RAS Techniques and Instruments

RASTAI **2,** 45–61 (2023) https://doi.org/10.1093/rasti/rzad001
Advance Access publication 2023 January 23

ESA-Ariel Data Challenge NeurIPS 2022: introduction to exo-atmospheric studies and presentation of the Atmospheric Big Challenge (ABC) Database

Quentin Changeat[1,2★†] and Kai Hou Yip[2]

[1] *European Space Agency (ESA), ESA Office, Space Telescope Science Institute (STScI), 3700 San Martin Drive, Baltimore, MD 21218, USA*
[2] *Department of Physics and Astronomy, Gower St., London WC1E 6BT, UK*

Accepted 2023 January 10. Received 2023 January 5; in original form 2022 June 29

ABSTRACT

This is an exciting era for exo-planetary exploration. The recently launched JWST, and other upcoming space missions such as Ariel, Twinkle, and ELTs are set to bring fresh insights to the convoluted processes of planetary formation and evolution and its connections to atmospheric compositions. However, with new opportunities come new challenges. The field of exoplanet atmospheres is already struggling with the incoming volume and quality of data, and machine learning (ML) techniques lands itself as a promising alternative. Developing techniques of this kind is an inter-disciplinary task, one that requires domain knowledge of the field, access to relevant tools and expert insights on the capability and limitations of current ML models. These stringent requirements have so far limited the developments of ML in the field to a few isolated initiatives. In this paper, We present the Atmospheric Big Challenge Database (ABC Database), a carefully designed, organized, and publicly available data base dedicated to the study of the inverse problem in the context of exoplanetary studies. We have generated 105 887 forward models and 26 109 complementary posterior distributions generated with Nested Sampling algorithm. Alongside with the data base, this paper provides a jargon-free introduction to non-field experts interested to dive into the intricacy of atmospheric studies. This data base forms the basis for a multitude of research directions, including, but not limited to, developing rapid inference techniques, benchmarking model performance, and mitigating data drifts. A successful application of this data base is demonstrated in the NeurIPS Ariel ML Data Challenge 2022.

Key words: planets and satellites: atmospheres – telescopes – methods: data analysis.

1 CONTEXT

The field of exoplanet has come a long way since the discovery of the first exoplanet in 1994 (Wolszczan & Frail 1992). With the launch of dedicated telescopes for the detection of exoplanets, such as the Convection, Rotation et Transits planétaires (*CoRoT*; Pätzold et al. 2012), the *Kepler* (Borucki et al. 2010), and the *Transiting Exoplanet Survey Satellite* (*TESS*; Ricker et al. 2015) space telescopes, we now have basic characteristics, such as planetary radii or masses, for more than 5000 alien worlds. From the observed population, we deduced that, while exoplanets are ubiquitous (Cassan et al. 2012; Batalha 2014), the architecture of our Solar system does not appear to be a typical outcome of planetary formation. For instance, the first detected exoplanet around a sun-like star is classified as a hot Jupiter (Mayor & Queloz 1995), a planet of a similar size to Jupiter (e.g. about 10 times the size of Earth) but orbiting so close to its host star that it completes a full revolution in about 4 d. Such planet does not exist in our Solar system and so are the majority of the observed planets, referred as sub-Neptunes due to their size being between the size of Earth and Neptune (Howard et al. 2010; Fulton et al. 2017; Petigura et al. 2022). To answer the most fundamental questions of the field, such as 'what are exoplanets made of?' or 'how do planets form?', one must obtain complementary information to planetary masses and radii.

In the last decade, astronomers have therefore turned their attention to exoplanetary atmospheres, or exo-atmospheres, in the quest for further constraints on these worlds (Charbonneau et al. 2002; Tinetti et al. 2007; Swain, Vasisht & Tinetti 2008; Kreidberg et al. 2014; Schwarz et al. 2015; Sing et al. 2016; Stevenson et al. 2017; de Wit et al. 2018; Hoeijmakers et al. 2018; Tsiaras et al. 2018, 2019; Brogi & Line 2019; Welbanks et al. 2019; Edwards et al. 2020; Changeat & Edwards 2021; Roudier et al. 2021; Yip et al. 2021a; Changeat et al. 2022; Chen et al. 2022; Edwards et al. 2022; Estrela, Swain & Roudier 2022; Mikal-Evans et al. 2022). The study of exoplanet atmospheres has been enabled by the use of space-based instrumentation, such as the *Hubble Space Telescope* (*HST*), the retired *Spitzer Space Telescope*, and ground-based facilities such as the Very Large Telescope (VLT). Many discoveries were made. We, for instance, know that water vapour is present in many hot Jupiter atmospheres, and we have recently recovered evidence for links between atmospheric chemistry and formation pathways. However, with the recent launch of the NASA/ESA/CSA *James Webb Space Telescope* (*JWST*; Greene et al. 2016), and the upcoming ESA Ariel Mission (Tinetti et al. 2021) and BSSL Twinkle Mission (Edwards

★ E-mail: quentin.changeat.18@ucl.ac.uk
† ESA Research Fellow.

et al. 2019b), the field of exoplanetary atmosphere will undergo a revolution. The quality and quantity of atmospheric data will be multiplied exponentially, bearing many new challenges.

One of the main challenge in the study of exo-atmospheres, even today, concerns with the reliable extraction of information content from observed data. Atmospheres are complex dynamical systems, involving many physical processes (chemical and cloud reactions, energy transport, fluid dynamics) that are coupled, poorly understood, and difficult to reproduce on Earth. Astronomers have therefore attempted to interpret observations of atmospheres using retrieval techniques: simplified models (or reduced order models) for which the parameter space of possible solutions is explored using a statistical framework (Irwin et al. 2008; Madhusudhan & Seager 2009; Line et al. 2012, 2013; Waldmann et al. 2015a,b; Lavie et al. 2017; Gandhi & Madhusudhan 2018; Mollière et al. 2019; Zhang et al. 2019; Min et al. 2020; Al-Refaie et al. 2021; Harrington et al. 2022). With current observational data, state-of-the-art retrieval models use sampling based Bayesian techniques, such as MCMC or Nested Sampling, with non-informative (uniform) priors to obtain the posterior distributions of between 10 and 30 free parameters (Changeat et al. 2021a). The number of free parameters depends on the information content available in the observational data and the chosen atmospheric model. As of today, there is no consensus on the most appropriate atmospheric model to employ, and we cannot obtain in-situ observations (e.g. we cannot travel there). Sampling-based techniques typically require between 10^5 and 10^8 forward model calls to reach convergence, meaning that only models providing spectra of the order of seconds are viable. With the increase in data quality, thanks to *JWST*, Ariel, and Twinkle, it will enable a wider range of atmospheric processes to be probed by the observations, implying that forward models must grow in complexity and so does the dimensionality of the problem (The JWST Transiting Exoplanet Community Early Release Science Team et al. 2022). As such, interpreting next-generation telescope data is currently a real issue, which has been highlighted multiple times by studies relying on simulations, that will require a revolution in both our models and information extraction techniques (Rocchetto et al. 2016; Caldas et al. 2019; Changeat et al. 2019; Taylor et al. 2020, 2021; Yip et al. 2020; Changeat et al. 2021a; Al-Refaie et al. 2022a; Yip et al. 2022a).

In recent years, the community started to explore alternative approaches to circumvent the bottleneck with sampling based approaches. Machine learning (ML) models land itself as a promising candidate with its high flexibility and rapid inference time. Waldmann (2016) pioneered the use of deep learning network in the context of atmospheric retrieval, training a Deep Belief Network to identify molecules from simulated spectra. On the other end, Márquez-Neila et al. (2018) led the first attempt to train a Random Forest regressor to predict planetary parameters directly. Since then, the field has started to look at different ML methodologies to bypass the lengthy and computationally intensive retrieval process (Soboczenski et al. 2018; Zingales & Waldmann 2018; Cobb et al. 2019; Hayes et al. 2020; Nixon & Madhusudhan 2020; Oreshenko et al. 2020; Ardevol Martinez et al. 2022; Haldemann et al. 2022; Himes et al. 2022; Yip et al. 2022a). Pushed by astronomers' need for explainable solutions, other groups have also looked into the information content of exoplanetary spectra with AI (Guzmán-Mesa et al. 2020; Yip et al. 2021b).

The publicly available Atmospheric Big Challenge (ABC) Database of forward models and retrievals aims to provide the resources to address aforementioned issue via participation of external communities and encourage novel, cross-disciplinary solutions. It is constructed as a permanent data repository for further investigations. The data base is accessible at the following link: https://doi.org/10.5281/zenodo.6770103.

Since the creation of similar data base constitutes a major barrier to anyone interested in applying ML in the domain of exoplanet atmospheres, we emphasize on its release as a community asset. The organization and creation of this data set poses a challenge on its own because of the following:

(i) It requires a cross-disciplinary collaboration. The problem requires domain knowledge (atmospheric chemistry, radiative transfer, atmospheric retrievals) to ensure the data product represents a meaningful science case rather than a trivial example. At the same time, it requires ML expertise to ensure the data product is representative of the problem at hand, and ideally, one that adequately reflects the reality.

(ii) It requires access to the relevant tools which is often exclusive to communities in exoplanet: atmospheric retrieval and chemistry codes as well as instrument noise simulators.

(iii) It requires significant computing resources. For this project, more than 2000 000 CPUh were used. Simulations of this scale have never been attempted before.

This paper is written to (1) provide non-field experts with a light-weighted introduction to the science behind the data generation process, (2) document the steps involved in the creation of the ABC data base, and (3) to provide a carefully curated, well-organized, and scientifically relevant data set for any research community. This manuscript complements the data challenge proposal description (Yip et al. 2022b) accepted as a NeurIPS 2022 data challenge. It is intended to provide the required domain knowledge for non-field experts. We present a simplified jargon-free introduction to the most commonly employed techniques in the field of exo-atmospheres in Appendix A.

2 DATA GENERATION

For the data generation, we employed *Alfnoor* (Changeat et al. 2021a), a tool built to expand the forward model and atmospheric retrieval capabilities of *TauREx 3* (Al-Refaie et al. 2021) to large populations of exo-atmospheres. *Alfnoor* allows to automatize the generation or telescope simulations and perform large-scale standardized atmospheric retrievals. A lightweight description of the main concepts behind atmospheric studies of exoplanets are described in Appendix A. In the context of ESA-Ariel, we generated 105 887 simulated forward observations as well as 26 109 standardized retrieval outputs.

2.1 Source of input parameters

To model those extrasolar systems, some preliminary assumptions were required. In particular, all the parameters that are not linked to the atmospheric chemistry needed to be fixed to realistic values. Those parameters include, but are not limited to, stellar radius (R_{s}), distance to Earth (d), star magnitude K (K_{mag}), planetary radius (R_{p}), planetary mass (M_{p}), planet equilibrium temperature (T), and transit duration (t_{14}).

The planetary objects in this data base were selected from the list of confirmed known exoplanets and the list of *TESS* exoplanet candidates (TOIs). This list was constructed as part of the ESA-Ariel Target list initiative (Edwards et al. 2019a; Edwards & Tinetti 2022), frozen to 2022 March 1 for this data base. For the TOIs, we are aware that some of those objects will not be exoplanets; however, the observation of their transit by *TESS* and the first preliminary checks

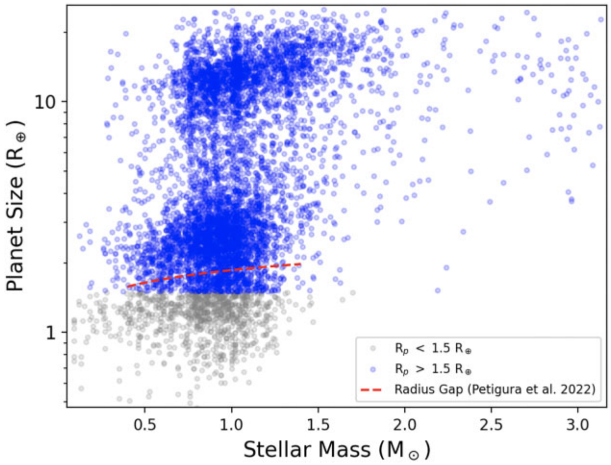

Figure 1 Size of the considered planets versus the mass of their host star. We exclude planets that have radii below 1.5 $R_\oplus$ marked in grey and approximately corresponding to the lower limit on the Radius Valley.

of their inferred properties make them compelling objects. Follow-up observations will allow us to classify their nature, but for the purpose of building this data base, they are as close as possible to what the reality looks like. As radial velocity follow-ups cannot and is not systematically conducted for all targets, the mass of some of those objects is unknown. In this case, as in Edwards & Tinetti (2022), we replace the planetary mass by an estimate from the relation described in Chen & Kipping (2017). To those lists of objects, we filtered all the planets with radius below 1.5 $R_\oplus$, the conservative value for the middle of the Radius Valley (Fulton et al. 2017; Cloutier & Menou 2020; Petigura et al. 2022). This is because the atmospheric composition of small planets would require a much more complex treatment (e.g. the assumption of hydrogen-dominated atmosphere is not theoretically sound) than is proposed here. In total, we obtained data for 2972 confirmed exoplanets and 2928 candidate exoplanets, thus bringing our total to 5900 unique objects. The list of selected planets for this database is shown in Fig. 1.

