# Lessons Learned from Ariel Data Challenge 2022 Inferring Physical Properties of Exoplanets From Next-Generation Telescopes

**Kai Hou Yip**[*] KAI.HOU.YIP@UCL.AC.UK
**Quentin Changeat**[†] QUENTIN.CHANGEAT.18@UCL.AC.UK
**Ingo Waldmann**[*] INGO.WALDMANN@UCL.AC.UK
**Eyup B. Unlu**[‡] EYUP.UNLU@UFL.EDU
**Roy T. Forestano**[‡] ROY.FORESTANO@UFL.EDU
**Alexander Roman**[‡] ALEXROMAN@UFL.EDU
**Katia Matcheva**[‡] MATCHEVA@UFL.EDU
**Konstantin T. Matchev**[‡] MATCHEV@UFL.EDU
**Stefan Stefanov** STEFAN.STEFANOV.ML@GMAIL.COM
**Ondřej Podsztavek**[§] PODSZOND@FIT.CVUT.CZ
**Mario Morvan**[*] MARIO.MORVAN.18@UCL.AC.UK
**Nikolaos Nikolaou**[*] N.NIKOLAOU@UCL.AC.UK
**Ahmed Al-Refaie**[*] AHMED.AL-REFAIE.12@UCL.AC.UK
**Clare Jenner**[*] C.JENNER@UCL.AC.UK
**Chris Johnson**[¶] C.JOHNSON@EPCC.ED.AC.UK
**Angelos Tsiaras**[‖*] ANGELOS.TSIARAS@INAF.IT
**Billy Edwards**[**] B.EDWARDS@SRON.NL
**Catarina Alves de Oliveira**[††] CATARINA.ALVES@ESA.INT
**Jeyan Thiyagalingam**[‡‡] T.JEYAN@STFC.AC.UK
**Pierre-Olivier Lagage** PIERRE-OLIVIER.LAGAGE@CEA.FR
**James Cho** JAMESCHO@BRANDEIS.EDU
**Giovanna Tinetti**[*] G.TINETTI@UCL.AC.UK

**Editors:** Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

[*] Department of Physics and Astronomy, Gower Street, University College London, UK

[†] European Space Agency (ESA), ESA Office, Space Telescope Science Institute (STScI), 3700 San Martin Drive, Baltimore MD 21218, USA

[‡] Physics Department, University of Florida, Gainesville, FL 32611, USA

[§] Faculty of Information Technology, Czech Technical University in Prague, Thákurova 9, 160 00 Prague 6, Czechia

[¶] EPCC, The University of Edinburgh

[‖] INAF - Osservatorio Astrofisico di Arcetri, Florence, Italy

[**] Netherlands Space Research Institute (SRON), Leiden, The Netherlands

[††] European Space Agency, ESAC, Villanueva de la Cãnada, E-28692 Madrid, Spain

[‡‡] STFC Scientific Computing, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Campus, OX11 0QX.
Centre Energie Atomique, Saclay, France
Martin A. Fisher School of Physics, Brandeis University, 415 South Street, Waltham, MA 02453, USA

## Abstract

Exo-atmospheric studies, i.e. the study of exoplanetary atmospheres, is an emerging frontier in Planetary Science. To understand the physical properties of hundreds of exoplanets, astronomers have traditionally relied on sampling-based methods. However, with the growing number of exoplanet detections (i.e. increased data quantity) and advancements in technology from telescopes such as JWST and Ariel (i.e. improved data quality), there is a need for more scalable data analysis techniques. The Ariel Data Challenge 2022 aims to find interdisciplinary solutions from the NeurIPS community. Results from the challenge indicate that machine learning (ML) models have the potential to provide quick insights for thousands of planets and millions of atmospheric models. However, the machine learning models are not immune to data drifts, and future research should investigate ways to quantify and mitigate their negative impact.

**Keywords:** Exoplanet atmospheres, Generative modelling, uncertainty quantification, approximate inference

## 1. Introduction

Exoplanets are planets that orbit stars other than our own Sun. The number of confirmed exoplanets has grown exponentially from 1 to more than 5000 in just under 30 years, thanks to dedicated ground and space based missions such as e.g. Super-WASP (Pollacco et al., 2006), NASA Kepler (Borucki et al., 2010) and the NASA TESS mission (Ricker et al., 2015). The next frontier is characterising these exoplanets, as understanding their atmospheric composition, dynamics and interior helps to provide clues to key questions such as "What are the evolution paths for exoplanets?", "How likely is it to find an Earth-like planet?" and "What are the conditions for life to emerge?". Answering questions such as these is crucial to our understanding of our place in the Universe.

Retrieving the physical properties of exoplanets from observations is a computationally demanding task. Astronomers have traditionally relied on statistical sampling algorithms such as MCMC or Nested Sampling (Skilling, 2006) to approximate the Bayesian posterior distribution of different atmospheric properties, such as the temperature of the planet or trace gas abundances (e.g. Madhusudhan, 2018). However, these algorithms, while precise, are not easily scalable to large datasets, and there have only been a handful of population-level analyses on the different classes of exoplanets  (e.g. Sing et al., 2016; Barstow et al., 2017; Tsiaras et al., 2018; Fisher and Heng, 2018; Pinhas et al., 2019; Mansfield et al., 2021; Roudier et al., 2021; Changeat et al., 2022; Edwards et al., 2022).

The launch of the Ariel Space Telescope in 2029, promises to provide thousands of high-quality observations for a wide range of exoplanets (Tinetti et al., 2021). Conventional sampling algorithms will soon become a major bottleneck to our understanding of planetary characteristics in our local galactic neighbourhood (Yip et al., 2022a; Ardevol Martinez et al., 2022; Matchev et al., 2022a). The field needs, more than ever, a scalable solution to efficiently analyze thousands of planets. The emergence of machine learning-based models makes it possible to analyze thousands, or even millions, of planets at scale within a reasonable amount of time.

However, due to the limited availability of real data, most AI/ML applications for exoplanetary atmosphere characterization are trained on simulated data (e.g. Cobb et al.,

. https://exoplanets.nasa.gov/

2019; Zingales and Waldmann, 2018; Yip et al., 2020; Ardevol Martinez et al., 2022). This means that the joint distribution on which our models are trained and evaluated on is likely to be different from that of real data. This phenomenon, commonly known as "concept drift" in the machine learning literature, often results in poor model performance on real data (e.g. Ditzler et al., 2015; Žliobaitė et al., 2016; Humphrey et al., 2022).

