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

Quentin Changeat,1,* and Kai Hou Yip1†
1Department of Physics and Astronomy, Gower St., London WC1E 6BT, United Kingdom

ABSTRACT

This is an exciting era for exo-planetary exploration. In the past two decades, astronomers have harvested data from all the observatories at their disposal. Those collective efforts allowed us to have a glimpse at the convoluted process of planetary formation and evolution and its connections to atmospheric compositions, but nevertheless remained limited by the low quality and scarcity of exo-atmospheric data. Now, the recently launched JWST, and other upcoming space missions such as Ariel, Twinkle and ELTs are set to a new the landscape, bringing fresh insights to these remote worlds. 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 interdisciplinary 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. As part of the data product of the NeurIPS 2022 Data challenge, we would like to present the Ariel Big Challenge Database (ABC Database), a carefully designed, organised and publicly available database. With 105,887 forward models, 26,109 complementary posterior distributions and an easy-to-understand documentation, this represents an unprecedented effort to invite cross-disciplinary experts to the study of the inverse problem in the context of exoplanetary studies.

Key words: exoplanet atmosphere – telescope data – inverse problem – machine learning

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 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 days. 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). 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 et al. 2008; Kreidberg et al. 2014; Schwarz et al. 2015; Sing et al. 2016; Stevenson et al. 2017; Hoeijmakers et al. 2018; de Wit 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. 2021; Changeat et al. 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. Many discovery 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 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

* Domain expert: quentin.changeat.18@ucl.ac.uk
† ML expert: kai.yip.13@ucl.ac.uk

© 2022 The Authors
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. 2015b,a; Lavie et al. 2017; Gandhi & Madhusudhan 2018; Mollière et al. 2019; Zhang et al. 2019; Min et al. 2020; Al-Refaie et al. 2021b; 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. 2021). 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 typically requires between $10^5$ and $10^8$ forward model calls to reach convergence, meaning that only models providing spectra in times 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. 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; Changeat et al. 2019; Caldas et al. 2019; Yip et al. 2020; Taylor et al. 2020, 2021; Changeat et al. 2021; Al-Refaie et al. 2021a; Yip et al. 2022).

In recent years, the community started to explore alternative approaches to circumvent the bottleneck with sampling based approaches. Machine Learning 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, where he trained 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 (Zingales & Waldmann 2018; Soboczenski et al. 2018; Cobb et al. 2019; Hayes et al. 2020; Oreshenko et al. 2020; Nixon & Madhusudhan 2020; Himes et al. 2022; Ardevol Martinez et al. 2022; Haldemann et al. 2022; Yip et al. 2022). 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. 2021).

Organised as part of the Conference on Neural Information Processing Systems (NeurIPS) Competition Track, the 2022 Ariel Data Challenge aims to highlight the above issue to external communities and encourage novel, cross-disciplinary solutions to the problem. For this Data Challenge, we constructed the open-access Ariel Big Challenge (ABC) Database of atmospheric forward models and retrievals. The database is made permanently accessible at the following link: https://doi.org/10.5281/zenodo.6779103.

Since the creation of similar database constitutes a major barrier to anyone interested in applying Machine Learning in the domain of exoplanet atmospheres, we emphasise on its release as a community asset. The organisation and creation of this dataset poses a challenge on its own because:

(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 machine learning expertise to ensure the data product is representative of the problem at hand, and ideally, one that adequate 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 2,000,000 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 database and 3.) To provide a carefully curated, well-organised, and scientifically relevant dataset for any research community. This manuscript complements the Data Challenge proposal description (Yip et al. 2022) accepted as a NeurIPS 2022 data challenge.

