution Time	Speed Up
TauREx3	>30 minutes	x 1
PERTURB	8 seconds	x 225

[Tab. 1] inference time



[Fig. 3] uncertainty characterisation

PERTURB:

Random sampling methods generate unphysical samples for correlated data.

Techniques designed for interpreting uncorrelated data **perform poorly**:
- Ceteris paribus
- SHAP

Perturb generates samples that are:
- **Physically motivated**
- Within **expected distribution** [Fig. 4]

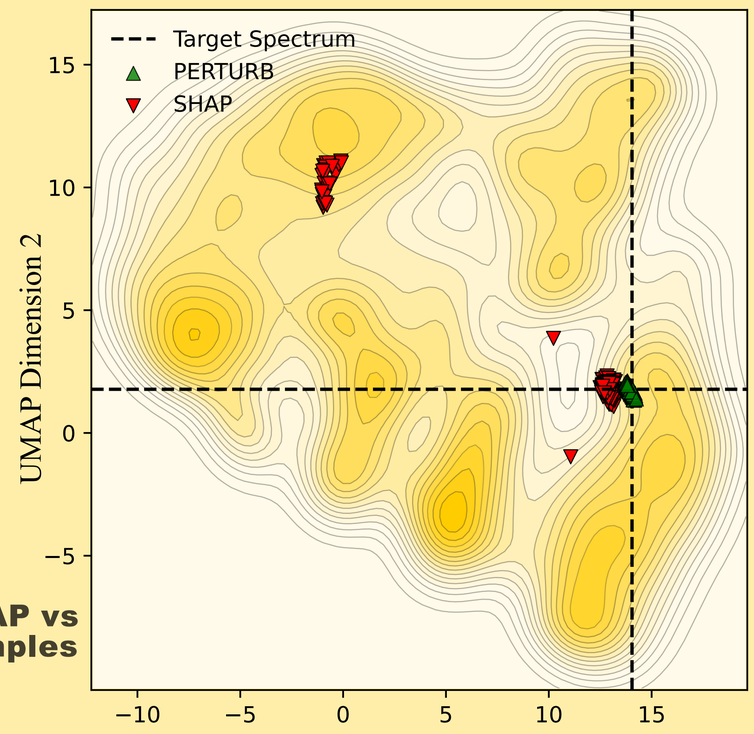
and provides a visualisation of:
- **Response magnitude**
- **Non-linearity** [Fig. 5-6]

VISUALISING FEATURE RESPONSE:

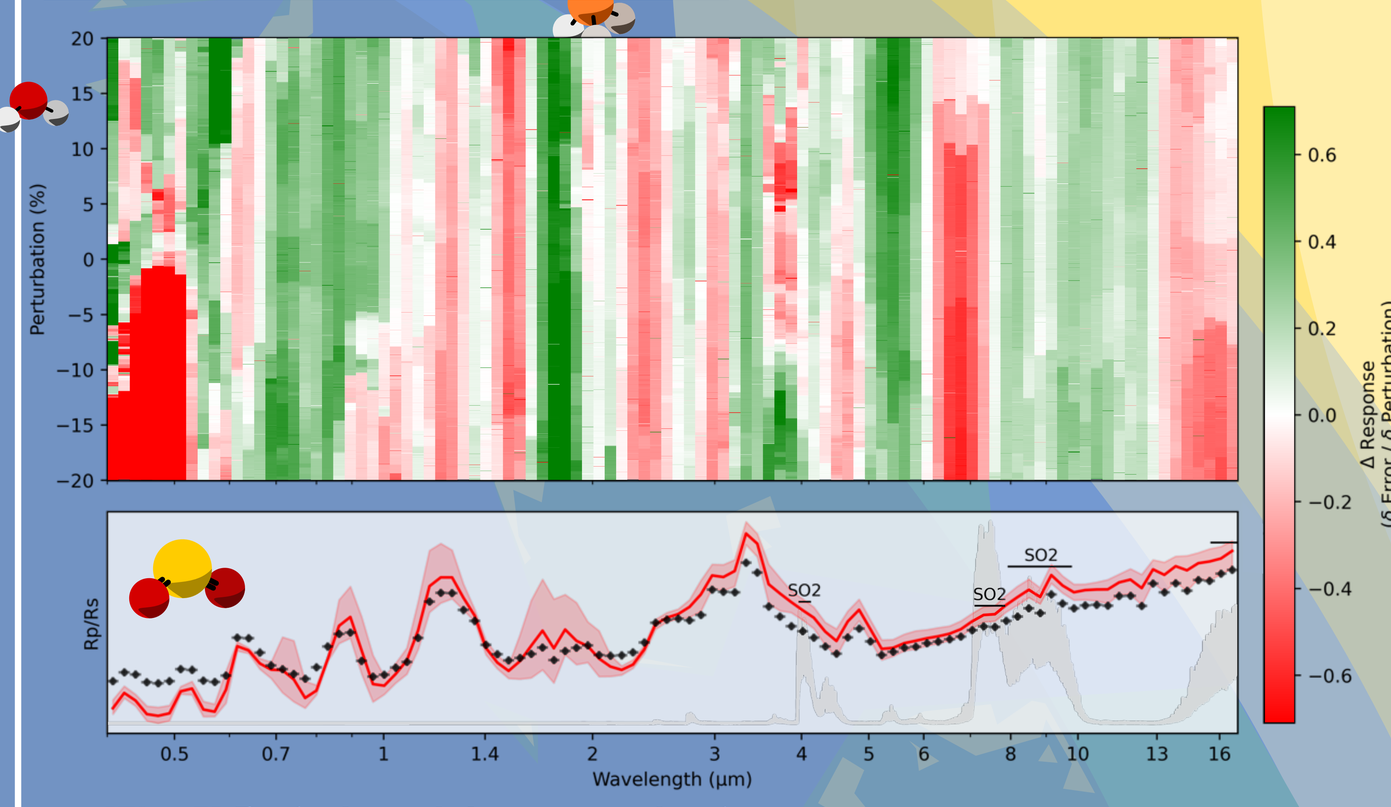
ML comes with the **black-box problem**:
- It is hard to know the relative effect of each observed datapoint on the retrieved result.