Fig. 2 shows the distributions of nine selected stellar and planetary parameters. These values are taken from the actual planetary system and therefore follows the current observed demographics, these values remains unchanged thorough the data generation process. However, relying on currently known planets is a double-edged knife. While it saved us from making unverified assumptions, our data are prone to selection bias stemmed from the observation technique, strategy, and instrument specification. These biases can be easily spot from Fig. 2. For instance, the distribution of orbital period tends to be shorter (peaks around ~3 d) as their proximity to the host star makes them easier to discover.

2.2 The atmospheric forward model setup

We produce batches of randomized observations for the population described in the previous section. For each planet, the stellar (R_s, d, K_{mag}), orbital (t_{14}), and bulk parameters (R_p, M_p, T) are fixed to their literature values, while the chemistry of the atmosphere is randomly generated. The thermal profile is assumed to be isothermal (constant temperature) at the equilibrium temperature of the planet, and we simulate the planet's atmosphere from 10 to 10^{-10} bar using 100 layers (divided uniformly in log-pressure space).

For the chemistry, we assume a primary atmosphere made mainly from hydrogen and helium (He/H_2 = 0.17), to which we add trace gases. The trace gases are H_2O (Polyansky et al. 2018), CH_4 (Yurchenko et al. 2017; Chubb et al. 2021), CO (Li et al. 2015), CO_2 (Yurchenko et al. 2020), and NH_3 (Coles, Yurchenko & Tennyson 2019), selected based on our current understanding of exoplanetary chemistry (Agúndez et al. 2012; Venot & Agúndez 2015; Drummond et al. 2016; Madhusudhan et al. 2016; Stock et al. 2018; Woitke et al. 2018; Venot et al. 2020; Baeyens et al. 2022; Al-Refaie et al. 2022a). The mixing ratio, or trace abundance, of those gases is randomly chosen using a Log Uniform law and depends on the molecule considered. The Log Uniform law is chosen rather than a more informative law (such as equilibrium chemistry) because we are looking for solutions that are unbiased to our current, most likely limited, understanding of atmospheric chemistry. Such training set is suitable to produce ML solutions behaving in a similar way to the so-called free chemistry retrievals. If correlation exists in a real population (e.g. between the chemistry of the atmosphere and its thermal structure), such method should allow the extraction of this trend without the need to implicitly make a physical assumption. Note that this is required in the cases where data have undergone a data shift (in this case, when the data are generated using a different atmospheric assumption). Another important point to consider involves the detection capabilities of Ariel for each molecule and the degeneracy between molecular species. For instance, CO shares similar features to CO_2 in Ariel but it is a much harder molecule to detect due to its weaker absorption properties. Due to those differences in the strength of spectral features and guided by the Ariel Tier-2 detection limits investigated in Changeat et al. (2020a), we select different bounds for the randomized chemical abundances. This process allows us to balance our data set and ensure that a significant fraction of the planets have detectable amount of CO. The bounds employed for this data set are as follows:

(i) H2O: RandomLogUniform(min = −9, max = −3).
(ii) CO: RandomLogUniform(min = −6, max = −3).
(iii) CO2: RandomLogUniform(min = −9, max = −4).
(iv) CH4: RandomLogUniform(min = −9, max = −3).
(v) NH3: RandomLogUniform(min = −9, max = −4).

For each parametrized atmosphere, we compute the radiative transfer (see Appendix A) layer by layer, including the contributions from molecular absorption, Collision Induced Absorption (CIA), and Rayleigh Scattering.

Each spectrum is first computed at a high resolution,[1] before being convolved with an Ariel instrument simulation. For each planet, we employed the *TauREx* plugin for ArielRad (Mugnai et al. 2020), the official Ariel noise simulator, to estimate the noise on observation at each wavelength. With ArielRad, we force each observation to satisfy the criteria for Ariel Tier 2 observations (Tinetti et al. 2021), meaning that the observations have a specific resolution profile (e.g: $R \approx 10$ for $1.10 < \lambda < 1.95$ μm; $R \approx 50$ for $1.95 < \lambda < 3.90$ μm; $R \approx 15$ for $3.90 < \lambda < 7.80$ μm) and that the signal-to-noise ratio (SNR) of the observations must be higher than 7 on average. The SNR is here defined on the atmospheric signal (e.g. the second part of equation A2). To produce the simulated spectra, we select the minimum number of transit that allow to reach this threshold, meaning that our sample of observations contains a wide range of final SNR. Since we used real objects for those simulations and that all planets are not favourable targets for Ariel, this means that some targets require

[1] Spectra have to be computed at high-resolution ($\mathcal{R}$) since instrumental binning is done on the received photons, e.g. the recorded transit depth Δ. In our case, we used $\mathcal{R} = \lambda/\Delta\lambda = 10\,000$.

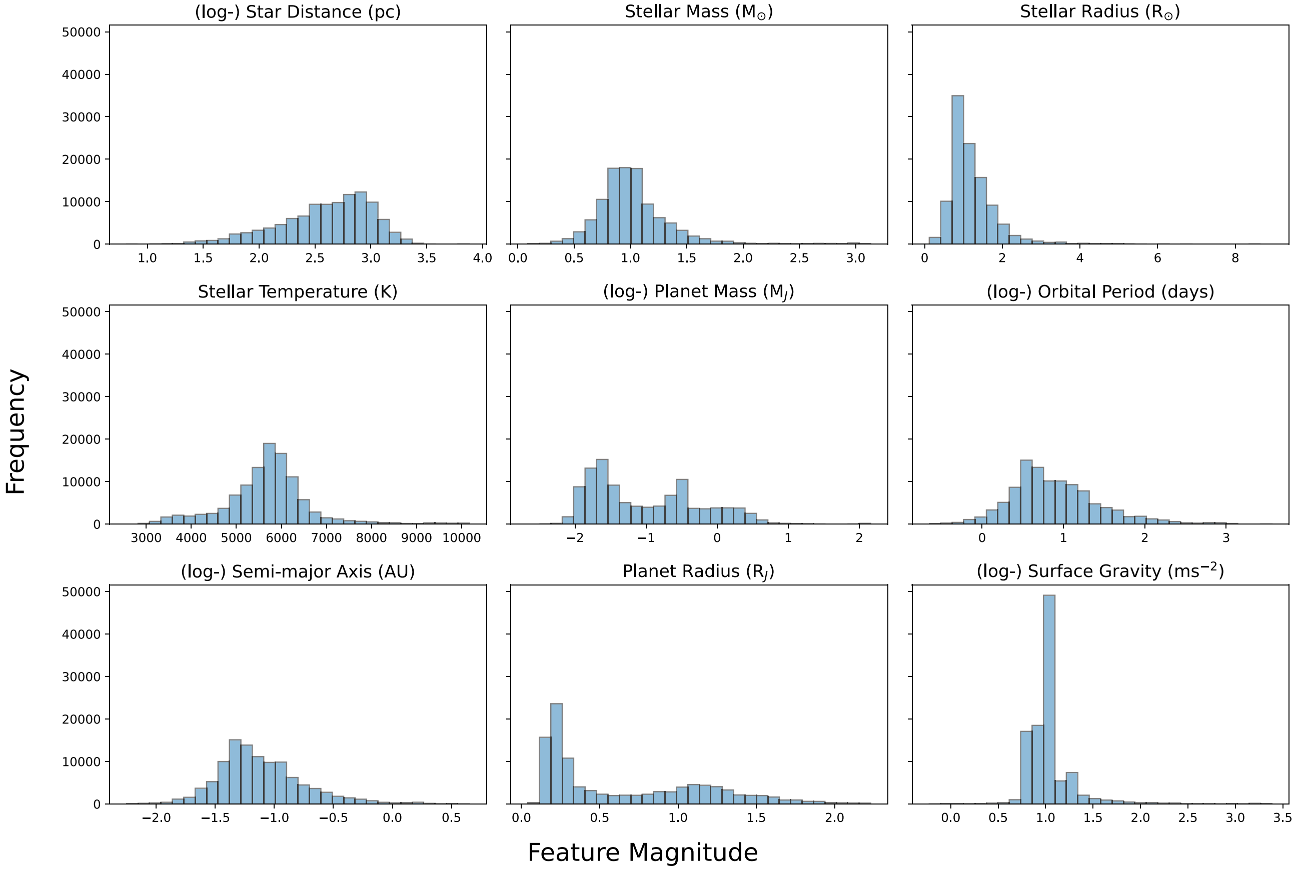

Figure 2 Distribution of nine stellar and planetary parameters used to generate the synthetic spectra. These distributions follow closely to the actual demographic of the currently known population of exoplanets, and therefore they are also subject to biases presented in the original population.

an un-realistic number of observations to reach the SNR condition of Tier 2. However, this does not affect the purpose of this data set, providing independent instances of realistic noise profiles.

Following those steps, we obtain a realistic Ariel simulated observation for each planet and each randomized chemistry. We show an example of such simulated observation in Fig. 3. In total, we produced 105 887 simulated observations for the ABC Database.

2.3 The atmospheric retrieval setup

For 26 109 (25 per cent) of the simulated observations generated at the previous step, we perform the traditional inversion technique using *Alfnoor*.

For the model to optimize, we kept the same setup as presented in the previous section and performed parameter search on the following free parameters: isothermal temperature (T), log abundances for H_2O, CO_2, CH_4, CO, and NH_3. The priors are made wide and uninformative, with the atmospheric temperature being fitted between 100 and 5500 K and the chemical abundances between 10^{-12} and 10^{-1} in Volume Mixing Ratios. The widely used Nested Sampling Optimizer, *MultiNest* (Feroz, Hobson & Bridges 2009), was employed with 200 live points and an evidence tolerance of 0.5.

For a single example on Ariel data, we provide the best-fitting spectrum in Fig. 3. From the optimization process, we are able to extract the traces of each parameters and the weights of the corresponding models. This allows to construct the posterior distribution of the free parameters with, for instance, *corner*. The posterior distribution of the same example is shown in Appendix B, Fig. C1. Processing of the posterior distribution also allows to extract statistical indicators describing the chemical properties of the planet, such as mean, median, and quantiles for each of the investigated parameters.

2.4 Data overview

Following the data generation process outlined above, we have generated a total of 105 887 forward models in Ariel Tier-2 resolution. Of them, 26 per cent of them are complemented with results from atmospheric retrieval (following a generic setting as described in Section 2.3).

Fig. 4 shows the distribution of mean transit depth (red) overlapped with the distribution of feature height (orange). The former served as a proxy of the diverse planetary classes present in the data set. The characteristic dichotomy stemmed from current demographics studies[2] and selection bias in our observation technique.[3] The latter is calculated from the difference between the maximum and minimum transit depth of each spectrum, it served as a proxy of the 'strength' of the spectroscopic features presented in the spectra, e.g. the peaks and troughs as seen in Figs A2 and 3. We note that an SED with linear slope will also produce a non-negligible feature height value, which is still considered as spectroscopic feature in our case. The two quantities are closely linked to our targets of interest, which means that any successfully model not only need to account for the inter-variation

[2] Latest studies show that Super-Earth sized planets are prevalent while there is a deficiency in the population of sub-Neptunes.

[3] Transit technique tends to favour larger planets.