## 2. The Ariel Data Challenge

The Ariel Data Challenge (ADC) is an annual event that seeks innovative solutions aimed at tackling pressing issues faced by the Ariel Space Telescope and the exoplanet community in general. Each year the ADC focuses on a different issue concerning either the technical or scientific aspect of the mission. A summary of the first ADC and its top-ranked solutions can be found in Nikolaou et al. (2020). This year's ADC invited innovative solutions to the problem of atmospheric retrieval. Participants were given a 4-month window to train a model that can infer physical properties from simulated observations from Ariel. The following subsections contain a brief description of the challenge; we refer the interested reader to Yip et al. (2022b) for more details.

### 2.1. Tasks

The competition's goal is to develop a model capable of predicting six atmospheric properties of the exoplanet given a simulated observation from Ariel. These parameters are: molecular abundance of five gases, namely $H_2O$, $CH_4$, $CO_2$, CO, $NH_3$, as well as the mean atmospheric temperature at the planet's terminator. The exact targets to predict would vary depending on the specific participation track chosen. The Light Track asked participants to submit their predictions for the 16th, 50th and 84th percentile of each of the six properties of interest, i.e. 18 targets in total. The Regular Tracks asked participants to submit a 6-dimensional conditional distribution for each test example.

As not all participants may be familiar with the problem domain and scientific background, we provided extensive documentation on the background of the data challenge. This included relevant scientific literature and a challenge starter kit containing the baseline solutions.

### 2.2. Data

Each spectroscopic observation is generated following a 3-step approach. First, a planet configuration is randomly selected from the catalogue of discovered planets, and, based on the chosen planet's configuration a randomly generated atmospheric profile and trace gasses produced. We then used the atmospheric modelling software `TauREx` (Al-Refaie et al., 2021) to produce a theoretical atmospheric model of the exoplanet. The forward model is further processed by `ArielRad` (Mugnai et al., 2020) to generate realistic observations expected by the Ariel Space Mission. This process is made automatic via the software `Alfnoor` (Changeat et al., 2020; Mugnai et al., 2021). A total of 100,000+ simulated Ariel observations were generated for this competition.

. https://github.com/ucl-exoplanets/NeurIPS2022_Baseline

As for the ground truth, we have generated posterior distributions for ∼26% (21,988) of the simulated observations using a Bayesian Nested Sampling (NS) algorithm (i.e. `MultiNest` Feroz and Hobson, 2008; Feroz et al., 2019). These data points are thus ‘fully-labelled’ and the remainder -for which only forward model input and output pairs are available- are considered ‘weakly-labelled’. For more details please refer to Changeat and Yip (2022).

### 2.2.1. Design of test data

| | Set 1 | Set 2 | Set 3 | Set 4 |
|---|---|---|---|---|
| Planetary Configuration | In-Range | Out-Range | In-Range | Out-Range |
| Atmospheric Properties | In-Range | In-Range | Out-Range | Out-Range |

The test set is purposely designed to reflect the fact that actual observations from Ariel will be different from any existing simulated models, and as part of our investigation, we want to know how this change may impact the model’s performance. To this end, we have divided the test-set data into 4 different sets (See Table 1). Set 1 test data are generated with the same atmospheric assumptions and planetary configurations as seen in the training data, and therefore is the most similar to the training data, while Set 4 contains unseen atmospheric assumptions and planetary configurations, we therefore expect it to be most dissimilar. Here In-Range (Out-Range) corresponds to test set information being included (excluded) in the training data.

## 2.3. Metric

### 2.3.1. Light Track

Submissions to the Light Track were evaluated based on the sum of the relative RMSE of all targets $t$ and their individual quartiles $q$, e.g. $RMSE_n = \sqrt{\sum_l^3 \sum_t^6 \left(\frac{q_{l,t}-\hat{q}_{l,t}}{q_{l,t}}\right)^2}$. The final score was calculated by subtracting 1000 from the average performance over the entire test set, i.e.,

$$\text{score}_{Light} = 1000 - \frac{\sum_n^N RMSE_n}{N} \tag{1}$$

### 2.3.2. Regular Track

Submissions to the Regular Track were evaluated using the scaled Wasserstein-2 distance between the predicted conditional distribution $f$ and the NS-generated posterior distribution $\hat{f}$.

$$W_{2,t}(f, \hat{f}) = \inf_{\pi \in \Gamma(f,\hat{f})} \int_{\mathbb{R}\times\mathbb{R}} (x-y)^2 d\pi(x,y) \tag{2}$$

where $\Gamma(f, \hat{f})$ represents the set of probability distributions on the metric space $\mathbb{R} \times \mathbb{R}$, whose marginal distributions are $f$ and $\hat{f}$ on first and second factor, respectively. The total distance, $W_{2,n}$, for each test example is the sum of the individual $W_{2,t}$ from each dimension, scaled by the size of their respective prior bounds, $B_t$ i.e. $W_{2,n} = \sum_t W_{2,t}/B_t$. The final score was calculated by subtracting 1000 from the average score over the entire test set, i.e.

$$\text{score}_{Regular} = 1000 - \frac{\sum_{n=1}^{N} W_{2,n}}{N} \tag{3}$$

Note that inputs to the metric were subjected to a number of preprocessing steps before being admitted by the metric function. Those conditions were 1.) Values outside the target's prior bounds were replaced with boundary values of the respective bounds 2.) the number of samples per individual test examples must not exceed 5000.

## 3. Standout Solutions

### 3.1. Gators' Solution

Unlu, Forestano, Roman, Matcheva and Matchev (under their corporate username "gators") drew inspiration for their solution from transformer architectures with self-attention layers. The model is built from several fully connected neural networks some of which use as their inputs concatenations or products of the outputs of previous modules. The model includes a "modification layer" which modifies the spectrum using all available features, an "attention layer" which enhances (suppresses) relevant (redundant) information, an "auxiliary layer" tapping into the information contained in the auxiliary parameters, followed by the main layer which predicts the learned parameters. The code for the "gators" solution can be accessed on GitHub at https://github.com/EyupBunlu/ArielDataChallenge2022Gators.

A significant improvement of the model's prediction was made by preprocessing the data using physics-motivated feature engineering. For example, to isolate the effect of the atmosphere, for each planet $i$, the contribution of the opaque planet disk was subtracted from the observed modulation $M_{i\lambda}$ at each wavelength bin $\lambda$ as (Matchev et al., 2022b)

$$M'_{i\lambda} = M_{i\lambda} - \left(\frac{R_{ip}}{R_{is}}\right)^2, \tag{4}$$

where $R_{ip}$ and $R_{is}$ are the respective planet and star radii. Subsequently, the features (4) and the noise were rescaled as $M''_{i\lambda} = M'_{i\lambda}/\max_\lambda(M'_{i\lambda})$ and $\epsilon'_{i\lambda} = \epsilon_{i\lambda}/\max_\lambda(M'_{i\lambda})$. New dimensionless features motivated by the dimensional analysis of Matchev et al. (2022a) were also added to the inputs, e.g., ratios of the planet radius, scale height, distance to the host star $D$, etc. Finally, the auxiliary features were standardized and the modulation data was cleaned by modifying unphysical values (outliers).