We need to better understand the predictions made by atmospheric retrieval models to:
- Enable more **widespread adoption**
- Ensure ML can be a **reliable tool** for astrophysicists

$$\text{base prediction: } Y = f(X)$$
$$\text{gaussian window: } w_{\text{gauss}}(x_i, x_j) = \exp\left(-\frac{4\ln 2\,(x_i - x_j)^2}{\text{FWHM}^2}\right)$$
$$\text{stochasticity modifier: } \boldsymbol{\epsilon} \sim \mathcal{N}\left(x_i^j, \sigma_i\right)$$
$$\text{augmented data: } \bar{X} = X + \boldsymbol{\epsilon} \cdot w_{\text{gauss.}}(x_i, x_j)$$
$$\text{augmented predictions: } \bar{Y} = f\left(\bar{X}\right)$$
$$\text{perturbation response: } R = Y - \bar{Y}$$

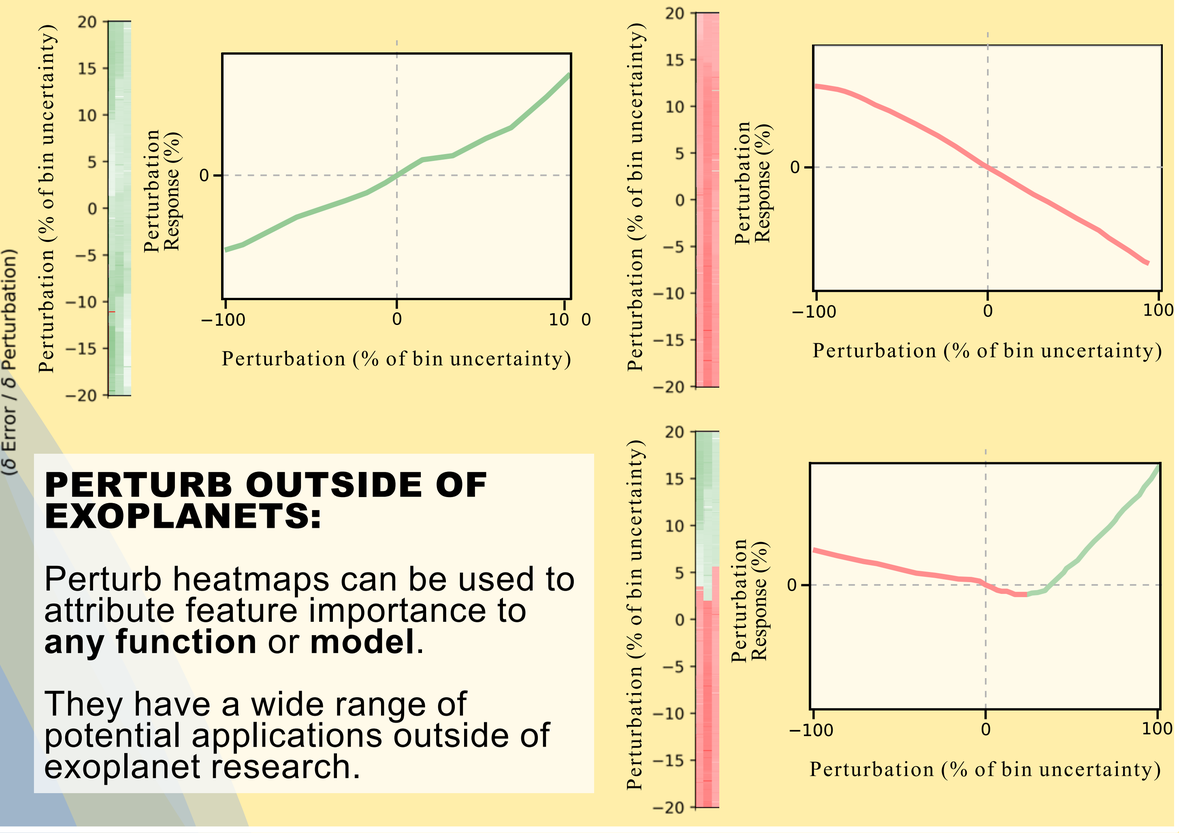


[Fig. 4] SHAP vs Perturb samples



[Fig. 5] Perturb model response heatmap

[Fig. 6] how to read Perturb heatmaps



PERTURB OUTSIDE OF EXOPLANETS:

Perturb heatmaps can be used to attribute feature importance to **any function** or **model**.

They have a wide range of potential applications outside of exoplanet research.