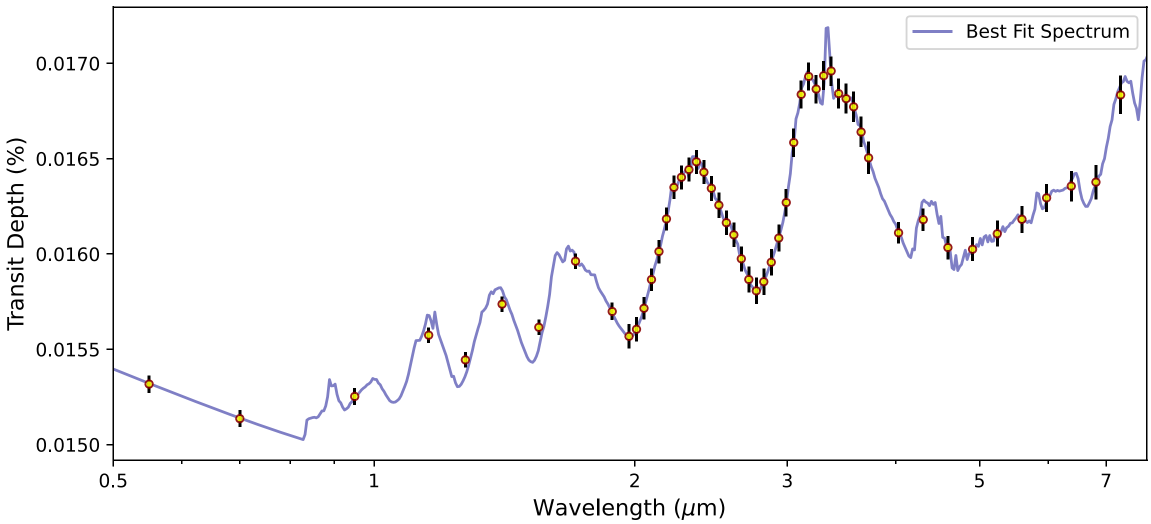

Figure 3 Example of a simulated Ariel observation with errorbars (data points) for a randomized chemistry. The best-fitting model obtained using atmospheric retrieval is also shown (solid line). The slope at the lowest wavelengths arises from Rayleigh Scattering, while most of the other spectral modulations in this example can be attributed to CH_4. The data points around 4.5 μm are associated with CO and CO_2 absorption. Note the difference in wavelength coverage (0.5–7.8 μm) as compare to the *HST* spectrum (1.1–1.7 μm) in Fig. A2, which allows us to extract precise information for many molecules.

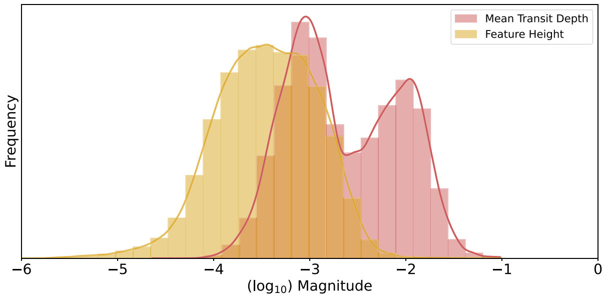

Figure 4 Distribution of the mean transit depth (red) overlapped with the distribution of the feature height (orange), both measured in a logarithm scale. The dichotomy displayed in mean transit depth distribution stemmed from the observational demographics of planet radius, showing the diversity of currently known exoplanets in our data set. On the other hand, the feature height documents the 'strength' of spectroscopic features in each spectrum (such as absorption features or strong trends induced by Rayleigh Scattering). Any successful model must be able to account for the variations in both scales.

between different spectra, it also needs to take into account the intra-variation across wavelength channels, which is always one to three orders of magnitude smaller than the variation in mean transit depth.

Next we will look at results from atmospheric retrieval. The quality of the retrieved product is closely related to the information content of individual spectrum, which is a function of the wavelength coverage, size of the spectral bin, observational uncertainties and the abundance of the molecule. Fig. 5 compares the retrieval results against the input values of the six targets of interest (H_2O, CO_2, CH_4, CO, NH_3, temperature). Each data point in every subplot represents a single spectrum and is coloured in accordance to the size of the inter-quartile range (IQR).[4] Points lying along the diagonal line – those that are retrieved correctly – tend to have tighter constraint, while points that deviate from the diagonal line tend to entail larger uncertainties. For most gases there is a transition region where molecules at certain abundance level starts to depart from the diagonal line. The extent and onset of the transition region is a function of the instrument specification (e.g. its detection limits), the composition of the atmosphere and the strength of the molecular absorption. Changeat et al. (2020a) pioneered an initial study of this transition region and derived the detection limit for each gas based on the size of the errorbar obtained. Here, we find similar results, and the detection limits of Ariel correspond to the region where all the retrieved values from Fig. 5 deviate from the diagonal line (associated with colours from green to red).

Appendix D continues our discussion into other aspects of the data product.

2.5 Structure of the ABC data base

The data base contains two levels of data product: The first level is for general use and the second level is designed specifically for the competition. We will describe each level next.

2.5.1 Level 1: cleaned data

Level 1 contains data products for general use. As *TauREx 3* performs forward modelling and retrieval on a planet-by-planet basis. The data are pre-processed to provide an unified structure for effective data navigation and a foundation for further processing. Below is the list of operations we performed:

(i) Removed any spectra with `NaN` values.
(ii) Removed spectra with transit depth larger than 0.1 in any wavelength bins.
(iii) Removed spectra with transit depth smaller than 1×10^{-8} in any wavelength bins.
(iv) Standardized units and data formats.
(v) Extracted all Stellar, Planetary and Instrumental metadata.
(vi) Combined all instances into a single, unified file.

Level 1 data are organized into `all_data.csv`, `observations.hdf5` and `all_target.hdf5`. `all_data.csv` contains information on the planetary system and the input values for

[4] Here we define IQR as the difference between the 84th and the 16th percentile.

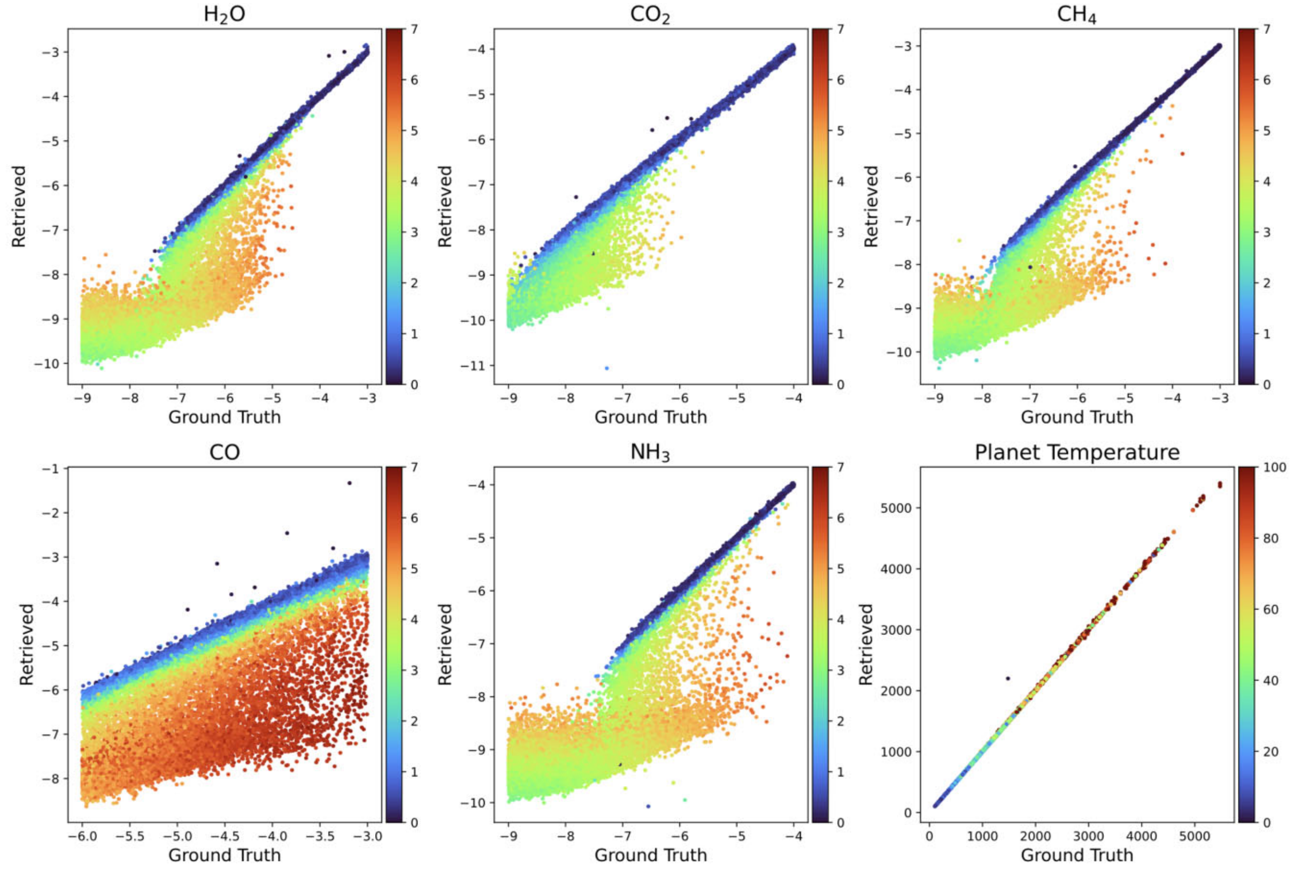

Figure 5 Comparison of the retrieved values against the input values for six different targets. Each data point represents a single instance and is colour-coded according to the respective size of the IQR. Ariel data at Tier-2 resolution are able to place tight constrain on the temperature and most molecules down to a certain abundance. Beyond that, the retrieved values starts to deviate from the diagonal and becomes less constrained, highlighting the limitations of the telescope.

the generation process, `observations.hdf5` contains information on individual observations and `all_target.hdf5` contains the corresponding retrieval results (posterior distributions of each atmospheric targets). In total, there are 105 887 planet instances, 25 per cent of them (26 109) has complementary retrievals from Nested Sampling.

2.5.2 *Level 2: curated data for model training*

The following section is designed for statistical model training. In order to allow for the broadest possible participation and minimize the overhead for non-field experts, we pre-processed the data set with our domain knowledge so that the end product is ready for model development. At the same time, we have tailored the train/test split procedure in order to allow a diverse array of solutions and research directions. Here we outlined the list of operations we performed:

(i) Removed data with less than 1500 points in the tracedata. This is to allow for more accurate comparison.

(ii) Removed un-informative and duplicated astrophysical and instrumental features.[5]

(iii) Split data into training and test sets (more details in Appendix E).

After performing the above operations, the training data has 91 392 planet instances with 21 988 of them has complementary retrievals results. The test data have 2997 instances, all of which are complemented with retrieval results. There is a notable difference in terms of the volume of data between Level 1 and Level 2 data *due to the pre-processing step and train/test split*. We have devoted a section in Appendix E to describe the Level 2 data in details.

2.6 Additional resources

Published along with the data base, we provide a series of complementary resources. In particular the data base is provided with a Jupyter Notebook describing the data structure, how to load the data set, and demonstrating its main characteristics. We also include a dedicated *TauREx 3* tutorial for those eager to learn the practical aspects of building forward models and performing atmospheric retrievals. All those resources are available under the same link as the data base.

3 OPEN CHALLENGES

With the constructed data set, we intended to accelerate and incentivize dedicated efforts to tackle a number of open challenges common to both the exoplanet field and the ML field.

[5] Including star_magnitudeK, star_metallicity, star_type, planet_type, star_type, star_mass_kg, star_radius_m, planet_albedo, planet_impact_param, planet_mass_kg, planet_radius_m, planet_transit_time, instrument_nobs.

3.1 Fast and accurate Bayesian Inference

One of the aims of the data base is to enable the development of advanced inference methods that are (1) able to produce posterior distributions, but at the same time, will not require as much computational resources compared to conventional sampling-based methods. This activity is proposed as part of the goal of the NeurIPS 2022 competition with simplified atmospheric cases and has already proven very successful (Yip et al. in preparation).

3.2 Estimating and mitigating the effect of data shifts

ML models are prone to potential performance degradation when the incoming data are different from the training distribution. This phenomenon is commonly known as data shifts (e.g. Lu et al. 2018; Bayram, Ahmed & Kassler 2022).

Any ML application to the study of exoplanetary atmosphere are likely to experience data shifts. Most ML models in the literature are currently limited to simulation-based inference as the amount of actual spectroscopic observations are fall short for model training, which has to be supplemented by simulations. The discrepancy between our simplistic atmospheric models and the actual atmosphere means that data shift is inevitable (Humphrey et al. 2022).

To emulate this situation, the test set in level 2 data are specifically designed to include chemical equilibrium forward models for which the provided ground truth from atmospheric retrievals assumed free chemistry. In some cases, clouds are included in the forward model to force degenerate behaviours in the test set (Line & Parmentier 2016; Pinhas & Madhusudhan 2017; Mai & Line 2019; Barstow 2020; Mukherjee, Batalha & Marley 2021; Changeat et al. 2021b). Those offsets between training and test sets were voluntarily introduced to evaluate whether the performance of ML solutions remain robust and consistent under ‘unseen’ distributions (this is typically the case in real life since we know little about real exo-atmospheres) and if they had correctly learned to faithfully reproduce the Bayesian retrieval technique.