The training was done only on the 21,988 labeled samples with an 80/20 train-test split and a dropout layer. The Adam optimiser (Kingma and Ba, 2017) was used with different learning rates at various stages of the training. The six-dimensional population of the labels was parameterized with the following ansatz

$$\rho(T, \boldsymbol{x}; T_p, \mu_i, \sigma_i, A_i, m_i) = \left[\frac{\Theta(T_p - T)}{\sigma_{T_1}\sqrt{2\pi}} e^{-\frac{(T-T_p)^2}{2\sigma_{T_1}^2}} + \frac{\Theta(T - T_p)}{\sigma_{T_2}\sqrt{2\pi}} e^{-\frac{(T-T_p)^2}{2\sigma_{T_2}^2}}\right] \times \prod_{i=1}^{5} \left[\frac{A_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} + (1 - A_i)\frac{\Theta(x_i + 12)\Theta(m_i - x_i)}{m_i + 12}\right],$$

which involves a Bigaussian distribution for the temperature (with mean $T_p = T_s\sqrt{\frac{R_s}{2D}}$ and standard deviations $\sigma_{T_1}$ and $\sigma_{T_2}$), and a sum of a gaussian distribution (mean $\mu_i$ and standard deviation $\sigma_i$) and a uniform distribution (from $-12$ to $m_i$) for $x_i$, the log of each concentration. Thermal equilibrium between the star and the planet was used to relate $T_p$ to the temperature of the host star $T_s$. The model predicted the parameters of the ansatz. Their training values were obtained by fitting to the provided trace data. The posterior distributions are then sampled from the ansatz with the learned values of the parameters.

### 3.2. Stefan Stefanov's Solution

Stefan Stefanov's solution represents each atmospheric property distribution as an independent mixture model of a normal and a uniform distribution. The 5 mixture model parameters: $\mu$ and $\sigma$ of the normal distribution, $a$, $b$ (upper and lower bound of the uniform distribution) and $\alpha$ - the mixing coefficient are predicted by a neural network. The neural network architecture is designed as a 1D convolutional autoencoder. The inputs to the encoder are the spectrum, the spectrum with subtracted mean (Ardevol Martinez et al., 2022) and the spectrum noise. These inputs are normalized by Deep Adaptive Input Normalization (Passalis et al., 2019) which constitutes the trainable normalization layer part of the network. The encoder contains 4 ResNet (He et al., 2016) blocks adapted for 1D inputs. The features produced by the ResNet blocks are concatenated with the provided auxiliary features and fed into a dense block consisting of 2 fully-connected layers followed by a head layer which outputs the predicted parameters. Target values for the 5 parameters described above are obtained from fitting mixture models to the provided trace data with expectation-maximization using the external library *pomegranate* (Schreiber, 2018).

A multi-task learning approach is applied for the neural network's training. In addition to the mixture model parameters the neural network is trained to predict the forward model parameters, the quartiles for the light track and a spectrum reconstruction. The neural network has a decoder component that performs input spectrum data reconstruction. The inputs to the decoder are the predictions from the encoder for the mixture, forward model parameters and quartiles concatenated with the auxiliary features. The decoder consists of 1D transposed convolution layers where the final layer outputs reconstruction of the input spectrum data. This autoencoder architecture allows for applying semi-supervised learning where test data is also included for model training.

The predictions for the regular track are sampled from the mixture models with parameters predicted by the neural network. The final solution is an ensemble of 10 neural networks trained with different random initializations (Lakshminarayanan et al., 2017). Each model provides one-tenth of the submission samples. The code for the solution is available at: `https://github.com/stefanistefanov/NeurIPS2022_Ariel_Challenge`.

### 3.3. Podsztavek's Solution

Podsztavek (under his username "podondra") designed a solution based on *deep ensembles* (Lakshminarayanan et al., 2017). A deep ensemble is an ensemble of $M$ neural networks that predict means and variances, i.e. normal distributions (Nix and Weigend, 1994). The deep ensemble used in this solution consisted of $M = 20$ convolutional neural networks (CNNs). Therefore, its output was a probability density function (PDF), a mixture of 20 equally weighted normal distributions. This solution's code is available online on GitHub, at https://github.com/podondra/ariel-data-challenge.

The inputs of the CNNs were *spectra* of transit depths with corresponding *auxiliary data* comprising all individual features (e.g. host star distance, exoplanet mass). Standardisation was applied to both spectra and auxiliary data. Each spectrum was standardised so that it had zero mean and unit variance. This standardisation reduced differences in the ranges of transit depths of individual spectra. Therefore, CNNs could focus on the shapes of spectra since the abundance of molecules in atmospheres determine them. Auxiliary data were standardised feature-wise, i.e. each auxiliary data feature was subtracted by its training set's mean and divided by its training set's standard deviation. The training set included all 21988 annotated exoplanets (i.e. spectra, auxiliary data, and annotations). The original annotations were weighted traces (i.e. samples) from distributions of targets. Such annotations would make the training of CNNs difficult. Therefore, the annotations were simplified to be 6 normal distributions fitted to the weighted traces independently for each target.

Each CNN was a modification of the VGG Net-A CNN (Simonyan and Zisserman, 2015). It consisted of a convolutional part (6 convolutional and 4 max pooling layers) and fully connected (FC) part (7 FC layers, each with 1024 neurons). The convolutional part processed spectra; its output was concatenated with auxiliary data into a vector processed by the FC part. The activation function of all layers (except the last one) was the rectified linear unit (ReLU). The last layer outputted the 6 normal distributions, i.e. 6 means and 6 variances. The softplus function outputted these 6 variances, and a minimal variance of $10^{-6}$ was added for numerical stability. All CNNs were trained with Kullback–Leibler divergence as the loss function using Adam optimiser (Kingma and Ba, 2017) with a learning rate of $10^{-4}$ and batch size of 256. These and other hyperparameters were optimised on a separate validation set (20 % of the training set) using early stopping on the light score. However, after optimising them, data from both training and validation sets were used to train the final CNNs for 2048 epochs. This number of epochs ensured sufficient convergence of CNNs.