2 INTRODUCTION TO ATMOSPHERIC STUDIES OF EXOPLANETS.

2.1 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, which 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 Figure 1. Transit events are periodic, so they can easily be distinguished from astrophysical sources of noise (stellar variations\(^1\), instrument systematics, 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 ($\Delta$). The transit depth is the normalised difference between the flux received from the star when the planet is out-of-transit ($F_{\text{out}}$) and when the planet is in-transit ($F_{\text{in}}$).

\[
\Delta = \frac{F_{\text{out}} - F_{\text{in}}}{F_{\text{out}}} = \left( \frac{R_p}{R_s} \right)^2
\]

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, we can replace the planetary radius $R_p$ by $R_p + h$, where $h$ is the

\(^1\) for example, stars’ brightness could vary from time to time.
chemistry of the atmosphere plays a part in the scaling factor $g$. On the opposite, if the temperature increases the atmosphere contracts under gravity. This makes sense and it is because the atmosphere contracts under gravity.

The scale height is defined by:

$$H = \frac{k_B T}{\mu g}$$

where $k_B$ is the Boltzmann constant, $T$ is the temperature, $\mu$ 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 behaviour of atmospheric composition, the rest is filled by gases like Hydrogen and Helium, in standard solar ratios (H/O = 2). On the opposite, 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\(^2\).

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.

2.2 Modelling exoplanet atmospheres

Observing exoplanetary transits at various wavelengths, meaning obtaining $\Delta$ as a function of $\lambda$, 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, we add a spectrum to the 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$_2$O) 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)\(^3\). On top of the molecular absorption from water vapour, we also consider three additional absorption processes: Collision Induced Absorption (CIA), Rayleigh Scattering and Grey Clouds (not considered in this version of the ABC database).

Equipped with this model, we set the following cases for which the spectra are available in Figure 2:

- **Case 1 (black):** Planet Radius $R_p = 1.0$, Temperature $T$ = 1200K, Mixing ratio of H$_2$O = $10^{-3}$ and No clouds.

\(^2\) we cannot observe any non-opaque (transparent) part of the atmosphere.

\(^3\) The trace molecules (like H$_2$O 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 (Jupiter has a Primary Atmospheres) as opposed to Secondary Atmospheres, which are principally made of heavier elements (Earth has a Secondary Atmosphere).
• Case 2 (blue): Same as Case 1 but the temperature is decrease to \( T = 500 \text{K} \).

• Case 3 (purple): Same as Case 1 but the water content is decrease to \( \text{H}_2\text{O} = 4.10^{-5} \) while the planetary radius is increased to \( R_P = 1.0085 \).

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

• Case 5 (green): Same as Case 2 but with an increased planetary radius to \( R_P = 1.013 \).

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\(\mu\text{m} \) to 2\(\mu\text{m} \) 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\(\mu\text{m} \) to 2\(\mu\text{m} \) range. Finally, Case 5 has a lower temperature (600K) 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\(\mu\text{m} \) to 1.6\(\mu\text{m} \) and reaches errors of the order 30ppm\(^4\). 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.

### 2.3 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 in information often means that the inverse mapping function \( M^{-1} \) is unknown and no longer 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 Figure 3 for a typical setup of an inverse problem.

Our goal is to estimate the set of parameters \( \Theta \) that best explain the observed spectrum \( D \) under a given atmospheric model \( M \). 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); Buchner (2021) for more information.

To find the conditional distribution of the model parameters, given the observation, also known as the posterior distribution \( P(\Theta|D, M) \) in Bayes’ Theorem (Bayes & Price 1763). The posterior distribution can be computed via the following formulation:

\[
P(\Theta|D, M) = \frac{P(D|\Theta, M)P(\Theta)}{P(D)}
\]

where \( P(D|\Theta, M) \) represents the likelihood function under a given model, \( P(\Theta) \) represents the prior and \( P(D) \) represents the normalising 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 \( M \), i.e.

\[
\mathbb{E}[\log P(\Theta|D, M)] = \mathbb{E}[\log(P(D|\Theta, M)P(\Theta))]
\]

\[
= \mathbb{E}[\log \left( \frac{1}{\sqrt{2\pi\epsilon^2}} \exp \left( -\frac{1}{2} \left( \frac{D - S}{\epsilon^2} \right)^2 \right) \right)]
\]

where \( S \) is our simulated spectrum and \( \epsilon \) represents the observation uncertainty/noise. Thanks to the forward model \( M \), we have an unique mapping from a set of parameters to a simulated spectrum \( S \) such that:

\[
S = M(\Theta).
\]

The relation between the observed spectrum \( D \) and the simulated spectrum \( S \) is:

\[
D = S + \epsilon.
\]

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 uniform prior).