3.3 Adaptation to other atmospheric assumptions

Atmospheric models are physical models built on varying level of complexities and modelling assumptions. ML models, however, are trained to optimize their performance with respect to the provided training set/training assumptions. In this data set, we have included forward models built from two different modeling assumptions, Simple chemsitry and Equilibrium Chemistry. It remains an open question as to how easy one can ‘switch’ from one model assumption to another. In terms of ML terminology, this kind of learning falls under the umbrella of transfer learning/domain adaptation, where one strives to adopt to from source domain (original training set) to the target domain with a limited number of training examples (Wilson & Cook 2020).

3.4 Benchmarking different retrieval techniques

The built data set can be used for more traditional code comparisons. The *TauREx* retrieval code was rigorously benchmarked against other established codes (Barstow et al. 2020, 2022). With this data set, the exoplanet community now has access to a wide range of well-referenced forward models and retrieval runs that can be used for standard benchmarking of atmospheric models (e.g. forward models) and a diverse array of retrieval techniques (e.g. MCMC, Nested Sampling, Normalizing Flows: Foreman-Mackey et al. 2013; Feroz et al. 2009; Buchner 2021; Yip et al.2022a).

4 FUTURE EXPANSION OF ABC DATABASE

The data base currently builds on highly simplified atmospheric model assumptions (constant or equilibrium chemistry, isothermal temperature, clear atmosphere). This is done to (1) gauge the success of such initiatives and (2) provide a rich data set to complete the required task.

Future iterations could explore more complex atmospheres with much more limited amount of training examples. This is because, as more complexity is embedded into the model (e.g. GCMs, complex chemistry, stellar activity effects), the computation of a single sample can take months. In this instance traditional parameter sampling is not an option, and faster AI accelerated techniques will be required. We therefore plan to further extend this data base over the coming years and provide new training/test sets to develop both exoplanet and ML activities. For example, future instances of this data base could feature the following:

(i) *JWST* and *HST* complementary data sets: This would allow to develop telescope-independent ML techniques and evaluate information content between the different data sets.

(ii) Other classes of exoplanets: The current set focuses on gaseous exoplanets. Future data releases could include small rocky exoplanets with secondary atmospheres, or water worlds.

(iii) More complex processes: Alternative chemical model (with more complete species sets, with dis-equilibrium processes: Stock et al. 2018; Woitke et al. 2018; Venot et al. 2020; Al-Refaie et al. 2022b) could be provided to study retrieval biases and develop chemistry robust ML methods. Similarly, complementary sets could include stellar activity, for which the relevance of AI methods has already been shown (Nikolaou et al. 2020), or even complex cloud models (Ackerman & Marley 2001; Kawashima & Ikoma 2018; Gao et al. 2020; Ma et al.2022).

(iv) More complex models: *Eclipse* observations or phase-curve observations produced using Global Climate Model could be included. This would allow to extend this to new observations as well as studying three-dimensional effects (Cho et al. 2003; Rauscher et al. 2008; Dobbs-Dixon, Cumming & Lin 2010; Showman, Cho & Menou 2010; Cho, Polichtchouk & Thrastarson 2015; Caldas et al. 2019; Komacek & Showman 2020; Skinner & Cho 2022) and to develop fast recovery techniques for phase-curve data. Current approaches to retrieve phase-curve data are limited by computational resources (Feng, Line & Fortney 2020; Irwin et al. 2020; Cubillos et al. 2021; Changeat et al. 2021a; Changeat 2022; Chubb & Min 2022) and can require up to 10 million samples (e.g. weeks on HPC facilities) to fully explore the parameter space of solution with Hubble data (Changeat et al.2021a).

5 CONCLUSIONS

We present here the publicly available ABC Database (https://doi.org/10.5281/zenodo.6770103), a data base of atmospheric forward and inverse models dedicated to the development of ML approaches in the field of exoplanets. In this paper, we introduce, for a non-expert community, the basic physical and chemical processes involved in the creation of such data base, describing the utilized tools,[6] and clearly

[6] The main simulation code, TAUREX 3, is open-source and publicly available at: https://github.com/ucl-exoplanets/TauREx3_public.

stating the adopted hypothesis. The constructed set includes about 105 887 forward models and 26 109 atmospheric retrievals from conventional sampling techniques, and should serve as a community asset to explore novel techniques to solve the inverse problem of retrieving chemical composition from spectroscopic data. This data base was used to support the third instalment of the Ariel Data Challenge, conducted as part of the NeurIPS Conference,[7] which led to new innovative ML-based solutions to infer posterior distributions from Ariel spectra. With this effort, and with future updates of this permanent data base, we hope to facilitate the development and adoption of ML solutions to a pressing issue for the next generation of space telescopes.

ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758892, ExoAI), the Science and Technology Funding Council grants ST/S002634/1 and ST/T001836/1, and the UK Space Agency grant ST/W00254X/1. QC is funded by the European Space Agency under the 2022 ESA Research Fellowship Program. The author thanks Ingo P. Waldmann, Giovanna Tinetti, and Ahmed F. Al-Refaie for their useful recommendations and discussions. The authors wish to thank two anonymous referees for their useful comments.

This work utilized resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

DATA AVAILABILITY

The data underlying this paper are available as a Zenodo Digital Repository, at https://doi.org/10.5281/zenodo.6770103.

REFERENCES

Ackerman A. S., Marley M. S., 2001, ApJ, 556, 872
Agúndez M., Venot O., Iro N., Selsis F., Hersant F., Hébrard E., Dobrijevic M., 2012, A&A, 548, A73
Agúndez M., Martínez J. I., de Andres P. L., Cernicharo J., Martín-Gago J. A., 2020, A&A, 637, A59
Al-Refaie A. F., Changeat Q., Waldmann I. P., Tinetti G., 2021, ApJ, 917, 37
Al-Refaie A. F., Changeat Q., Venot O., Waldmann I. P., Tinetti G., 2022a, ApJ, 932, 123
Al-Refaie A. F., Venot O., Changeat Q., Edwards B., 2022b, preprint (arXiv:2209.11203)
Ardevol Martinez F., Min M., Kamp I., Palmer P. I., 2022, A&A, 662, A108
Baeyens R., Konings T., Venot O., Carone L., Decin L., 2022, MNRAS, 512, 4877
Baker J., Fearnhead P., Fox E. B., Nemeth C., 2019, Stat. Comput., 29, 599
Barstow J. K., 2020, MNRAS, 497, 4183
Barstow J. K., Changeat Q., Garland R., Line M. R., Rocchetto M., Waldmann I. P., 2020, MNRAS, 493, 4884
Barstow J. K., Changeat Q., Chubb K. L., Cubillos P. E., Edwards B., MacDonald R. J., Min M., Waldmann I. P., 2022, Exp. Astron., 53, 447
Batalha N. M., 2014, Proc. Natl. Acad. Sci., 111, 12647
Bayes T., Price n., 1763, Phil. Trans. R. Soc. A, 53, 370
Bayram F., Ahmed B. S., Kassler A., 2022, Knowl.-Based Syst., 245, 108632
Borucki W. J. et al., 2010, Science, 327, 977
Brogi M., Line M. R., 2019, AJ, 157, 114
Buchner J., 2021, preprint (arXiv:2101.09675)
Caldas A., Leconte J., Selsis F., Waldmann I. P., Bordé P., Rocchetto M., Charnay B., 2019, A&A, 623, A161
Cassan A. et al., 2012, Nature, 481, 167
Changeat Q., 2022, AJ, 163, 106
Changeat Q., Edwards B., 2021, ApJ, 907, L22
Changeat Q., Edwards B., Waldmann I. P., Tinetti G., 2019, ApJ, 886, 39
Changeat Q., Al-Refaie A., Mugnai L. V., Edwards B., Waldmann I. P., Pascale E., Tinetti G., 2020a, AJ, 160, 80
Changeat Q., Edwards B., Al-Refaie A. F., Morvan M., Tsiaras A., Waldmann I. P., Tinetti G., 2020b, AJ, 160, 260
Changeat Q., Al-Refaie A. F., Edwards B., Waldmann I. P., Tinetti G., 2021a, ApJ, 913, 73
Changeat Q., Edwards B., Al-Refaie A. F., Tsiaras A., Waldmann I. P., Tinetti G., 2021b, Exp. Astron., 53, 391
Changeat Q. et al., 2022, ApJS, 260, 3
Charbonneau D., Brown T. M., Noyes R. W., Gilliland R. L., 2002, ApJ, 568, 377
Chen J., Kipping D., 2017, ApJ, 834, 17
Chen T., Fox E., Guestrin C., 2014, Proc. 31st International Conference on Machine Learning, Vol. 32. PMLR, Bejing, p. 1683
Chen C., Carlson D., Gan Z., Li C., Carin L., 2016, Proc. 19th International Conference on Artificial Intelligence and Statistics, Vol. 51. PMLR, Cadiz, p. 1051
Chen G., Wang H., van Boekel R., Pallé E., 2022, AJ, 164, 173
Cho J. Y.-K., Menou K., Hansen B. M. S., Seager S., 2003, ApJ, 587, L117
Cho J. Y.-K., Polichtchouk I., Thrastarson H. T., 2015, MNRAS, 454, 3423
Chubb K. L., Min M., 2022, A&A, 665, A2
Chubb K. L. et al., 2021, A&A, 646, A21
Cloutier R., Menou K., 2020, AJ, 159, 211
Cobb A. D. et al., 2019, AJ, 158, 33
Coles P. A., Yurchenko S. N., Tennyson J., 2019, MNRAS, 490, 4638
Cubillos P. E. et al., 2021, ApJ, 915, 45
de Wit J. et al., 2018, Nat. Astron., 2, 214
Dobbs-Dixon I., Cumming A., Lin D. N. C., 2010, ApJ, 710, 1395
Drummond B., Tremblin P., Baraffe I., Amundsen D. S., Mayne N. J., Venot O., Goyal J., 2016, A&A, 594, A69
Edwards B., Tinetti G., 2022, AJ, 164, 15
Edwards B., Mugnai L., Tinetti G., Pascale E., Sarkar S., 2019a, AJ, 157, 242
Edwards B. et al., 2019b, Exp. Astron., 47, 29
Edwards B. et al., 2020, AJ, 160, 8
Edwards B. et al., 2022, preprint (arXiv:2211.00649)
Estrela R., Swain M. R., Roudier G. M., 2022, ApJ, 941, L5
Feng Y. K., Line M. R., Fortney J. J., 2020, AJ, 160, 137
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Fulton B. J. et al., 2017, AJ, 154, 109
Gal Y., Ghahramani Z., 2015, preprint (arXiv:1506.02142)
Gandhi S., Madhusudhan N., 2018, MNRAS, 474, 271
Gao P. et al., 2020, Nat. Astron., 4, 951
Greene T. P., Line M. R., Montero C., Fortney J. J., Lustig-Yaeger J., Luther K., 2016, ApJ, 817, 17
Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, Proc. 34th International Conference on Machine Learning, Vol. 70. PMLR, p. 1321
Guzmán-Mesa A. et al., 2020, AJ, 160, 15
Haldemann J. et al., 2022, preprint (arXiv:2202.00027)
Harrington J. et al., 2022, Planet. Sci. J., 3, 80
Hayes J. J. C. et al., 2020, MNRAS, 494, 4492
Himes M. D. et al., 2022, Planet. Sci. J., 3, 91
Hoeijmakers H. J. et al., 2018, Nature, 560, 453
Homan M. D., Gelman A., 2014, J. Mach. Learn. Res., 15, 1593
Howard A. W. et al., 2010, Science, 330, 653

[7] https://neurips.cc/Conferences/2022/CompetitionTrack.