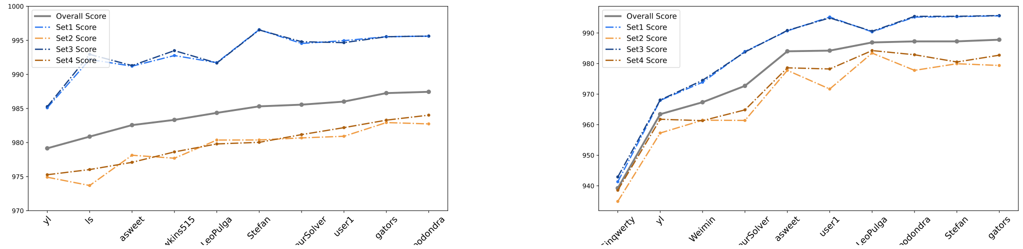
The deep ensemble of 20 CNNs generated samples from the predicted distribution: 250 samples were sampled from 6 normal distributions outputted by each CNN. Therefore, there were 5000 samples in total for the Regular Track. Then, the sample quartiles were computed for the Light Track.

## 4. Discussion

### 4.1. Learning Paradigms

In this section we will discuss our observations based on the submissions we received on the final evaluation round.

**Learning strategies** Whilst there was no restriction on the learning strategies imposed by the hosts, most participants leveraged the availability of labelled data and trained their models in a supervised manner. Few have made use of the weakly-labelled data to pre-train their models. Most participants trained the same or almost identical models for both tracks, as similarities between the two tracks allowed for it.



$(a)$ Light Track

$(b)$ Regular Track

Figure 1: Performance on both tracks for top-10 ranking solutions on the final evaluation set

**Model architecture** Many top-ranking solutions have opted for generative models such as Mixture Density Models (MDNs) and, in some cases, Normalising Flows based networks, possibly due to the fact that the regular track explicitly asked for a multivariate conditional distribution, where generative models tend to excel. Another reason could be the prevalence of Gaussian and Uniform distributions (or a mixture of both) in the dataset, which makes MDNs a good approximation. Some solutions went further and trained ensemble models using MDNs as their base models. However, the winner of the competition did not rely on generative models and instead went for a discriminative model with physics-motivated data engineering.

**Data Pre- and post- processing** There are significant differences in how data is prepared and processed among the solutions offered by participants with different backgrounds. Those with expertise in astronomy tended to preprocess the data using established theories in the relevant area. On the other hand, participants from an AI background tended to standardize the data without considering any particular domain knowledge. The former approach seems to make the learning task easier, and it allows for more flexibility in the learning approach, as demonstrated by the winning team. On the other hand, the latter approach places more emphasis on the design of the network itself. The competition's results indicated that having domain-specific knowledge can be advantageous in these types of problem. However, it is also possible that with a more diverse and larger training set, one may be able to compensate for the lack of such knowledge. A similar divide in terms of emphasis on data preprocessing vs. learning algorithm was also observed in past ADCs (Nikolaou et al., 2020).

### 4.2. General Performance on the test set

Figure 1 shows the average performance of the qualifying teams in both tracks on different test datasets. It is evident that the average performance across the different test sets (see Section 2.2.1 for more information on the differences between different sets) is not homogeneous. Every team (in the top-10 places), regardless of their ranking, tended to perform

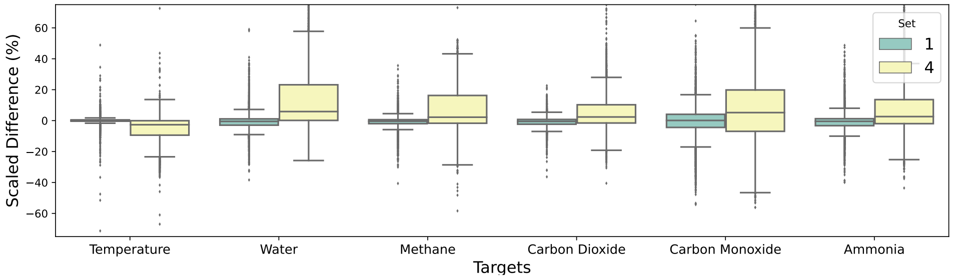


Figure 2: Distribution of % difference on each targets for top-10 winners on Set 1 and 4

better on Set 1 and 3 (blue and purple), but worse on Set 2 and 4 (orange and brown). The inclusion of intermediate sets (Set 2 and 3) helps to identify the confounding factor here - changes in atmospheric assumptions have a far larger impact on model performance than changes in planetary configuration. The impact is not homogeneous across all teams' models, though; while top-ranked models generally suffer a more significant drop in performance on Set 2 and 4, the gap generally narrows with lower-ranked models.

The observed discrepancy between different sets has important implications. The systematic drop in performance (on Set 2 and 4) across all models suggests that, regardless of different training procedures and model architectures, machine learning models tend to perform worse on data originating from a different distribution. This finding adds to the pile of evidence indicating that concept drift tends to deteriorate model performance (Žliobaitė et al., 2016).

This problem is particularly relevant in the context of space missions, as the corresponding ML models have to be trained on simulated data in anticipation of the actual data being collected. The two data distributions are certainly different, and this shift may be identifiable by comparing the model's prediction to the result from MCMC methods and contrasting it with its baseline performance. However, doing so would contradict our original goal of avoiding lengthy MCMC integration.

### 4.2.1. Performance on physical parameters

Standing from a physical point of view, we can gain more insights into how the different atmospheric assumptions may have have influenced the models' performance. Figure 2 compares the scaled difference of each target, stratified by their respective test sets (Set 1 and 4). We chose to compute the signed difference between the model's prediction and the ground truth, scaled by the respective ground truth i.e. $(y - \hat{y})/y$. This metric helps reveal the direction of the bias, in addition to the degree of deviation from the targets. As expected, most models performed well on Set 1, with the predictions for most targets being within 10% of the ground truth, and Carbon Monoxide (CO) being the most challenging to predict. This finding is consistent with existing literature (Changeat et al., 2020; Yip et al., 2020; Changeat and Yip, 2022).

. As a quick reminder, Set 1 data is the most similar to the training data while Set 4 data is the most dissimilar to the training data.