Unfortunately, in most cases Equation 4 cannot be computed analytically. The main reason lies with the Bayesian Evidence, \( P(D) = \int P(D|\Theta)d\Theta \), 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 samples to estimate the maximum likelihood and the Bayesian evidence. There are many optimizing strategies available, these includes 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 utilises data sub-sampling

\(^4\) parts per million, \( 10^{-6} \)
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 (Mandt et al. 2017; Chen et al. 2016; 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 et al. (2014) introduce SG-HMC, a fusion between SG-MCMC and HMC, to provide further speed up to the algorithm.

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

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

3 DATA GENERATION

For the Conference on Neural Information Processing Systems 2022 (NeurIPS), the Ariel Data Challenge on Inferring Physical Properties of Exoplanets From Next-Generation Telescopes was accepted. The goal of this competition is to identifying a reliable and scalable method to perform planetary characterisation. For the data generation we employed Alfnoor (Changeat et al. 2021), a tool built to expand the forward model and atmospheric retrieval capabilities of TauREx3 (Al-Refaie et al. 2021b) to large populations of exo-atmospheres. Alfnoor allows to automate the generation or telescope simulations and perform large scale standardised atmospheric retrievals. In the context of ESA-Ariel, we generated 105,887 simulated forward observations as well as 26,109 standardised retrieval outputs.
3.1 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 stellar radius ($R_*$), distance to Earth (d), star magnitude K (Kmag), planetary radius ($R_p$), planetary mass ($M_p$), planet equilibrium temperature ($T$) and transit duration ($t_{14}$).

The objects in this database 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 the 1st of March 2022 for this database. 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 of their inferred properties make them compelling objects. Follow-up observations will allow to classify their nature, but for the purpose of building this database, they are as close as possible to what the reality looks like. To those lists of objects, we filtered all the planets with radius below 1.5 $R_e$. 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 2,972 confirmed exoplanets and 2,928 candidate exoplanets, thus bringing our total to 5,900 unique objects.

3.2 The atmospheric forward model setup

To generate simulated telescope observations, we employed Afnoor. We produce batches of randomised observations for the population described in the previous section. For each planet the stellar parameters ($R_*$, d, Kmag), 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 isothermal (constant temperature) at the equilibrium temperature of the planet, and we simulate the planet’s atmosphere from 10 bar 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$_2$O (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 et al. 2019), selected based on our current understanding of exoplanetary chemistry (Agúndez et al. 2012; Venot & Agúndez 2015; Madhusudhan et al. 2016; Drummond et al. 2016; Woitke et al. 2018; Stock et al. 2018; Venot et al. 2020; Al-Refaie et al. 2021a; Baeyens et al. 2022). 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. However, due to the differences in strength of spectral features and the capabilities of ESA-Ariel (e.g. its wavelength coverage), we select different bounds for the randomised chemical abundances. The bounds employed for this dataset are:

(i) H$_2$O: RandomLogUniform(min=-9, max=-3).
(ii) CO: RandomLogUniform(min=-6, max=-3).
(iii) CO$_2$: RandomLogUniform(min=-9, max=-4).
(iv) CH$_4$: RandomLogUniform(min=-9, max=-3).
(v) NH$_3$: RandomLogUniform(min=-9, max=-4).

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

Each spectrum is first computed at high-resolution$^5$, 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. 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 and that the signal-to-noise ratio (SNR) of the observations must be higher than 7. Since our chosen sample of planet is made from real objects and that all are not favourable targets for Ariel, this means that some targets require an unrealistic number of observations to reach the SNR condition of Tier 2. However, this does not affect the purpose of this dataset, providing independent instances of realistic noise profiles.