Humphrey A., Kuberski W., Bialek J., Perrakis N., Cools W., Nuyttens N., Elakhrass H., Cunha P., 2022, MNRAS, 517, L116

Irwin P. G. J. et al., 2008, J. Quant. Spec. Radiat. Transf., 109, 1136

Irwin P. G. J., Parmentier V., Taylor J., Barstow J., Aigrain S., Lee E. K. H., Garland R., 2020, MNRAS, 493, 106

Izmailov P., Vikram S., Hoffman M. D., Wilson A. G., 2021, preprint (arXiv:2104.14421)

Kawashima Y., Ikoma M., 2018, ApJ, 853, 7

Komacek T. D., Showman A. P., 2020, ApJ, 888, 2

Kreidberg L. et al., 2014, Nature, 505, 69

Lakshminarayanan B., Pritzel A., Blundell C., 2017, Proc. 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, p. 6405

Lavie B. et al., 2017, AJ, 154, 91

Li G., Gordon I. E., Rothman L. S., Tan Y., Hu S.-M., Kassi S., Campargue A., Medvedev E. S., 2015, ApJS, 216, 15

Line M. R., Parmentier V., 2016, ApJ, 820, 78

Line M. R., Zhang X., Vasisht G., Natraj V., Chen P., Yung Y. L., 2012, ApJ, 749, 93

Line M. R. et al., 2013, ApJ, 775, 137

Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G., 2018, IEEE Trans. Knowl. Data Eng., 31, 2346

Ma Y., Ma Y.-A., Chen T., Fox E. B., 2015, NIPS

Ma S., Ito Y., Al-Refaie A. F., Changeat Q., Edwards B., Tinetti G., 2022, preprint (arXiv:2301.13708)

Maddox W. J., Garipov T., Izmailov P., Vetrov D., Wilson A. G., 2019, A Simple Baseline for Bayesian Uncertainty in Deep Learning. Curran Associates Inc., Red Hook, NY

Madhusudhan N., Seager S., 2009, ApJ, 707, 24

Madhusudhan N., Agúndez M., Moses J. I., Hu Y., 2016, Space Sci. Rev., 205, 285

Mai C., Line M. R., 2019, ApJ, 883, 144

Mandt S., Hoffman M. D., Blei D. M., 2017, J. Mach. Learn. Res., 18, 1

Márquez-Neila P., Fisher C., Sznitman R., Heng K., 2018, Nat. Astron., 2, 719

Mayor M., Queloz D., 1995, Nature, 378, 355

Mikal-Evans T. et al., 2022, Nat. Astron., 6, 471

Min M., Ormel C. W., Chubb K., Helling C., Kawashima Y., 2020, A&A, 642, A28

Mollière P., Wardenier J. P., van Boekel R., Henning T., Molaverdikhani K., Snellen I. A. G., 2019, A&A, 627, A67

Mugnai L. V., Pascale E., Edwards B., Papageorgiou A., Sarkar S., 2020, Exp. Astron., 50, 303

Mukherjee S., Batalha N. E., Marley M. S., 2021, ApJ, 910, 158

Neal R., 2011, Handbook of Markov Chain Monte Carlo. Chapman & Hall, New York

Nemeth C., Fearnhead P., 2019, preprint (arXiv:1907.06986)

Nikolaou N. et al., 2020, preprint (arXiv:2010.15996)

Nixon M. C., Madhusudhan N., 2020, MNRAS, 496, 269

Oreshenko M. et al., 2020, AJ, 159, 6

Pätzold M. et al., 2012, A&A, 545, A6

Pearce T., Leibfried F., Brintrup A., Zaki M., Neely A., 2020, 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020

Petigura E. A. et al., 2022, AJ, 163, 179

Pinhas A., Madhusudhan N., 2017, MNRAS, 471, 4355

Polyansky O. L., Kyuberis A. A., Zobov N. F., Tennyson J., Yurchenko S. N., Lodi L., 2018, MNRAS, 480, 2597

Potthast R., 2006, Inverse Probl., 22, R1

Rauscher E., Menou K., Cho J. Y.-K., Seager S., Hansen B. M. S., 2008, ApJ, 681, 1646

Ricker G. R. et al., 2015, J. Astron. Telesc. Instrum. Syst., 1, 014003

Ritter H., Botev A., Barber D., 2018, 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings, Vol. 6

Rocchetto M., Waldmann I. P., Venot O., Lagage P. O., Tinetti G., 2016, ApJ, 833, 120

Roudier G. M., Swain M. R., Gudipati M. S., West R. A., Estrela R., Zellem R. T., 2021, AJ, 162, 37

Schwarz H., Brogi M., de Kok R., Birkby J., Snellen I., 2015, A&A, 576, A111

Sharma S., 2017, ARA&A, 55, 213

Showman A. P., Cho J. Y.-K., Menou K., 2010, in Seager S., ed., Atmospheric Circulation of Exoplanets. Space Science Series of the Univ. Arizona Press, Tucson, AZ

Sing D. K. et al., 2016, Nature, 529, 59

Skilling J., 2006, Bayesian Anal., 1, 833

Skinner J. W., Cho J. Y.-K., 2022, MNRAS, 511, 3584

Soboczenski F. et al., 2018, preprint (arXiv:1811.03390)

Speagle J. S., 2020, MNRAS, 493, 3132

Stevenson K. B. et al., 2017, AJ, 153, 68

Stock J. W., Kitzmann D., Patzer A. B. C., Sedlmayr E., 2018, MNRAS, 479, 865

Swain M. R., Vasisht G., Tinetti G., 2008, Nature, 452, 329

Taylor J., Parmentier V., Irwin P. G. J., Aigrain S., Lee E. K. H., Krissansen-Totton J., 2020, MNRAS, 493, 4342

Taylor J., Parmentier V., Line M. R., Lee E. K. H., Irwin P. G. J., Aigrain S., 2021, MNRAS, 506, 1309

The JWST Transiting Exoplanet Community Early Release Science Team, 2022, preprint (arXiv:2208.11692)

Tinetti G. et al., 2007, Nature, 448, 169

Tinetti G. et al., 2021, preprint (arXiv:2104.04824)

Trotta R., 2017, preprint (arXiv:1701.01467)

Tsiaras A. et al., 2018, AJ, 155, 156

Tsiaras A., Waldmann I. P., Tinetti G., Tennyson J., Yurchenko S. N., 2019, Nat. Astron., 3, 1086

Venot O., Agúndez M., 2015, Exp. Astron., 40, 469

Venot O., Cavalié T., Bounaceur R., Tremblin P., Brouillard L., Lhoussaine Ben Brahim R., 2020, A&A, 634, A78

Waldmann I. P., 2016, ApJ, 820, 107

Waldmann I. P., Rocchetto M., Tinetti G., Barton E. J., Yurchenko S. N., Tennyson J., 2015a, ApJ, 813, 13

Waldmann I. P., Tinetti G., Rocchetto M., Barton E. J., Yurchenko S. N., Tennyson J., 2015b, ApJ, 802, 107

Wang M., Deng W., 2018, Neurocomputing, 312, 135

Welbanks L., Madhusudhan N., Allard N. F., Hubeny I., Spiegelman F., Leininger T., 2019, ApJ, 887, L20

Welling M., Teh Y. W., 2011, Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11. Omnipress, Madison, WI, p. 681

Wilson G., Cook D. J., 2020, ACM Trans. Intell. Syst. Technol., 11, 51

Woitke P., Helling C., Hunter G. H., Millard J. D., Turner G. E., Worters M., Blecic J., Stock J. W., 2018, A&A, 614, A1

Wolszczan A., Frail D. A., 1992, Nature, 355, 145

Xing C., Arpit D., Tsirigotis C., Bengio Y., 2018, preprint (arXiv:1802.08770)

Yao J., Pan W., Ghosh S., Doshi-Velez F., 2019, Proceedings at the International Conference on Machine Learning: Workshop on Uncertainty & Robustness in Deep Learning (ICML).

Yip K. H., Tsiaras A., Waldmann I. P., Tinetti G., 2020, AJ, 160, 171

Yip K. H., Changeat Q., Edwards B., Morvan M., Chubb K. L., Tsiaras A., Waldmann I. P., Tinetti G., 2021a, AJ, 161, 4

Yip K. H., Changeat Q., Nikolaou N., Morvan M., Edwards B., Waldmann I. P., Tinetti G., 2021b, AJ, 162, 195

Yip K. H., Changeat Q., Al-Refaie A., Waldmann I., 2022a, preprint (arXiv:2205.07037)

Yip K. H. et al., 2022b, The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS 2022)

Yurchenko S. N., Amundsen D. S., Tennyson J., Waldmann I. P., 2017, A&A, 605, A95

Yurchenko S. N., Mellor T. M., Freedman R. S., Tennyson J., 2020, MNRAS, 496, 5282

Zhang M., Chachan Y., Kempton E. M. R., Knutson H. A., 2019, PASP, 131, 034501

Zingales T., Waldmann I. P., 2018, AJ, 156, 268

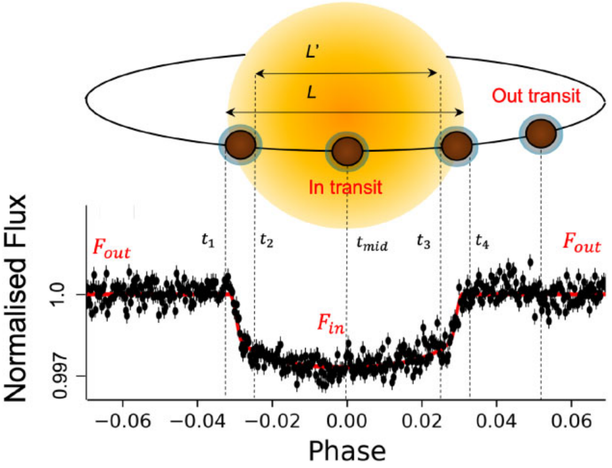

Figure A1 Diagram of an observation of the transiting exoplanet KELT-11 b (top panel) and the corresponding normalized flux from a real observation, also called a light curve (bottom panel). The phase, which labels the x-axis, is the position of the planet in its orbit with 0 (by convention) being the middle of the transit (t_{mid}). The transit starts at the event t_1 and finishes at the event t_4, spanning the transit duration t_{14}. The transit depth (Δ) is the observed normalized flux between in and out of transit situations. The observation is adapted from Changeat et al. (2020b).

APPENDIX A: INTRODUCTION TO ATMOSPHERIC STUDIES OF EXOPLANETS

This section provides a summary of the domain knowledge required to properly exploit the ABC Database. It is written as an introduction for non-exoplanet audience.

A1 Observations of transiting exoplanets

Exoplanets are detected using various methods, but the two most popular techniques used today are radial velocity and transit. In particular, transit is an indirect technique that relies on monitoring the host star's variations in brightness. A transit event occurs when the planet passes in front of the star, blocking a fraction of the light received here on Earth. Transit events can be observed, thus revealing the presence of the planet and its important properties, such as radius. A typical transit observation is described, along with the relevant quantities, in Fig. A1. Transit events are periodic, so they can easily be disentangled from other astrophysical sources of noise (stellar variations,[8] instrument systematics and observing conditions) when long-term monitoring is employed.

For most observatories, absolute measurements are challenging. This is especially true when the required precision is high, as it is the case for exoplanets. As such, for exoplanets, we prefer to rely on differential quantities such as transit depth (Δ). The transit depth is the normalized difference between the flux received from the star when the planet is out-of-transit (F_{out}) and when the planet is in-transit (F_{in}):

$$\Delta = \frac{F_{\mathrm{out}} - F_{\mathrm{in}}}{F_{\mathrm{out}}} = \left(\frac{R_{\mathrm{p}}}{R_{\mathrm{s}}}\right)^2, \qquad \text{(A1)}$$

where R_{p} is the radius of the planet and R_{s} is the radius of the star.

To first order, to account for the contribution of an atmosphere, one can simply replace the planetary radius R_{p} by $R_{\mathrm{p}} + h$, where h is the effective size of the atmosphere. Neglecting second order terms, this gives

$$\Delta = \left(\frac{R_{\mathrm{p}}}{R_{\mathrm{s}}}\right)^2 + \frac{2R_{\mathrm{p}}h}{R_{\mathrm{s}}^2}. \qquad \text{(A2)}$$

Now, crudely, the size of the atmosphere depends on the atmospheric scaleheight H such that $h = NH$, where N is a scaling factor encoding information regarding the atmospheric compositions. The scaleheight is defined as

$$H = \frac{k_{\mathrm{b}}T}{\mu g}, \qquad \text{(A3)}$$

where k_{b} is the Boltzmann constant, T is the temperature, μ is the mean molecular weight, and g is the gravity.

From those simple expressions, which here serve an illustrative purpose and are an oversimplification of the model used to build the ABC Database, we can deduce some standard behaviours of atmospheric properties. First, to extract information on the planet and its atmosphere, we will always require some knowledge of the host star. This is because the planet is not observed directly and the observed quantities (Δ) are a function of the stellar parameters (here the stellar radius R_{s}). In addition, we observe that for the more massive planet (larger g) the contribution of the atmosphere will be diminished, as the atmosphere contracts under gravity. In contrast, if the temperature increases the atmosphere will be inflated and thus the atmospheric signal will be larger. The chemistry of the atmosphere plays a part in the scaling factor N but their relation cannot be easily deduced here. Intuitively, molecules with larger abundance tend to make the atmosphere opaque at higher altitudes, therefore increasing the apparent size of the atmosphere.[9]

Those concepts, while useful to acquire an intuitive understanding of the behaviour of planetary atmospheres, are rather limiting and proper modelling is required to correctly interpret exo-atmospheric observations.