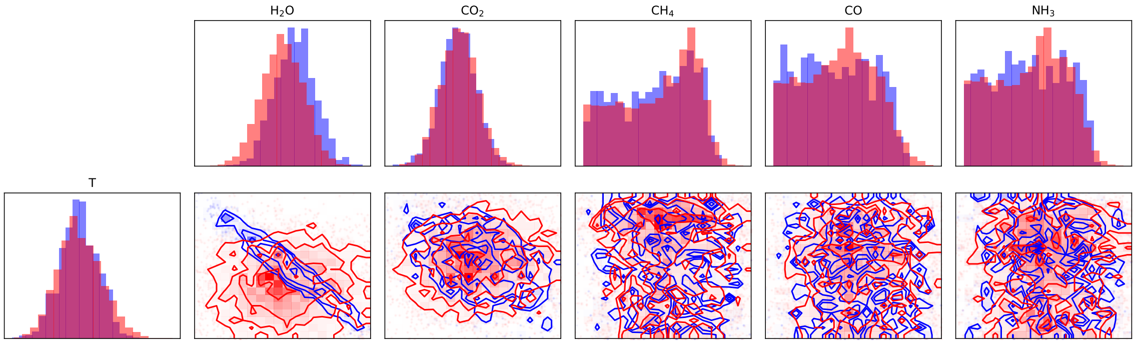


Figure 3: An example comparing the distributions generated from Nested Sampling (blue) and the submission from the participant (red). The first row shows the marginal distribution between individual targets, and the second row (except for the first column) shows the covariance between Temperature (T) and other targets. The Wasserstein-2 metric gave a high score due to the similarities between marginal distributions but failed to penalize the differences in correlations.

However, the performance of all models saw a sharp decline across all targets when using test data from Set 4, with water ($H_2O$), methane ($CH_4$), and carbon monoxide (CO) being the most affected. Furthermore, the corresponding boxplots for each individual target reveal that most models tend to underpredict the molecular abundances of gases while overpredicting the planet's temperature.

These observations seem to suggest that, regardless of the different training procedure and inductive biases the models use, most of them have learned the mapping from the feature (spectrum) space to target (physical properties) space under the atmospheric assumption provided by the training data. In other words, most models learned to relate features of the input to *individual* targets, without necessarily taking into account how each of the parameters may contribute to the full spectrum. On the other hand, targets inferred under the ground truth technique (atmospheric retrieval) are always constrained by the observation, i.e. it will always look for a set of physical parameters that minimises the distance between the observed and the theoretical spectrum generated under the training atmospheric assumption.

### 4.3. Choice of metric

Quantifying the distance between two 6-dimensional distributions is not an easy task. While the Wasserstein-2 metric is sensitive to differences between predictions and ground truth for individual targets, it fails to account for the covariance between different pairs of targets. This has resulted in reduced scientific yield, as submissions were unable to reproduce the covariance between atmospheric targets, despite achieving high similarity on the marginal (univariate) distributions (see Figure 3 for an example).

## 5. Summary and Future Outlook

Currently, sampling-based algorithms like MCMC or Nested Sampling are considered the best options for addressing inverse problems like atmospheric retrieval. However, these methods do not support scalability to handle large datasets. The Ariel Data Challenge 2022 represents one of the first organized attempts to develop scalable solutions for solving inverse problems in the field of exo-atmospheric studies.

An important message has emerged from the competition: *simulation-based inference models are prone to data drifts*, which are almost certain to occur in future instruments where real data is unavailable until after launch. Despite this limitation, participants achieved good performance by designing a well-crafted data processing pipelines and choosing clever machine learning architectures. Consistent with the outcomes of the Ariel Data Challenge in 2019 and 2021, the inclusion of domain knowledge in the pipeline has proven vital to improving model performance.

The significant decline in model performance when encountering an unfamiliar data distribution may be attributed to the insufficient variety of the training data. Future versions of the data challenge (or similar investigations) should consider how enhancing the diversity of the training data could improve the model's capacity to generalize beyond its training set. Additionally, it would be beneficial to explore alternative metrics that can explicitly and rigidly incorporate the physical constraints imposed by the atmospheric forward model.

## Acknowledgments

We express our gratitude to all the participants and the organizing chair of NeurIPS for their support in making this competition possible. We also extend our thanks to DiRAC - High Performance Computing Facilities for their generous contribution of computing resources, without which the competition data would not have been generated.

The authors acknowledge the use of the UCL Myriad High Performance Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

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) and the European Union's Horizon 2020 COMPET programme (grant agreement No 776403, ExoplANETS A). Furthermore, we acknowledge funding by the Science and Technology Funding Council (STFC) grants: ST/K502406/1, ST/P000282/1, ST/P002153/1 and ST/S002634/1.

Podsztavek acknowledges the support of his supervisor Prof. Pavel Tvrdík, co-supervisor Dr Petr Škoda, the European Regional Development Fund project "Research Center for Informatics" (No. CZ.02.1.01/0.0/0.0/16 019/0000765), and the Grant Agency of the Czech Technical University in Prague (No. SGS23/209/OHK3/3T/18).

JT's time on this work were supported by the EPRSRC grants EP/T10001569/1 via Alan Turing Institute and EP/V001310/1.

## References

A. F. Al-Refaie, Q. Changeat, I. P. Waldmann, and G. Tinetti. TauREx 3: A Fast, Dynamic, and Extendable Framework for Retrievals. , 917(1):37, August 2021. doi: 10.3847/1538-4357/ac0252.

Francisco Ardevol Martinez, Michiel Min, Inga Kamp, and Paul I. Palmer. Convolutional neural networks as an alternative to Bayesian retrievals. *arXiv e-prints*, art. arXiv:2203.01236, March 2022. doi: 10.48550/arXiv.2203.01236.

J. K. Barstow, S. Aigrain, P. G. J. Irwin, and D. K. Sing. A Consistent Retrieval Analysis of 10 Hot Jupiters Observed in Transmission. , 834(1):50, January 2017. doi: 10.3847/1538-4357/834/1/50.

William J. Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D. Cochran, Edna DeVore, Edward W. Dunham, Andrea K. Dupree, Thomas N. Gautier, John C. Geary, Ronald Gilliland, Alan Gould, Steve B. Howell, Jon M. Jenkins, Yoji Kondo, David W. Latham, Geoffrey W. Marcy, Søren Meibom, Hans Kjeldsen, Jack J. Lissauer, David G. Monet, David Morrison, Dimitar Sasselov, Jill Tarter, Alan Boss, Don Brownlee, Toby Owen, Derek Buzasi, David Charbonneau, Laurance Doyle, Jonathan Fortney, Eric B. Ford, Matthew J. Holman, Sara Seager, Jason H. Steffen, William F. Welsh, Jason Rowe, Howard Anderson, Lars Buchhave, David Ciardi, Lucianne Walkowicz, William Sherry, Elliott Horch, Howard Isaacson, Mark E. Everett, Debra Fischer, Guillermo Torres, John Asher Johnson, Michael Endl, Phillip MacQueen, Stephen T. Bryson, Jessie Dotson, Michael Haas, Jeffrey Kolodziejczak, Jeffrey Van Cleve, Hema Chandrasekaran, Joseph D. Twicken, Elisa V. Quintana, Bruce D. Clarke, Christopher Allen, Jie Li, Haley Wu, Peter Tenenbaum, Ekaterina Verner, Frederick Bruhweiler, Jason Barnes, and Andrej Prsa. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327(5968):977, February 2010. doi: 10.1126/science.1185402.