Following those steps, we obtain a realistic Ariel simulated observation for each planet and each randomised chemistry. We show an example of such simulated observation in Figure 4. In total, we produced 105,887 simulated observations for the ABC Database.

3.3 The atmospheric retrieval setup

For 26,109 (25%) of the simulated observations generated at the previous step, we perform the traditional inversion technique using Afnoor.

For the model to optimise, 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$_2$O, CO$_2$, CH$_4$, CO and NH$_3$. The priors are made wide and uninformative, with the atmospheric temperature being fitted between 100K and 5500K and the chemical abundances between 10^{-12} and 10^{-1} in Volume Mixing Ratios. The widely used Nested Sampling Optimizer, MultiNest (Feroz et al. 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-fit spectrum in Figure 4. 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, Figure B1. 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.

3.4 Data Overview

Following the data generation process outlined above, we have generated a total of 105,887 forward models in Ariel Tier-2 resolution. 26% of them are complemented with results from atmospheric retrieval (following a generic setting as described in Section 3.3). Figure 5 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 dataset. The characteristic dichotomy stemmed from current demo-

$^5$ Spectra have to be computed at high-resolution (R) since instrumental binning is done on the received photons, e.g. the recorded transit depth $\Delta$. In our case, we used $R = 4^4 = 10,000$. 

RASTI 000. 1–18 (2022)
Figure 4. Example of a simulated Ariel observation with errorbars (datapoints) for a randomised chemistry. The best-fit 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 CH₄. The datapoints around 4.5 μm are associated with CO and CO₂ absorption. Note the difference in wavelength coverage (0.5 μm to 7.8 μm) as compare to the HST spectrum (1.1 μm to 1.7 μm) in Figure 2, which allows us to extract precise information for many molecules.

Figure 5. Distribution of mean transit depth (red) overlapped with the distribution of the feature height (orange), both measured in 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 dataset. On the other hand, the feature height documents the “strength” of absorption feature in each spectrum. Any successful model must be able to account for the variations in both scales, graphics studies⁶ and selection bias in our observation technique.⁷ 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 molecular features presented in the spectra, e.g. the peaks and troughs as seen in Figure 2 and Figure 4. 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 between different spectra, it also needs to take into account the intra-variation across wavelength channels, which is always 1–3 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. Figure 6 compares the retrieval results against the input values of the 6 targets of interest (H₂O, CO₂, CH₄, CO, NH₃, Temperature). Each data point in every subplot represents a single spectrum and is colored in accordance to the size of the inter-quartile range (IQR).⁸ 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. Chang 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 error-bar obtained. Here, we find similar results, and the detection limits of Ariel correspond to the region where all the retrieved values from Figure 6 deviate from the diagonal line (associated with colors from green to red).

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

3.5 Structure of the ABC Database

The database contains 2 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 below:

3.5.1 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

⁶ Latest studies show that Super-Earth sized planets are prevalent while there is a deficiency in the population of sub-Neptunes.
⁷ Transit technique tends to favour larger planets.
⁸ Here we define IQR as the difference between the 84th and the 16th percentile.
Figure 6. 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 is 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.

is 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 $1 \times 10^{-8}$ in any wavelength bins.
(iv) Standardised units and data formats.
(v) Extracted all Stellar, Planetary and Instrumental metadata.
(vi) Combined all instances into a single, unified file.

Level 1 data is organised into all_data.csv, observations.hdf5 and all_target.hdf5. all_data.csv contains information on the planetary system and the input values for 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% of them (26,109) has complementary retrievals from Nested Sampling.