A2 Modelling exoplanet atmospheres

Observing exoplanetary transits at various wavelengths, meaning obtaining Δ as a function of λ, provides information about the atmospheric properties. This is because a planetary atmosphere contributes to the transit depth by absorbing the incoming stellar light (slightly) differently at different wavelengths (e.g. the atmospheric contribution is wavelength-dependent). The absorption profile of the atmosphere depends on its constituents (molecular species, clouds, hazes) and properties (thermal structure). To model the observed signal as a function of wavelengths, also called a spectrum, astronomers use simplified models of the relevant processes occurring in exoplanet atmospheres. In Appendix B, we describe the mathematical formulation of one such model for the transit geometry, commonly used as a parametrized 1D forward model. Put simply, the light from the host star is propagated through an atmosphere layer by layer and impacted according to the absorption of the atmosphere. In our case, the absorptions considered are absorptions by molecular species, Rayleigh Scattering, and Collision Induced Absorption (CIA).

Through this process, from a parametrized one-dimensional description of an atmosphere controlled by a finite number of parameters, one can compute the theoretical spectrum of an exoplanet. This process, called forward modelling, can be made relatively

[8] For example, stars' brightness could vary from time to time.

[9] we cannot observe any non-opaque (transparent) part of the atmosphere.

fast (on the order of seconds), but due to the non-linearity of the equations and the input spectroscopic data (cross-sections), it cannot be directly inverted. In the next section, we explain how traditional techniques (e.g. Bayesian sampling) are used to perform model fitting and retrieve the properties of an exo-atmosphere from its observed spectrum.

Before explaining the use of inversion techniques, or atmospheric retrievals, applied in the context of exoplanet atmospheres, we wish to present a series of simple models to illustrate further the sensitivity analysis made in the previous section. We have created a mock planet with a non-negligible atmosphere, and we will show how changing the values of some of the model parameters affects the observation (e.g. the spectrum).

For simplicity, we set the planet with an isothermal atmosphere, meaning the temperature of the atmosphere is constant with altitude (e.g. constant at all pressure levels) and therefore can be defined by a single parameter (*T*). To this atmosphere, we add a single trace molecule (H_2O) defined by its absolute abundance in volume (volume mixing ratio), and we fill the rest remaining atmosphere with hydrogen and helium in standard solar ratios ($H_2/He = 0.17$).[10] On top of the molecular absorption from water vapour, we also consider three additional absorption processes: CIA, Rayleigh scattering, and Grey Clouds (not considered in this version of the ABC data base). Equipped with this model, we set the following cases for which the spectra are available in Fig. A2:

(i) Case 1 (black): planetary radius $R_{\rm p} = 1.0R_{\rm J}$, temperature $T =$ 1200 K, nixing ratio of $H_2O = 10^{-3}$, and no clouds.

(ii) Case 2 (blue): same as Case 1 but the temperature is decrease to $T = 500$ K.

(iii) Case 3 (purple): same as Case 1 but the water content is decrease to $H_2O = 10^{-5}$, while the planetary radius is increased to $R_{\rm p} = 1.0085$.

(iv) Case 4 (red): same as Case 1 but with clouds (cloud top pressure is set at 0.01 bar).

(v) Case 5 (green): same as Case 2 but with an increased planetary radius to $R_{\rm p} = 1.013R_{\rm J}$.

From those specifically designed case, one can compare Case 1 and 2, for which only the temperature is changed. As a consequence of this change, the size of the atmosphere is decreased as explained in Section 2.1, and the atmospheric features are smaller, bringing the whole spectrum down. In this case, distinguishing between Case 1 and Case 2 would be relatively easy. For Cases 3, 4, and 5, however, the story can be a little more complicated as multiple parameters are changed, but those can be used to highlight degeneracies typically encountered in the interpretation of exoplanet spectra, therefore justifying the need for more sophisticated atmospheric retrieval techniques.

For those cases, the spectral features are reduced compared to Case 1, but they appear much closer in the 1–2 μm wavelength range. This is because Case 3 has less water compared to Case 1, which we expect to decrease the spectral features but thanks to the slightly larger radius, the spectrum is brought back to a similar level. Case 4 has opaque clouds, which 'cuts' the spectral features above a certain pressure level, making it look exactly like Case 3 in the 1–2 μm range. Finally, Case 5 has a lower temperature (600 K) and is brought back to the same level by an increased in radius. With current telescopes, such as *HST*, the wavelength coverage is relatively small. One typical instrument onboard *HST* is the Wide Field Camera 3 with its G141 Grism, which has a wavelength coverage from 1.1 to 1.6 μm and reaches errors of the order of 30 ppm.[11] Highlighting a typical observation with *HST* on the same figure, we show how difficult it would be to distinguish between Cases 3, 4, and 5. This highlight the requirement to next-generation space telescopes such as Ariel to constrain atmospheric properties.

A3 Solving the inverse problem for exo-atmospheres

The study of exo-atmosphere relies on spectroscopic observations to infer fundamental atmospheric properties that cannot be directly observed. This kind of problem is broadly described as the inverse problem (Potthast 2006), where one tries to uncover the cause (atmospheric properties) from the effect (observations). However, more often than not, the full effect is seldom observed, instead, observers often received a corrupted form of the effect, which is the observations. In terms of exoplanetary spectra, there are several sources of corruption, such as the presence of noise and limited spectroscopic coverage. The loss of information often means that the inverse mapping function $\mathcal{M}^{-1}$ is unknown and may no longer be uniquely defined, which generally give rise to more than one plausible causes, also known as model degeneracy (see section above). In some extreme cases, severe loss in information (like extremely low S/N observations) effectively means that the cause may no longer be recoverable. See Fig. A3 for a typical setup of an inverse problem.

Our goal is to estimate the set of parameters Θ that best explains the observed spectrum *D* under a given atmospheric model $\mathcal{M}$. There are different ways to approach this 'atmospheric retrieval' problem, but most of them involve a forward model (which includes our atmospheric assumptions) and an optimizer. Here we will briefly describe the problem in terms of Bayesian framework, for a more detailed discussion on Bayesian statistics, one can refer to Skilling (2006), Feroz et al. (2009), Foreman-Mackey et al. (2013), Sharma (2017), Trotta (2017), Speagle (2020), and Buchner (2021) for more information.

Our goal is to find the conditional distribution of the model parameters given the observation, also known as the posterior distribution ($P(\Theta|D,\mathcal{M}) = P(D|\Theta,\mathcal{M})P(\Theta)/P(D)$) in Bayes' theorem (Bayes & Price 1763). The posterior distribution can be computed via the following formulation:

$$P(\Theta|D,\mathcal{M}) = \frac{P(D|\Theta,\mathcal{M})P(\Theta)}{P(D)}, \tag{A4}$$

where $P(D|\Theta,\mathcal{M})$ represents the likelihood function under a given model, $P(\Theta)$ represents the prior, and $P(D)$ represents the normalizing constant, or the Bayesian Evidence.

The (log-) Gaussian likelihood function is commonly used to compare the observation *D* with the output from the forward model $\mathcal{M}$, i.e.

$$\mathbb{E}[\log P(D|\Theta,\mathcal{M}] = \mathbb{E}[\log(\mathcal{N}(D|S,\sigma)] \tag{A5}$$

$$= \mathbb{E}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(D-S)^2}{\sigma^2}\right)\right)\right], \tag{A6}$$

[10] The trace molecules (like H_2O only accounts for a very tiny portion of the atmospheric composition, the rest is filled by gases like hydrogen and helium, this kind of atmosphere is also known as Primary Atmosphere (e.g. Jupiter has a Primary Atmosphere) as opposed to Secondary Atmosphere, which is principally made of heavier elements (e.g. Earth has a Secondary Atmosphere).

[11] Parts per million, 10^{-6}.

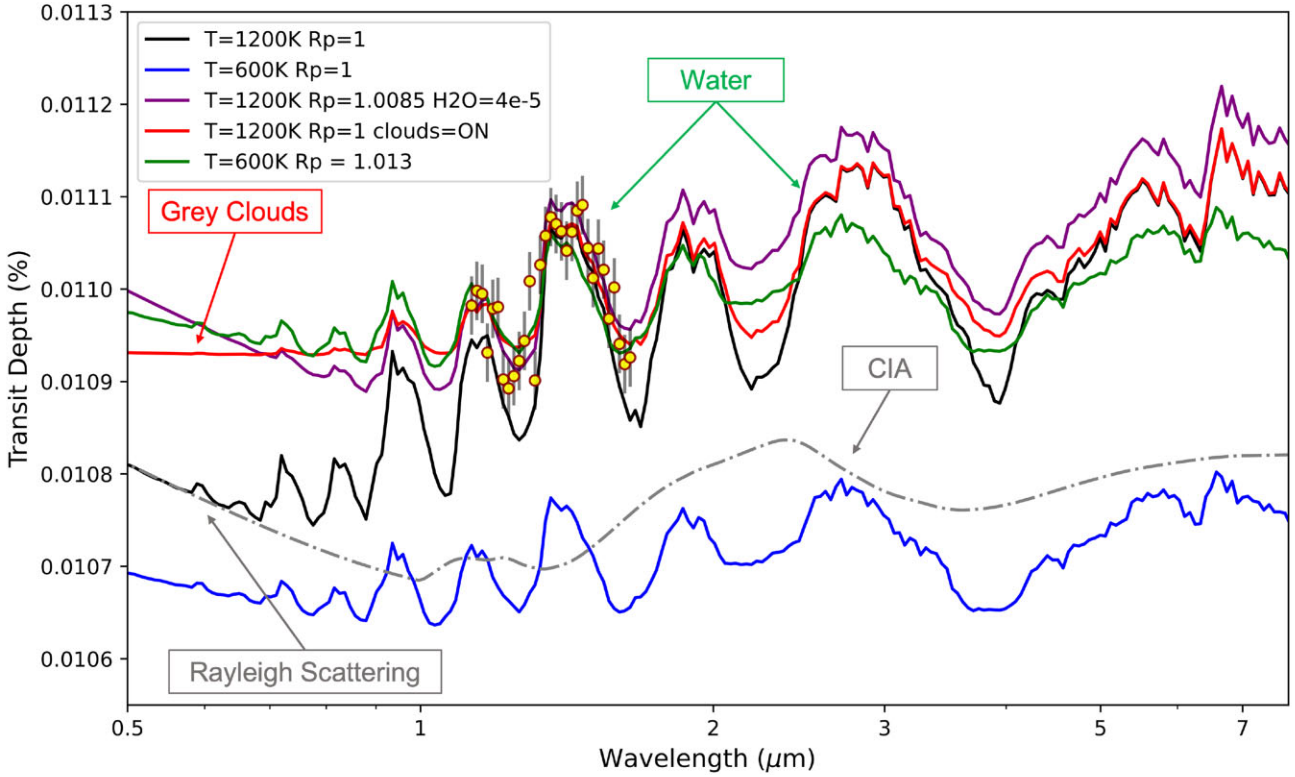

Figure A2 Spectra illustrating the sensitivity of atmospheric models to input parameters. In black: Model 1, in blue: Model 2, in purple: Model 3, in red: Model 4, and in green: Model 5. We also show in dashed grey line a model similar to Model 1 but without absorption of water, leaving only the continuum contribution of Rayleigh Scattering (short wavelengths) and CIA (long wavelengths). The red and yellow points represent a simulated observation with *Hubble Space Telescope* (*HST*) at 30 parts per million (ppm), highlighting the difficulty of constraining atmospheric properties from current data.

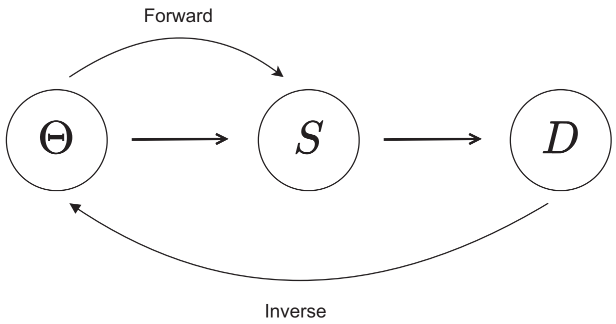

Figure A3 Schematic of a typical inverse problem setup. The forward process produces an effect (full spectrum, *S*) from a hidden cause (e.g. atmospheric parameters, Θ). However, the full effect is often unavailable to the observer due to the loss of information (such as instrument systematics, limitation in spectroscopic coverage, etc.). Instead, observers can only receive the partial effect (or otherwise known as the observation, *D*). The aim of the inverse problem is to recover the hidden cause that produces them in the first place.

where S is a simulated spectrum generated using the forward model $\mathcal{M}$ and σ represents the observation uncertainty/noise. Thanks to the forward model $\mathcal{M}$, we have an unique mapping from a set of parameters to a simulated spectrum S such that

$$S = \mathcal{M}(\Theta). \tag{A7}$$

The relation between the observed spectrum D and the simulated spectrum S is

$$D \approx S + \epsilon, \tag{A8}$$

where $\epsilon = N(0, \sigma^2)$. The approximation sign reflects the fact that the model remains an approximation of the real phenomena.