Q. Changeat, A. Al-Refaie, L. V. Mugnai, B. Edwards, I. P. Waldmann, E. Pascale, and G. Tinetti. Alfnoor: A retrieval simulation of the ariel target list. *The Astronomical Journal*, 160(2):80, Jul 2020. ISSN 1538-3881. doi: 10.3847/1538-3881/ab9a53. URL http://dx.doi.org/10.3847/1538-3881/ab9a53.

Q. Changeat, B. Edwards, A. F. Al-Refaie, A. Tsiaras, J. W. Skinner, J. Y. K. Cho, K. H. Yip, L. Anisman, M. Ikoma, M. F. Bieger, O. Venot, S. Shibata, I. P. Waldmann, and G. Tinetti. Five Key Exoplanet Questions Answered via the Analysis of 25 Hot-Jupiter Atmospheres in Eclipse. , 260(1):3, May 2022. doi: 10.3847/1538-4365/ac5cc2.

Quentin Changeat and Kai Hou Yip. ESA-Ariel Data Challenge NeurIPS 2022: Introduction to exo-atmospheric studies and presentation of the Ariel Big Challenge (ABC) Database. *arXiv e-prints*, art. arXiv:2206.14633, June 2022.

Adam D. Cobb, Michael D. Himes, Frank Soboczenski, Simone Zorzan, Molly D. O'Beirne, Atılım Güneş Baydin, Yarin Gal, Shawn D. Domagal-Goldman, Giada N. Arney, Daniel Angerhausen, and II 2018 NASA FDL Astrobiology Team. An Ensemble of Bayesian

Neural Networks for Exoplanetary Atmospheric Retrieval. , 158(1):33, July 2019. doi: 10.3847/1538-3881/ab2390.

Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25, 2015. doi: 10.1109/MCI.2015.2471196.

Billy Edwards, Quentin Changeat, Angelos Tsiaras, Kai Hou Yip, Ahmed F. Al-Refaie, Lara Anisman, Michelle F. Bieger, Amelie Gressier, Sho Shibata, Nour Skaf, Jeroen Bouwman, James Y-K. Cho, Masahiro Ikoma, Olivia Venot, Ingo Waldmann, Pierre-Olivier Lagage, and Giovanna Tinetti. Exploring the Ability of HST WFC3 G141 to Uncover Trends in Populations of Exoplanet Atmospheres Through a Homogeneous Transmission Survey of 70 Gaseous Planets. *arXiv e-prints*, art. arXiv:2211.00649, November 2022.

F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, Jan 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2007.12353.x. URL http://dx.doi.org/10.1111/j.1365-2966.2007.12353.x.

Farhan Feroz, Michael P. Hobson, Ewan Cameron, and Anthony N. Pettitt. Importance nested sampling and the multinest algorithm. *The Open Journal of Astrophysics*, 2(1), Nov 2019. ISSN 2565-6120. doi: 10.21105/astro.1306.2144. URL http://dx.doi.org/10.21105/astro.1306.2144.

Chloe Fisher and Kevin Heng. Retrieval analysis of 38 WFC3 transmission spectra and resolution of the normalization degeneracy. , 481(4):4698–4727, December 2018. doi: 10.1093/mnras/sty2550.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

A. Humphrey, W. Kuberski, J. Bialek, N. Perrakis, W. Cools, N. Nuyttens, H. Elakhrass, and P. A. C. Cunha. Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth. , 517(1):L116–L120, November 2022. doi: 10.1093/mnrasl/slac120.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Nikku Madhusudhan. Atmospheric Retrieval of Exoplanets. In Hans J. Deeg and Juan Antonio Belmonte, editors, *Handbook of Exoplanets*, page 104. 2018. doi: 10.1007/978-3-319-55333-7_104.

Megan Mansfield, Michael R. Line, Jacob L. Bean, Jonathan J. Fortney, Vivien Parmentier, Lindsey Wiser, Eliza M. R. Kempton, Ehsan Gharib-Nezhad, David K. Sing, Mercedes López-Morales, Claire Baxter, Jean-Michel Désert, Mark R. Swain, and Gael M. Roudier. A unique hot Jupiter spectral sequence with evidence for compositional diversity. *Nature Astronomy*, 5:1224–1232, October 2021. doi: 10.1038/s41550-021-01455-4.

Konstantin T. Matchev, Katia Matcheva, and Alexander Roman. Analytical Modeling of Exoplanet Transit Spectroscopy with Dimensional Analysis and Symbolic Regression. *The Astrophysical Journal*, 930(1):33, May 2022a. doi: 10.3847/1538-4357/ac610c.

Konstantin T. Matchev, Katia Matcheva, and Alexander Roman. Transverse Vector Decomposition Method for Analytical Inversion of Exoplanet Transit Spectra. *The Astrophysical Journal*, 939(2):95, November 2022b. doi: 10.3847/1538-4357/ac82f3.

Lorenzo V. Mugnai, Enzo Pascale, Billy Edwards, Andreas Papageorgiou, and Subhajit Sarkar. ArielRad: the Ariel Radiometric Model. *arXiv e-prints*, art. arXiv:2009.07824, September 2020.

Lorenzo V. Mugnai, Ahmed Al-Refaie, Andrea Bocchieri, Quentin Changeat, Enzo Pascale, and Giovanna Tinetti. Alfnoor: Assessing the Information Content of Ariel's Low-resolution Spectra with Planetary Population Studies. , 162(6):288, December 2021. doi: 10.3847/1538-3881/ac2e92.

Nikolaos Nikolaou, Ingo P. Waldmann, Angelos Tsiaras, Mario Morvan, Billy Edwards, Kai Hou Yip, Giovanna Tinetti, Subhajit Sarkar, James M. Dawson, Vadim Borisov, Gjergji Kasneci, Matej Petkovic, Tomaz Stepisnik, Tarek Al-Ubaidi, Rachel Louise Bailey, Michael Granitzer, Sahib Julka, Roman Kern, Patrick Ofner, Stefan Wagner, Lukas Heppe, Mirko Bunse, and Katharina Morik. Lessons Learned from the 1st ARIEL Machine Learning Challenge: Correcting Transiting Exoplanet Light Curves for Stellar Spots. *arXiv e-prints*, art. arXiv:2010.15996, October 2020.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60, 1994. doi: 10.1109/ICNN.1994.374138.

Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Deep adaptive input normalization for price forecasting using limit order book data. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Arazi Pinhas, Nikku Madhusudhan, Siddharth Gandhi, and Ryan MacDonald. $H_2O$ abundances and cloud properties in ten hot giant exoplanets. , 482(2):1485–1498, January 2019. doi: 10.1093/mnras/sty2544.