3.5.2 Level 2: Data for NeurIPS 2022 competition.

The following section is specific to the NeurIPS 2022 competition. In order to allow for the broadest possible participation and minimise the overhead for non-field experts, we pre-processed the dataset with our domain knowledge so that the end product is ready for model development. At the same time, we have kept as much auxiliary information as possible to allow a diverse array of solutions. Here we outlined the list of operations we performed:

(i) Removed data with less than 1500 points in the tracedata. This is to allow more accurate metric computation.
(ii) Removed un-informative and duplicated astrophysical or instrumental features$^9$.
(iii) Split data into training and test sets.
(iv) Re-organised the data to comply with competition format.

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 has 2,997 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. We have devoted a section in Appendix D to describe the Level 2 data in details.

9 including star_magnitudeK, star_metallicity, star_type, planet_type, star_mass_kg, star_radius_m, planet_albedo, planet_impact_param, planet_mass_kg, planet_radius_m, planet_transit_time, instrument_nobs.
3.6 Additional resources
Published along with the database, we provide a series of complementary resources. In particular the database is provided with a Jupyter Notebook describing the data structure, how to load the dataset, and demonstrating its main characteristics. We also include a dedicated TauREx3 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 database.

4 CONCLUSIONS
To support the 3rd installment of the Ariel Data Challenge, accepted as part of the NeurIPS Conference10, we present here the publicly available ABC Database (https://doi.org/10.5281/zenodo.6770103). This paper introduces, for a non expert community, the basic physical and chemical processes involved in the creation of such database, describes the utilised tools11, and clearly states 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. With this effort, 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 Union’s Horizon 2020 research and innovation programme (grant agreement No 758892, ExoAI), from the Science and Technology Funding Council grants ST/S002634/1 and ST/T001836/1 and from the UK Space Agency grant ST/W00254X/1. The author thanks Ingo P. Waldmann, Giovanna Tinetti and Ahmed F. Al-Refaie for their useful recommendations and discussions.

This work utilised 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 article are available as a Zenodo Digital Repository, at https://doi.org/10.5281/zenodo.6770103.

REFERENCES
Agúndez, M., Venot, O., Iro, N., Selsis, F., Hersant, F., Hébrard, E., & Dobrijevic, M., 2012. The impact of atmospheric circulation on the chemistry of the hot Jupiter HD 209458b, A&A, 548, A73.

Agúndez, M., Martínez, J. I., de Andres, P. L., Cernicharo, J., & Martín-Gago, J. A., 2020. Chemical equilibrium in AGB atmospheres: successes, failures, and prospects for small molecules, clusters, and condensates, A&A, 637, A59.

Al-Refaie, A. F., Changeat, Q., Venot, O., Waldmann, I. P., & Tinetti, G., 2021a. A comparison of chemical models of exoplanet atmospheres enabled by TauREx 3.1, arxiv e-prints, p. arXiv:2110.01271.

Al-Refaie, A. F., Changeat, Q., Waldmann, I. P., & Tinetti, G., 2021b. TauREx 3: A Fast, Dynamic, and Extendable Framework for Retrievals, ApJ, 917(1), 37.

Ardevel Martínez, F., Min, M., Kamp, I., & Palmer, P. I., 2022. Convolutional neural networks as an alternative to Bayesian retrievals, arXiv e-prints, p. arXiv:2203.01236.

Baeyens, R., Konings, T., Venot, O., Carone, L., & Decin, L., 2022. Grid of pseudo-2D chemistry models for tidally locked exoplanets - II. The role of photochemistry, MNRAS, 512(4), 4877–4892.

Baker, J., Fearnhead, P., Fox, E. B., & Nemeth, C., 2019. Control variates for stochastic gradient MCMC, Statistics and Computing, 29(3), 599–615.

Batalha, N. M., 2014. Exploring exoplanet populations with NASA’s Kepler Mission, Proceedings of the National Academy of Science, 111(35), 12647–12654.

Bayes, T. & Price, n., 1763. Li. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. Philosophical Transactions of the Royal Society of London, 53, 370–418.

Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., Caldwell, J., Christensen-Dalsgaard, J., Cochran, W. D., DeVore, E., Dunham, E. W., Dupree, A. K., Gautier, T. N., Geary, J. C., Gilliland, R., Gould, A., Howell, S. B., Jenkins, J. M., Kondo, Y., Latham, D. W., Marcy, G. W., Meibom, S., Kjeldsen, H., Lissauer, J. J., Monet, D. G., Morrison, D., Sasselov, D., Tarter, J., Boss, A., Brownlee, D., Owen, T., Burzi, D., Charbonneau, D., Doyle, L., Fortney, J., Ford, E. B., Holman, M. J., Seager, S., Steffen, J. H., Welsh, W. F., Rowe, J., Anderson, H., Buchhave, L., Ciardi, D., Walkowicz, L., Sherry, W., Horch, E., Isaacson, H., Everett, M. E., Fischer, D., Torres, G., Johnson, J. A., Endl, M., MacQueen, P., Bryson, S. T., Dotson, J., Haas, M., Kolodziejczak, J., Van Cleve, J., Chandrasekaran, H., Twicken, J. D., Quintana, E. V., Clarke, B. D., Allen, C., Li, J., Wu, H., Tenenbaum, P., Verner, E., Bruhweiler, F., Barnes, J., & Prsa, A., 2010. Kepler Planet-Detection Mission: Introduction and First Results, Science, 327(5968), 977.

Brogi, M. & Line, M. R., 2019. Retrieving Temperatures and Abundances of Exoplanet Atmospheres with High-resolution Cross-correlation Spectroscopy, AJ, 157(3), 114.

Buchner, J., 2021. Nested Sampling Methods, arxiv e-prints, p. arXiv:2101.09675.

Caldas, A., Leconte, J., Selsis, F., Waldmann, I. P., Bordé, P., Rocchetto, M., & Charnay, B., 2019. Effects of a fully 3D atmospheric structure on exoplanet transmission spectra: retrieval biases due to day-night temperature gradients, A&A, 623, A161.

Cassan, A., Kubas, D., Beaulieu, J. P., Dominik, M., Horne, K., Greenhill, J., Wambsganss, J., Menzies, J., Williams, A., Jørgensen, U. G., Udalski, A., Bennett, D. P., Alborn, M. D., Batista, V., Brillant, S., Caldwell, J. A. R., Cole, A., Coutures, C., Cook, K. H., Dieters, S., Dominis Prester, D., Donatowicz, J., Fouquè, P., Hill, K., Kains, N., Kane, S., Marquette, J. B., Martin, R., Pollard, K. R., Sahu, K. C., Venter, C., Warren, D., Watson, B., Zub, M., Sumi, T., Szymański, M. K., Kubia, M., Polišenský, R., Sozysński, I., Ulaczyk, K., Pietrzyński, G., & Wyrzykowski, Ł., 2012. One or more bound planets per Milky Way star from microlensing observations, Nature, 481(7380), 167–169.

Changeat, Q., Edwards, B., 2021. The Hubble WFC3 Emission Spectrum of the Extremely Hot Jupiter KELT-9b, ApJ, 907(1), L22.

Changeat, Q., Edwards, B., Waldmann, I. P., & Tinetti, G., 2019. Toward a More Complex Description of Chemical Profiles in Exoplanet Retrievals: A Two-layer Parameterization, ApJ, 886(1), 39.

Changeat, Q., Al-Refaie, A., Mugnai, L. V., Edwards, B., Waldmann, I. P., Pascale, E., & Tinetti, G., 2020a. Almro: A Retrieval Simulation of the Ariel Target List, AJ, 160(2), 80.

Changeat, Q., Edwards, B., Al-Refaie, A. F., Morvan, M., Tsiaras, A., Wald-
Zhang, M., Chachan, Y., Kempton, E. M. R., & Knutson, H. A., 2019. Forward Modeling and Retrievals with PLATON, a Fast Open-source Tool, *PASP*, **131**(997), 034501.