As for the prior function $P(\Theta)$, it represents our prior belief on the distribution of the random variables. With limited knowledge on the exo-atmosphere, the community always opt for an uninformative prior (also known as an uniform prior).

Unfortunately, in most cases, equation (A4) cannot be computed analytically. The main reason lies with the Bayesian Evidence, $P(D) = \int P(D, \Theta)\mathrm{d}\Theta$, the integral demands evaluation of the probability for every possible combinations, which makes the quantity intractable for any meaning cases.

A common strategy is to sample the parameter space, and use the distribution of the samples to compute the maximum likelihood estimation (MLE) and the Bayesian Evidence. There are many optimizing strategies available, including grid sampling, optimal estimation, Markov chain Monte Carlo models (MCMC), and Nested Sampling, amongst others. Those are however computationally intensive and require evaluation of millions of forward models.

There have been efforts from the ML community to develop scalable sampling algorithms. Stochastic Gradient MCMC (SG-MCMC) is a popular class of algorithms that utilizes data subsampling techniques to reduce computational time to construct the chain (Welling & Teh 2011; Ma et al. 2015; Baker et al. 2019; Nemeth & Fearnhead 2019). Stochastic Gradient Descent (SGD)'s link to approximate Bayesian Inference has prompted further investigation into its statistical properties (Chen et al. 2016; Mandt, Hoffman & Blei 2017; Xing et al. 2018); it has since been shown that SGD with constant step size (Constant-SGD) can approximate Bayesian

Posterior Distribution. Other algorithms, such as Hamiltonian Monte Carlo (HMC), incorporated information on the gradient within the proposal to improve the sampling efficiency (Neal 2011; Homan & Gelman 2014). Chen, Fox & Guestrin (2014) introduced SG-HMC, a fusion between SG-MCMC and HMC, to provide further speed up to the algorithm.

Other approaches focuses on architectural design or post-processing techniques to incorporate elements of Bayesian Inference, such as Dropout (Gal & Ghahramani 2015), Neural Network Ensembles (Lakshminarayanan, Pritzel & Blundell 2017; Cobb et al. 2019; Pearce et al. 2020), SWA-Gaussian (SWAG Maddox et al. 2019), KF-Laplace (Ritter, Botev & Barber 2018), and temperature-scaling (Guo et al. 2017).

The availability of many state-of-the-art algorithms prompts the need to benchmark their performances under different data sets and scenarios (Yao et al. 2019; Izmailov et al. 2021). Aligned with this objective, the aim of this data base and the machine learning (ML) challenge is to leverage recent developments in scalable Bayesian Inference and identify potential solutions forward.

APPENDIX B: ATMOSPHERIC TRANSMISSION MODEL IN *TauREx*

In this section, we describe the simplified transit (forward) model used in the code *TauREx 3* . The atmosphere of the planet is separated in N_L homogeneous layers following a one-dimensional plane-parallel geometry (see Fig. B1). The light rays from the host star are propagated through the atmospheric layers, being impacted by extinction processes (absorption and scattering) at the different wavelengths (λ).

The normalized differential flux (Δ_λ) or the transit depth at wavelength λ reaching the observer is simply the ratio of the surface area of the planet to the host star, which can be further simplified to the planet-to-star radius squared:

$$\Delta_\lambda = \frac{F_{\mathrm{out},\lambda} - F_{\mathrm{in},\lambda}}{F_{\mathrm{out},\lambda}} = \left(\frac{\pi R_\mathrm{p}(\lambda)}{\pi R_\mathrm{s}}\right)^2 = \left(\frac{R_\mathrm{p}(\lambda)}{R_\mathrm{s}}\right)^2, \tag{B1}$$

where $F_{\mathrm{out},\,\lambda}$ is the total flux received from the system out-of-transit, $F_{\mathrm{in},\,\lambda}$ is the total flux received in-transit, $R_\mathrm{p}(\lambda)$ is the wavelength-dependent radius that includes the atmospheric contribution, and R_s is the stellar radius. In our case, the atmospheric contribution consists in the absorption of the star light from the atmosphere (e.g. we do not include scattering processes), which follows the Beer–Lambert law.

The wavelength-dependent contribution of the atmosphere starts at the surface labelled R_0. Note that for gaseous planets (e.g. without solid surface), R_0 is a reference radius at which we consider the atmosphere is fully opaque at all wavelengths. We obtain

$$\begin{aligned} \pi R_\mathrm{p}(\lambda)^2 &= C_\mathrm{sur} + C_\mathrm{atm} \\ &= 2\pi \int_0^{R_0} r\,\mathrm{d}r + 2\pi \int_{R_0}^{\infty} r\left(1 - \mathrm{e}^{-\tau(r,\lambda)}\right) \mathrm{d}r, \end{aligned} \tag{B2}$$

where C_sur is the contribution to the planet surface, C_atm is the contribution from the atmosphere, and r is the radial coordinate. In most cases, the former term is assumed to be completely opaque and therefore can be simply evaluated as the surface area of the planet at

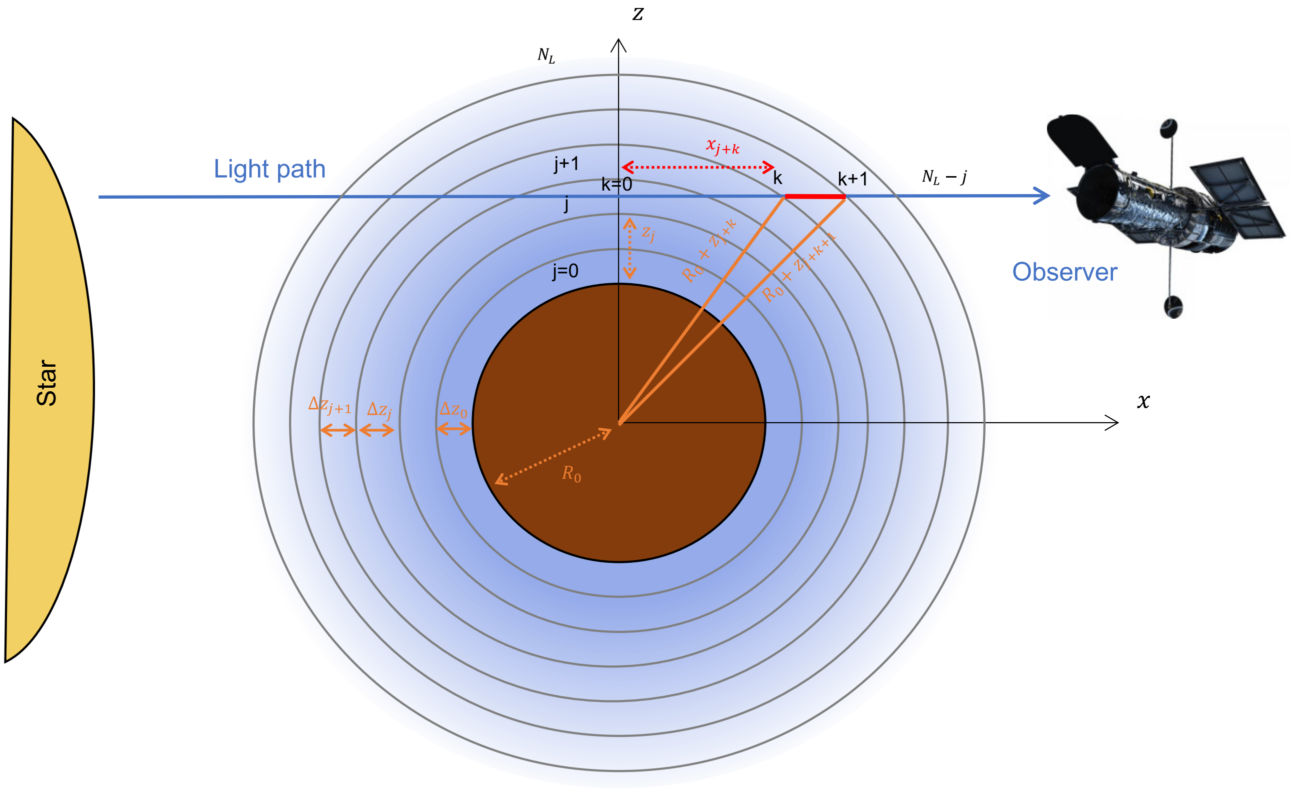

Figure B1 Illustration of the transmission of stellar radiation (left-hand side) through an exoplanet atmosphere (transit) towards an observer (right-hand side). R_0 is the reference radius at which the atmosphere becomes fully opaque. A light ray at altitude z propagates along the line-of-sight x. The atmosphere is separated in N_L layers for size Δ_z, which are labelled by the index $l = j + k$, where j refers to the z-component and k to the x-component. The discretized altitude z_l corresponds to the altitude at the lower boundary of the layer l.

radius R_0, the latter term involves the computation of the optical depth of the atmosphere at each layer, which summarizes the contribution from various processes happening within the atmosphere.

The optical depth $\tau(r,\,\lambda)$ is computed along the line of sight as follows:

$$\tau(r,\lambda) = 2\int_0^{x_\mathrm{f}} \sum_i^{N_\mathrm{G}} \chi_i(r')\rho(r')\sigma_i(r',\lambda)\mathrm{d}x. \tag{B3}$$

Here, χ_i is the mixing ratio (or abundance) of the ith species, ρ is the number density, and σ_i is the absorption cross-section of the ith species. The number of gases is noted N_G. The variable x_f is the maximum distance considered for the numerical integration.

Considering the one-dimensional geometry, the integration of τ along the x axis can be decomposed in unit elements $\tau(j,\,k)$, where j represents the y-axis indexes and k are the indexes along the x-axis. Physical quantities (e.g. the altitude z, the mixing ratio χ) defined at a layer l can then be related to the j, k indexes using $l = j + k$, and noting that k can only span the values from j to N_L. These are indexed with an additional subscript, for instance, $\chi_{i,\,l}$ is the mixing ratio of the ith species at layer l.

It follows that the unit path integral, labelled $\Delta x_{(j,\,k)}$ and identified by the red element in Fig. B1, can be expressed as

$$\Delta x_{(j,k)} = \sqrt{\left(R_0 + z_{j+k+1}\right)^2 - \left(R_0 + z_j + \frac{\Delta z_j}{2}\right)^2} - \sqrt{\left(R_0 + z_{j+k}\right)^2 - \left(R_0 + z_j + \frac{\Delta z_j}{2}\right)^2}, \tag{B4}$$

where z_l is the altitude at layer l and Δz_l are the changes in altitude at layer l.

Since the layer are equally spaced in log-pressure, we also have

$$\Delta z_l = -H_l \log\left(\frac{P_{l+1}}{P_l}\right), \tag{B5}$$

where H_l is the scaleheight at layer l and P_l is the pressure at layer l.

Expressing the optical layer element as

$$\tau_{(j,k)} = \sum_i^{N_\mathrm{G}} \chi_{i,j+k}\rho_{j+k}\sigma_{i,j+k}(\lambda)\Delta x_{(}j,k), \tag{B6}$$

one obtains the final contribution for the atmosphere as

$$C_\mathrm{atm} = 2\pi\sum_{j=0}^{N_\mathrm{L}}(R_0 + z_j)\left(1 - \exp\left[-2\sum_{k=0}^{N_\mathrm{L}-j}\tau_{(j,k)}\right]\right)\Delta z_j, \tag{B7}$$

and the transit depth as a function of wavelengths or the transmission spectrum can be computed following equation (B1). In this investigation, we produced a grid of transmission spectra through a randomized and uniform grid of free parameters.

The absorbing properties of the different molecules (H_2O, CO, CO_2, CH_4 and NH_3) and processes (Rayleigh Scattering, CIA) are encoded in the cross-sections (σ_i in equation B3). Cross-sections are temperature-, pressure-, and wavelength-dependent, and have a highly non-linear behaviour. In most codes, including the one used here, since the computation of cross-sections is a computationally intensive and complex process, they are pre-computed in tabulated files that are then interpolated to obtain the absorbing profile of the relevant molecules and processes at a given temperature, pressure, and wavelength.

For a more complete description of the employed code, we refer the reader to the original papers (Al-Refaie et al. 2021, 2022a) and the NeurIPS *TauREx* tutorial available at Zenodo: https://doi.org/10.5281/zenodo.6770103.

APPENDIX C: POSTERIOR DISTRIBUTION OF ATMOSPHERIC RETRIEVAL

Fig. C1 shows an example of posterior distribution resulting from a *TauREx* atmospheric retrieval. This posterior distribution corresponds to the data shown in Fig. 3.