D. L. Pollacco, I. Skillen, A. Collier Cameron, D. J. Christian, C. Hellier, J. Irwin, T. A. Lister, R. A. Street, R. G. West, D. R. Anderson, W. I. Clarkson, H. Deeg, B. Enoch, A. Evans, A. Fitzsimmons, C. A. Haswell, S. Hodgkin, K. Horne, S. R. Kane, F. P. Keenan, P. F. L. Maxted, A. J. Norton, J. Osborne, N. R. Parley, R. S. I. Ryans, B. Smalley, P. J. Wheatley, and D. M. Wilson. The WASP Project and the SuperWASP Cameras. , 118(848):1407–1418, October 2006. doi: 10.1086/508556.

George R. Ricker, Joshua N. Winn, Roland Vanderspek, David W. Latham, Gáspár Á. Bakos, Jacob L. Bean, Zachory K. Berta-Thompson, Timothy M. Brown, Lars Buchhave, Nathaniel R. Butler, R. Paul Butler, William J. Chaplin, David Charbonneau, Jørgen Christensen-Dalsgaard, Mark Clampin, Drake Deming, John Doty, Nathan De Lee, Courtney Dressing, Edward W. Dunham, Michael Endl, Francois Fressin, Jian Ge, Thomas Henning, Matthew J. Holman, Andrew W. Howard, Shigeru Ida, Jon M. Jenkins, Garrett Jernigan, John Asher Johnson, Lisa Kaltenegger, Nobuyuki Kawai, Hans Kjeldsen, Gregory Laughlin, Alan M. Levine, Douglas Lin, Jack J. Lissauer, Phillip MacQueen, Geoffrey Marcy, Peter R. McCullough, Timothy D. Morton, Norio Narita, Martin Paegert, Enric Palle, Francesco Pepe, Joshua Pepper, Andreas Quirrenbach, Stephen A. Rinehart, Dimitar Sasselov, Bun'ei Sato, Sara Seager, Alessandro Sozzetti, Keivan G. Stassun, Peter Sullivan, Andrew Szentgyorgyi, Guillermo Torres, Stephane Udry, and Joel Villasenor. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003, January 2015. doi: 10.1117/1.JATIS.1.1.014003.

Gael M. Roudier, Mark R. Swain, Murthy S. Gudipati, Robert A. West, Raissa Estrela, and Robert T. Zellem. Disequilibrium Chemistry in Exoplanet Atmospheres Observed with the Hubble Space Telescope. , 162(2):37, August 2021. doi: 10.3847/1538-3881/abfdad.

Jacob Schreiber. pomegranate: Fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, 18(164):1–6, 2018. URL http://jmlr.org/papers/v18/17-636.html.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

David K. Sing, Jonathan J. Fortney, Nikolay Nikolov, Hannah R. Wakeford, Tiffany Kataria, Thomas M. Evans, Suzanne Aigrain, Gilda E. Ballester, Adam S. Burrows, Drake Deming, Jean-Michel Désert, Neale P. Gibson, Gregory W. Henry, Catherine M. Huitson, Heather A. Knutson, Alain Lecavelier Des Etangs, Frederic Pont, Adam P. Showman, Alfred Vidal-Madjar, Michael H. Williamson, and Paul A. Wilson. A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. , 529(7584): 59–62, January 2016. doi: 10.1038/nature16068.

John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4): 833 – 859, 2006. doi: 10.1214/06-BA127. URL https://doi.org/10.1214/06-BA127.

Giovanna Tinetti, Paul Eccleston, Carole Haswell, Pierre-Olivier Lagage, Jérémy Leconte, Theresa Lüftinger, Giusi Micela, Michel Min, Göran Pilbratt, Ludovic Puig, Mark Swain, Leonardo Testi, Diego Turrini, Bart Vandenbussche, Maria Rosa Zapatero Osorio, Anna Aret, Jean-Philippe Beaulieu, Lars Buchhave, Martin Ferus, Matt Griffin, Manuel Guedel, Paul Hartogh, Pedro Machado, Giuseppe Malaguti, Enric Pallé, Mirek Rataj, Tom Ray, Ignasi Ribas, Robert Szabó, Jonathan Tan, Stephanie Werner, Francesco Ratti, Carsten Scharmberg, Jean-Christophe Salvignol, Nathalie Boudin, Jean-Philippe Halain, Martin Haag, Pierre-Elie Crouzet, Ralf Kohley, Kate Symonds, Florian Renk, Andrew Caldwell, Manuel Abreu, Gustavo Alonso, Jerome Amiaux, Michel Berthé, Georgia Bishop,