Zingales, T. & Waldmann, I. P., 2018. ExoGAN: Retrieving Exoplanetary Atmospheres Using Deep Convolutional Generative Adversarial Networks, *AJ*, **156**(6), 268.
APPENDIX A: ATMOSPHERIC TRANSMISSION MODEL IN TauREX

For this exercise, we describe the simplified transit model used in the code TauREX 3. The atmosphere is separated in $N_L$ homogeneous layers following a one-dimensional plane-parallel geometry (see Figure A1). 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 ($\lambda$). The normalised differential flux reaching the observer is:

$$\Delta \equiv \frac{F_{\text{out}} - F_{\text{in}}}{F_{\text{out}}} = \left( \frac{R_p(\lambda)}{R_s} \right)^2,$$

where $R_p(\lambda)$ is the wavelength-dependent radius which 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 get:

$$\pi R_p(\lambda)^2 = C_{\text{sur}} + C_{\text{atm}} = 2\pi \int_0^{R_0} r dr + 2\pi \int_{R_0}^{\infty} r \left( 1 - e^{-\tau(r,\lambda)} \right) dr,$$

where $C_{\text{sur}}$ is the contribution to the planet surface, $C_{\text{atm}}$ is the contribution from the atmosphere and $r$ is the radial coordinate.

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

$$\tau(r, \lambda) = 2 \int_0^{x_f} \frac{N_G}{j} \chi_i(r') \rho(r') \sigma_i(r', \lambda) dx.$$

(A3)

Here, $\chi_i$ is the mixing ratio (or abundance) of the $i^{th}$ species, $\rho$ the number density, and $\sigma_i$ the absorption cross-section of the $i^{th}$ 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 $\tau$ 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 $z_j$, the mixing ratio $\chi$) 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 $i^{th}$ species at layer $l$.

It follows that the unit path integral, labelled $\Delta \chi_{i,j,k}$ and identified by the red element in Figure A1, can be expressed as:

$$\Delta \chi_{i,j,k} = \sqrt{\left( R_0 + z_k \right)^2 - \left( R_0 + z_j + \frac{\Delta z_l}{2} \right)^2} - \sqrt{\left( R_0 + z_k \right)^2 - \left( R_0 + z_j + \frac{\Delta z_l}{2} \right)^2},$$

where $z_l$ is the altitude at layer $l$ and $\Delta z_l$ is the changes in altitude at layer $l$.

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

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

(A5)

where $H_l$ is the scale height at layer $l$ and $P_l$ is the pressure at layer $l$.

Expressing the optical layer element as

$$\tau(j,k) = \sum_{i} x_{i,j+k} \rho \sigma_i (\lambda) \Delta \chi_{i,j,k},$$

(A6)

one gets the final contribution for the atmosphere as:

$$C_{\text{atm}} = 2\pi \sum_{j=0}^{N_G} \left( R_0 + z_j \right) \left( 1 - \exp \left( -2 \sum_{k=0}^{N_L-j} \tau(j,k) \right) \right) \Delta z_j,$$

(A7)

and the transit depth as a function of wavelengths can be computed.

The absorbing properties of the different molecules ($\text{H}_2\text{O}$, CO, CO$_2$, CH$_4$ and NH$_3$) and processes (Rayleigh Scattering, CIA) are encoded in the cross-sections ($\sigma_i$ in Equation A2). 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, which are then interpolated to obtain the absorbing profile of the relevant molecules and processes at a given temperature, pressure and wavelength.

APPENDIX B: POSTERIOR DISTRIBUTION OF ATMOSPHERIC RETRIEVAL

Figure B1 shows an example of posterior distribution resulting from a TauREX atmospheric retrieval. This posterior distribution corresponds to the data shown in Figure 4.

APPENDIX C: 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. Figure C1 shows the demographics of nine selected stellar and planetary parameters used to generate the population of forward model. However, relying on currently known planets is a double edged knife. While it saved us from making unverified assumptions, our data is prone to selection bias stemmed from the observation technique, strategy and instrument specification. These biases can be easily spot from Figure C1. For instance, the distribution of orbital period tends to be shorter (peaks around ~3 days) as their proximity to the host star makes them easier to discover. Furthermore, the bi-modal distribution of planet mass and radius contributes to the dichotomy seen in Figure 5.