The posterior distribution shows the correlation between the free parameters of the model (here atmospheric temperatures and abundances of five gases). In particular, this inverse problem is challenging for ML solutions as, due to the high level of degeneracies between the parameters of interest, the exoplanet community is interested in obtaining full probability distributions rather than a unique guess. Solutions to this inverse problem would be required to (i) correctly identify the abundances of detectable molecules (see CO_2, CH_4, and CO); (ii) characterize the correlation between parameters (see e.g. the negative correlation between temperature and abundance of CH_4); (iii) constrain upper limits for the parameters that cannot be determined (see e.g. NH_3 distribution); and (4) identify multimodal solutions (not shown in this example).

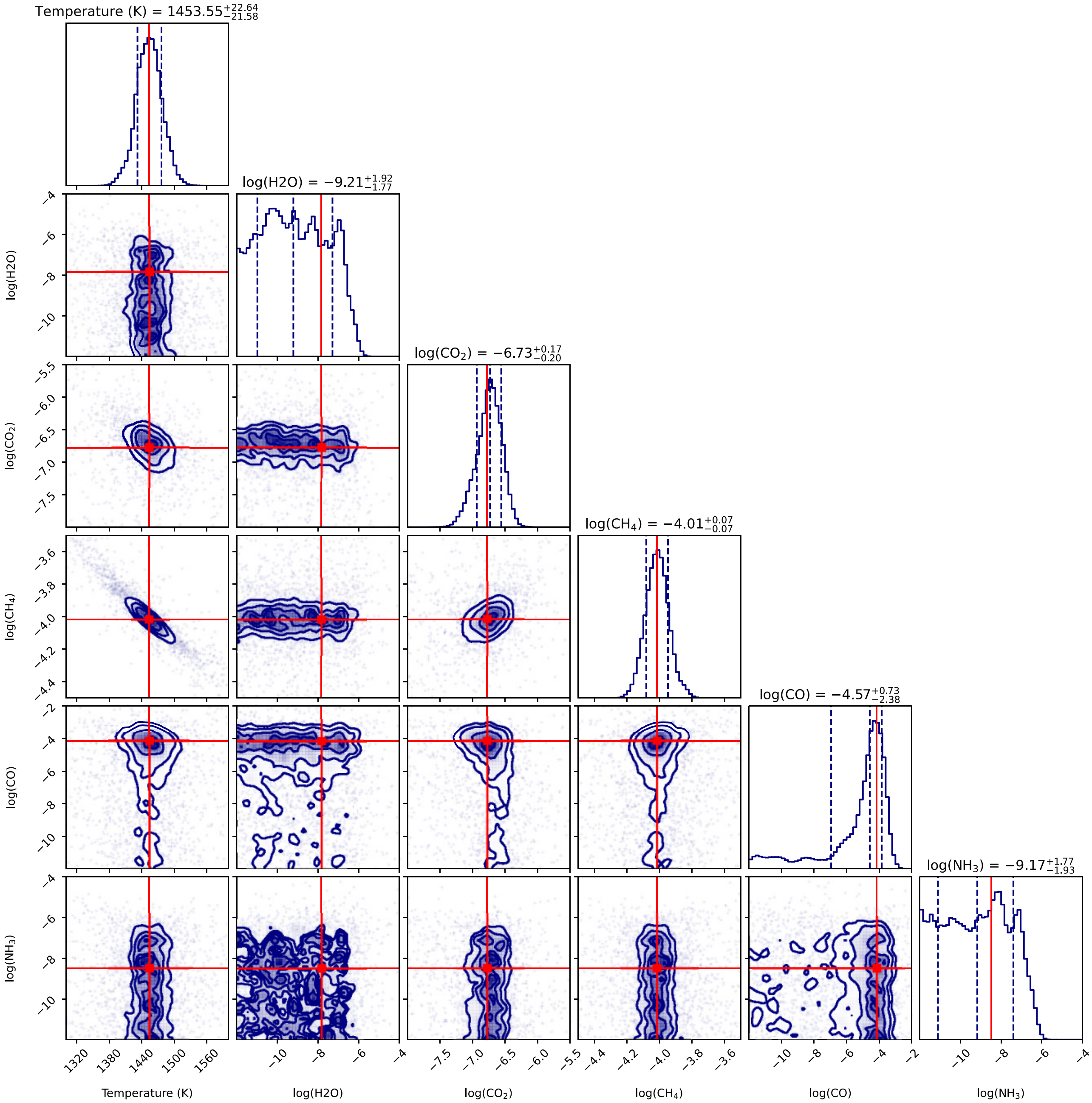

Figure C1 Example of posterior distribution obtained with *TauREx 3* on a simulated Ariel observation. This correlation map is constructed using the Nested Sampling traces and weights, with the *corner* library.

APPENDIX D: DATA OVERVIEW – CONTINUED

A strength of this large-scale data generation lies with the use of currently known demographics as the source of planetary candidates. Planet formation and evolution remains an actively researched area and there are contrasting theories as to how and why certain planets are more prevalent than others. By relying solely on observed planets we avoided producing fictitious planets that are otherwise impossible to form. Furthermore, the bi-model distribution of planet mass and radius contributes to the dichotomy seen in Fig. 4.

Due to the extremely low S/N ratio with exoplanetary observation and non-linear instrument systematics, actual observations are usually accompanied with non-negligible measurement errors. These errors are specific to the brightness of the host star, the data reduction process, the instrument onboard, and its separation from us. *ArielRad*, an official radiometric simulator dedicated to the Ariel Space Mission, is designed specifically to account for the aforementioned effects and provide realistic estimation of the observational uncertainties (Mugnai et al. 2020). Fig. D1 shows the distribution of (log-) observational uncertainties across the 52 wavelength channels. All of them displayed a non-Gaussian distribution, some even presented a bi-model distribution. There are also noticeable difference in terms of the shape and magnitude across different channels. For instance, uncertainties associated with the blue end of the spectrum tend to be smaller than the red end

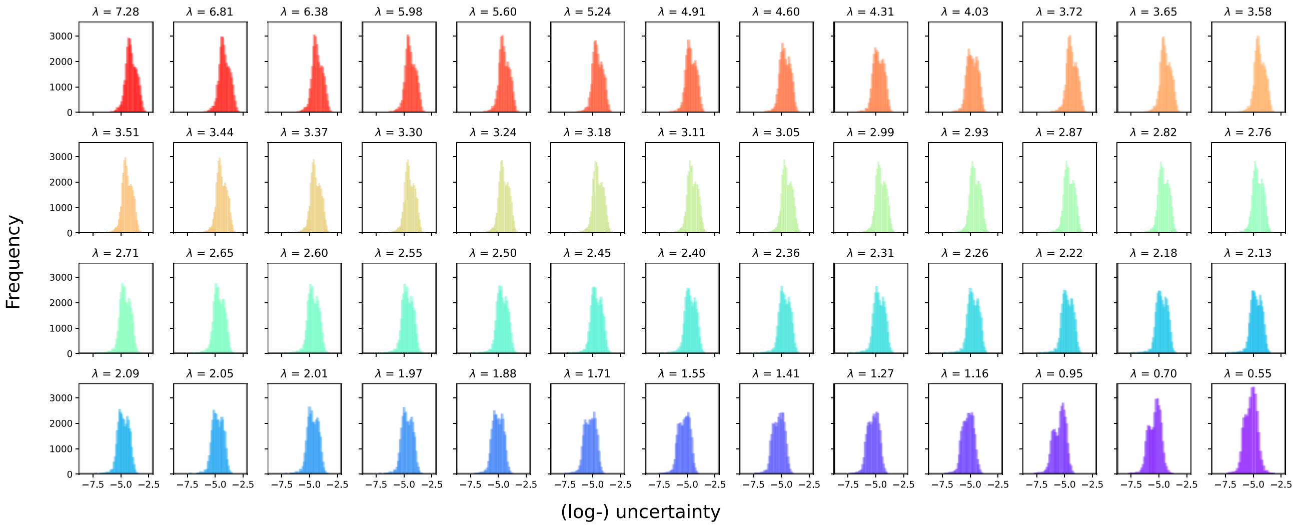

Figure D1 Distribution of (log-) uncertainty across different wavelength channels used by Ariel-Tier 2 resolution. These uncertainties are generated using ArielRad, which accounts for the different instrumentation onboard Ariel, stellar properties, as well as planetary properties. Since the SNR requirement of Ariel Tier-2 data is on the atmospheric signal, those distribution are approximately offset by one order of magnitude compared to the 'Feature height' distribution in Fig. 4.

of the spectrum, as the blue end of the spectrum is coming from a photometer.

APPENDIX E: LEVEL 2 DATA – DETAILED DESCRIPTIONS

E1 Structure

Level 2 data are designed originally for the NeurIPS 2022 competition, but the data structure can be re-used for general model training. It is consisting of a training and test set. The two sets shares the same structure and the aim is to allow better readability to non-field experts:

(i) `AuxillaryTable.csv`: contains supplementary astrophysical parameters.

(ii) `SpectralData.hdf5`: contains details of the spectroscopic observations

(iii) `Ground Truth Package`: contains the ground truth targets for the competition.

(a) `TraceData.hdf5` records the traces of the empirical distribution obtained from Nested Sampling; it is primarily used for the Regular Track.

(b) `QuartilesTable.csv` records the 16th, 50th, and 84th percentile of the posterior distribution; it is mainly used as a target for the Light Track.

(c) `FM_Parameter_Table.csv` records the model values that generate the spectra in the first place. While it could be different from the ground truth; it can be used as a soft label.

E2 Train-test split

With our long separation from any exoplanets and limitations from current technologies, it is almost impossible to ascertain the true nature of the target exo-atmosphere. In other words, *our test distribution will always be different from the training distribution, also known as domain shift in ML literature* (Wang & Deng 2018; Wilson & Cook 2020) To reflect this limitation, the train/test split is designed to uncover solutions that can maintain their performance even under unknown situations (unseen atmospheric behaviour and/or unseen planets).

Table E1. Summary of the different configurations of the four subsets.

	Subset 1	Subset 2	Subset 3	Subset 4
Planetary configuration	In-Range	Out-Range	In-Range	Out-Range
Atmospheric properties	In-Range	In-Range	Out-Range	Out-Range

To support this goal, we abandoned the usual practice of dividing a data set randomly into training and test set, which tests the model's ability to generalize under a *homogeneous* distribution. Instead, the test set is designed to contain In-Training Parameter Ranges (In-Range) and Out-of-Training Range Parameters (Out-Range) components. In-Range samples represent examples that come from the same distribution as the training data. Out-Range represent samples that are unseen by the model during training, this includes unseen planetary and atmospheric properties.

As a result, some of the planets are purposely removed from the training set to create *unseen* planetary properties, any theoretical spectra created using from those planets are also taken away from the training, causing a slight drop in the amount of available training data, as compared to Level 1 data. We further generated 5461 spectra under equilibrium chemistry scheme (Agúndez et al. 2012, 2020) by assuming solar elemental ratios to create *unseen* atmospheric properties. As stated above, these spectra are unseen and thus are *not* included in the training set.

The test set is stratified into four subsets, each representing varying degree of similarity to the training data. Table E1 summarizes the different configurations of the four subsets. Subset 1 is the most similar to the training set as all the components are In-Range, while Subset 4 is the most dissimilar as all the components are Out-Range. All of them contain roughly the same proportion (∼25 per cent)in the test set.

All examples, regardless of their initial atmospheric assumptions or planetary properties, are homogeneously retrieved using the free chemistry settings outlined in Section 2.3. By doing so, our retrievals will be, on purpose, biased and will not be retrieving the input chemistry (ground truth). Participants are tasked to reproduce the same results from our biased retrievals.

The combined effect of these two changes means that any proposed solution will have to maintain reliable and consistent behaviour when exposed to distributions that are unknown and unseen to their training distribution. We explicitly did not include any spectra generated with equilibrium chemistry assumption in the training set, as a proxy of the actual situation – our atmospheric models cannot adequately describe the actual atmosphere.

The stratification of test examples provides flexibility to future investigation. The test set can be used to test the trained model under different testing conditions, for instance, one can test their models on set 1 examples to understand the model's performance under homogeneous cases. In any cases, spectra generated with either free chemistry and/or equilibrium chemistry is available online for any interested parties to construct their own training and test set.

The data set for NeurIPS 2022 competition contains a similar data structure, but featured more diverse, unseen atmospheric assumptions than the one presented here. A discussion of these atmospheric assumption is outside the scope of this paper, and the readers are advised to refer to Yip et al. (2022b) for a more detailed description of the respective test set for the data challenge.

This paper has been typeset from a TEX/LATEX file prepared by the author.