Neil Bowles, Manuel Carmona, Deirdre Coffey, Josep Colomé, Martin Crook, Lucile Désjonqueres, José J. Díaz, Rachel Drummond, Mauro Focardi, Jose M. Gómez, Warren Holmes, Matthijs Krijger, Zsolt Kovacs, Tom Hunt, Richardo Machado, Gianluca Morgante, Marc Ollivier, Roland Ottensamer, Emanuele Pace, Teresa Pagano, Enzo Pascale, Chris Pearson, Søren Møller Pedersen, Moshe Pniel, Stéphane Roose, Giorgio Savini, Richard Stamper, Peter Szirovicza, Janos Szoke, Ian Tosh, Francesc Vilardell, Joanna Barstow, Luca Borsato, Sarah Casewell, Quentin Changeat, Benjamin Charnay, Svatopluk Civiš, Vincent Coudé du Foresto, Athena Coustenis, Nicolas Cowan, Camilla Danielski, Olivier Demangeon, Pierre Drossart, Billy N. Edwards, Gabriella Gilli, Therese Encrenaz, Csaba Kiss, Anastasia Kokori, Masahiro Ikoma, Juan Carlos Morales, João Mendonça, Andrea Moneti, Lorenzo Mugnai, Antonio García Muñoz, Ravit Helled, Mihkel Kama, Yamila Miguel, Nikos Nikolaou, Isabella Pagano, Olja Panic, Miriam Rengel, Hans Rickman, Marco Rocchetto, Subhajit Sarkar, Franck Selsis, Jonathan Tennyson, Angelos Tsiaras, Olivia Venot, Krisztián Vida, Ingo P. Waldmann, Sergey Yurchenko, Gyula Szabó, Rob Zellem, Ahmed Al-Refaie, Javier Perez Alvarez, Lara Anisman, Axel Arhancet, Jaume Ateca, Robin Baeyens, John R. Barnes, Taylor Bell, Serena Benatti, Katia Biazzo, Maria Błecka, Aldo Stefano Bonomo, José Bosch, Diego Bossini, Jeremy Bourgalais, Daniele Brienza, Anna Brucalassi, Giovanni Bruno, Hamish Caines, Simon Calcutt, Tiago Campante, Rodolfo Canestrari, Nick Cann, Giada Casali, Albert Casas, Giuseppe Cassone, Christophe Cara, Manuel Carmona, Ludmila Carone, Nathalie Carrasco, Quentin Changeat, Paolo Chioetto, Fausto Cortecchia, Markus Czupalla, Katy L. Chubb, Angela Ciaravella, Antonio Claret, Riccardo Claudi, Claudio Codella, Maya Garcia Comas, Gianluca Cracchiolo, Patricio Cubillos, Vania Da Peppo, Leen Decin, Clemence Dejabrun, Elisa Delgado-Mena, Anna Di Giorgio, Emiliano Diolaiti, Caroline Dorn, Vanessa Doublier, Eric Doumayrou, Georgina Dransfield, Luc Dumaye, Emma Dunford, Antonio Jimenez Escobar, Vincent Van Eylen, Maria Farina, Davide Fedele, Alejandro Fernández, Benjamin Fleury, Sergio Fonte, Jean Fontignie, Luca Fossati, Bernd Funke, Camille Galy, Zoltán Garai, Andrés García, Alberto García-Rigo, Antonio Garufi, Giuseppe Germano Sacco, Paolo Giacobbe, Alejandro Gómez, Arturo Gonzalez, Francisco Gonzalez-Galindo, Davide Grassi, Caitlin Griffith, Mario Giuseppe Guarcello, Audrey Goujon, Amélie Gressier, Aleksandra Grzegorczyk, Tristan Guillot, Gloria Guilluy, Peter Hargrave, Marie-Laure Hellin, Enrique Herrero, Matt Hills, Benoit Horeau, Yuichi Ito, Niels Christian Jessen, Petr Kabath, Szilárd Kálmán, Yui Kawashima, Tadahiro Kimura, Antonín Knížek, Laura Kreidberg, Ronald Kruid, Diederik J. M. Kruijssen, Petr Kubelík, Luisa Lara, Sebastien Lebonnois, David Lee, Maxence Lefevre, Tim Lichtenberg, Daniele Locci, Matteo Lombini, Alejandro Sanchez Lopez, Andrea Lorenzani, Ryan MacDonald, Laura Magrini, Jesus Maldonado, Emmanuel Marcq, Alessandra Migliorini, Darius Modirrousta-Galian, Karan Molaverdikhani, Sergio Molinari, Paul Mollière, Vincent Moreau, Giuseppe Morello, Gilles Morinaud, Mario Morvan, Julianne I. Moses, Salima Mouzali, Nariman Nakhjiri, Luca Naponiello, Norio Narita, Valerio Nascimbeni, Athanasia Nikolaou, Vladimiro Noce, Fabrizio Oliva, Pietro Palladino, Andreas Papageorgiou, Vivien Parmentier, Giovanni Peres, Javier Pérez, Santiago Perez-Hoyos, Manuel Perger, Cesare Cecchi Pestellini, Antonino Petralia, Anne Philippon, Arianna Piccialli, Marco Pignatari, Giampaolo Piotto, Linda Podio, Gianluca Polenta, Giampaolo Preti, Theodor Pribulla, Manuel Lopez Puertas, Monica Rainer, Jean-Michel Reess, Paul Rim-

mer, Séverine Robert, Albert Rosich, Loic Rossi, Duncan Rust, Ayman Saleh, Nicoletta Sanna, Eugenio Schisano, Laura Schreiber, Victor Schwartz, Antonio Scippa, Bálint Seli, Sho Shibata, Caroline Simpson, Oliver Shorttle, N. Skaf, Konrad Skup, Mateusz Sobiecki, Sergio Sousa, Alessandro Sozzetti, Judit Šponer, Lukas Steiger, Paolo Tanga, Paul Tackley, Jake Taylor, Matthias Tecza, Luca Terenzi, Pascal Tremblin, Andrea Tozzi, Amaury Triaud, Loïc Trompet, Shang-Min Tsai, Maria Tsantaki, Diana Valencia, Ann Carine Vandaele, Mathieu Van der Swaelmen, Adibekyan Vardan, Gautam Vasisht, Allona Vazan, Ciro Del Vecchio, Dave Waltham, Piotr Wawer, Thomas Widemann, Paulina Wolkenberg, Gordon Hou Yip, Yuk Yung, Mantas Zilinskas, Tiziano Zingales, and Paola Zuppella. Ariel: Enabling planetary science across light-years. *arXiv e-prints*, art. arXiv:2104.04824, April 2021.

A. Tsiaras, I. P. Waldmann, T. Zingales, M. Rocchetto, G. Morello, M. Damiano, K. Karpouzas, G. Tinetti, L. K. McKemmish, J. Tennyson, and S. N. Yurchenko. A Population Study of Gaseous Exoplanets. , 155(4):156, April 2018. doi: 10.3847/1538-3881/aaaf75.

Kai Hou Yip, Quentin Changeat, Nikolaos Nikolaou, Mario Morvan, Billy Edwards, Ingo P. Waldmann, and Giovanna Tinetti. Peeking inside the Black Box: Interpreting Deep Learning Models for Exoplanet Atmospheric Retrievals. *arXiv e-prints*, art. arXiv:2011.11284, November 2020.

Kai Hou Yip, Quentin Changeat, Ahmed Al-Refaie, and Ingo Waldmann. To Sample or Not To Sample: Retrieving Exoplanetary Spectra with Variational Inference and Normalising Flows. *arXiv e-prints*, art. arXiv:2205.07037, May 2022a. doi: 10.48550/arXiv.2205.07037.

Kai Hou Yip, Ingo P. Waldmann, Quentin Changeat, Mario Morvan, Ahmed F. Al-Refaie, Billy Edwards, Nikolaos Nikolaou, Angelos Tsiaras, Catarina Alves de Oliveira, Pierre-Olivier Lagage, Clare Jenner, James Y-K. Cho, Jeyan Thiyagalingam, and Giovanna Tinetti. ESA-Ariel Data Challenge NeurIPS 2022: Inferring Physical Properties of Exoplanets From Next-Generation Telescopes. *arXiv e-prints*, art. arXiv:2206.14642, June 2022b.

Tiziano Zingales and Ingo P. Waldmann. ExoGAN: Retrieving Exoplanetary Atmospheres Using Deep Convolutional Generative Adversarial Networks. , 156(6):268, December 2018. doi: 10.3847/1538-3881/aae77c.

Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. *Big data analysis: new algorithms for a new society*, pages 91–114, 2016.