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.
\[ j=0 \quad \text{and} \quad k=0 \]
\[ j \quad j+1 \]
\[ N \quad ! \quad k \quad k+1 \]
\[ R \quad ! \quad + \quad z \quad " \quad # \quad $ \quad % \quad R \quad ! \quad + \quad z \quad " \quad # \quad $ \quad \Delta z \quad ! \quad \Delta z \quad " \quad \Delta z \quad # \quad % \quad R \quad ! \quad \text{Star} \quad x \quad \text{Light path} \quad z \quad \text{observer} \quad x \quad ! \quad z \quad " \quad # \quad$ \quad N \quad ! \quad - \quad j \quad Observer \]

**Figure A1.** Illustration of the transmission of stellar radiation (left side) through an exoplanet atmosphere (transit) towards an observer (right 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 \( \Delta z \), which are labelled by the index \( l = j + k \), where \( j \) refers to the \( z \) component and \( k \) to the \( x \) component. The discretised altitude \( z_l \) corresponds to the altitude at the lower boundary of the layer \( l \).

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). Figure C2 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 differences in terms of the shape and magnitude across different channels. For instance, uncertainties associated with the blue end of spectrum tend to be smaller than the red end of the spectrum.

### APPENDIX D: LEVEL 2 DATA - DETAILED DESCRIPTIONS

#### D1 Structure

Level 2 data is designed specifically for NeurIPS 2022 competition. It is consisted of a training and test set. The two sets share 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**: The 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 generates the spectra in the first place. While it could be different from the ground truth, it can be used as a soft label.

#### D2 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 order words, our test distribution will always be different from the training distribution, also known as domain shift in Machine Learning literature (Wang & Deng 2018; Wilson & Cook 2020) To reflect this limitation, the goal of this year’s competition is to uncover solutions that can perform atmospheric retrieval even under unknown situations (unseen atmospheric behaviour and/or unseen planets).

To support this goal, we abandoned the usual practise of dividing a dataset randomly into training and test set, which tests the model’s
ability to generalise 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 came from the same distribution as the training data. Out-Range represents 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 hidden from the training set to create unseen planetary properties, hence any Forward Model produced from those planets are taken away from the training, causing a slight drop in the amount of available training data. We further generated 5461 spectra under equilibrium chemistry scheme\textsuperscript{12} (Agúndez et al. 2012, 2020) to create unseen atmospheric properties. These spectra are not included in the training set.

Unlike the training set, all test examples will be retrieved using the free chemistry settings outlined in Section 3.3. By doing so

\textsuperscript{12} The full settings of the equilibrium chemistry scheme will be revealed at a later date when the competition has ended.

---

\textbf{Figure B1.} Example of posterior distribution obtained with \textit{TauREx3} on a simulated Ariel observation. This correlation map is constructed using the Nested Sampling traces and weights, with the \textit{corner} library.

Temperature (K) = 1453.55 ± 22.64

log(H\textsubscript{2}O) = 9.21 ± 1.92

log(CH\textsubscript{4}) = 4.01 ± 0.07

log(CO) = 4.57 ± 0.73

log(NH\textsubscript{3}) = 9.17 ± 1.77
Figure C1. Distribution of nine stellar and planetary parameters used to generate the synthetic spectra. These distributions follow closely to the actual demographic of currently known population of exoplanets, and therefore they are also subject to biases presented in the original population.

Figure C2. 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 on board Ariel, stellar properties as well as planetary properties.
our retrievals will be, on purpose, biased and will not be retrieving the input chemistry (ground truth). Participants are still tasked to reproduce the 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.

Interested readers can refer to Yip et al. (2022) for a more detailed description of the test set. Spectra generated with free chemistry as well as equilibrium chemistry will be available online for any interested parties to construct their own training and test set.

This paper has been typeset from a \LaTeX file prepared by the author.