NASA/TM–2018–XXXXX

NASA Frontier Development Lab: Astrobiology Team II

Michael D. Himes
Planetary Sciences Group, Department of Physics, University of Central Florida,

Molly D. O'Beirne
Department of Geology and Environmental Science, University of Pittsburgh,

Frank Soboczenski
School of Population Health and Environmental Sciences, King's College London,

Simone Zorzan,
ERIN Department, Luxembourg Institute of Science and Technology

Mentors:

Atılım Güneş Baydin, Adam Cobb
Department of Engineering Science, University of Oxford, UK,

Daniel Angerhausen,
Center for Space and Habitability (CSH), Universitat Bern, Sidlerstrasse 5, 3012 Bern, Switzerland,

Giada N. Arney,
NASA Astrobiology Institute, Virtual Planetary Laboratory Team, Seattle, Washington, USA, Planetary Systems Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA,

Shawn D. Domagal-Goldman,
NASA Astrobiology Institute, Virtual Planetary Laboratory Team, Seattle, Washington, USA Planetary Environments Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

August 2018

The NASA Frontier Development Lab is a public / private research partnership between NASA, the SETI Institute and leaders in commercial AI and private space.

Hosted in Silicon Valley by the SETI Institute, the NASA FDL is an applied artificial intelligence research accelerator developed in partnership with NASA's Ames Research Center. Founded in 2016, the NASA FDL aims to apply AI technologies to challenges in space exploration by pairing machine learning expertise with space science and exploration researchers from academia and industry. These interdisciplinary teams address tightly defined problems and the format encourages rapid iteration and prototyping to create outputs with meaningful application to the space program and humanity.

SETI Institute
189 N. Bernardo Ave Suite 200
Mountain View, CA 94043

NASA Ames Space Portal
556 Edquiba Rd
Mountain View, CA 94043

Private Partners:

Google Cloud is a leading cloud services provider, powered by the resources and technology of Google. Google Cloud provided funding and the significant compute resources required by the project.

For more information about the NASA Frontier Development Lab, see the following:

- FDL home page at `http://www.frontierdevelopmentlab.org`
- For media, please get in contact with Darryl Waller at darryl.e.waller@nasa.gov

NASA/TM–2018–XXXXX

NASA Frontier Development Lab: Astrobiology Team II

Challenge:

From Biohints to Confirmed Evidence of Life: Possible Metabolisms Within Extraterrestrial Environmental Substrates

Michael D. Himes
Planetary Sciences Group, Department of Physics, University of Central Florida,

Molly D. O'Beirne
Department of Geology and Environmental Science, University of Pittsburgh,

Frank Soboczenski
School of Population Health and Environmental Sciences, King's College London,

Simone Zorzan,
ERIN Department, Luxembourg Institute of Science and Technology

Atılım Güneş Baydin, Adam Cobb
Department of Engineering Science, University of Oxford, UK,

Daniel Angerhausen,
Center for Space and Habitability (CSH), Universitat Bern, Sidlerstrasse 5, 3012 Bern, Switzerland,

Giada N. Arney,
NASA Astrobiology Institute, Virtual Planetary Laboratory Team, Seattle, Washington, USA,
Planetary Systems Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA,

Shawn D. Domagal-Goldman,
NASA Astrobiology Institute, Virtual Planetary Laboratory Team, Seattle, Washington, USA
Planetary Environments Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

SETI Institute

189 N. Bernardo Ave Suite 200
Mountain View, CA 94043

NASA Ames Space Portal

556 Edquiba Rd

Mountain View, CA 94043

August 2018

Acknowledgments

We would like to thank our amazing mentors for their fantastic support and guidance; Geronimo Villanueva for offering extensive support in setting up PSG on Google Cloud; Massimo Mascaro for his help in utilizing the Google Cloud Platform for this project; Sara Jennings, Shyla Spicer, and James Parr for their countless hours spent organizing the NASA Frontier Development Lab program to ensure everything ran smoothly; the SETI Institute for hosting us and providing the coffee necessary for Frank to function; and the rest of the FDL participants for their friendship and support over the course of the program.

Abstract

Over the past decade, the field of exoplanets has shifted from their detection to the characterization of their atmospheres. Atmospheric retrieval, the inverse modeling technique used to determine an atmosphere's temperature and composition, is both time-consuming and compute-intensive, requiring complex algorithms that generate thousands to millions of atmospheric models, compare the model to the observational data, and build a posterior distribution that gives the most probable value and uncertainty for each model parameter. For rocky, terrestrial planets, the retrieved atmospheric composition can give insight into the surface fluxes of gaseous species necessary to maintain the stability of that atmosphere, which may in turn provide insight into the geological and/or biological processes active on the planet. These atmospheres contain many molecules, some of which are biosignatures, or molecules indicative of biological activity. Runtimes of traditional retrieval models scale with the number of model parameters, so as more and more molecular species are considered, runtimes can become prohibitively long. Machine learning (ML) offers a unique way to reduce the time to perform a retrieval by orders of magnitude, given a sufficient data set to train with. Here we present the Intelligent exoplaNet Atmospheric RetrievAl (INARA) code, the first ML retrieval model for rocky, terrestrial exoplanets, and a data set of 3,000,000 spectra of synthetic rocky exoplanets generated using the NASA Planetary Spectrum Generator.

1 Introduction

In recent years, exoplanetary science has shifted from the detection of exoplanets to the characterization of their atmospheres via inverse modeling techniques ("retrievals"). These involve computationally-expensive and time-consuming algorithms, usually in a Bayesian framework (e.g., Feroz and Hobson 2008, ter Braak and Vrugt 2008), that produce many forward models of varying atmospheric structure and composition, compare the spectra in the bandpasses corresponding to observations, and find a best-fit model with associated uncertainties in each model parameter (Madhusudhan 2018).

Accurate constraints on parameters requires spectroscopic observations spanning a wide range of wavelengths. Presently, such techniques can be reasonably applied to only a small subset of the total number of observed exoplanets, as most have only been observed by one or two instruments using wide-band photometry. New flagship telescopes such as the James Webb Space Telescope and the Extremely Large Telescope will see first light in the coming years, allowing for unprecedented characterization of exoplanetary atmospheres. Thus, it is imperative that a fast and accurate retrieval framework exists for this purpose. Machine learning (ML) offers a unique method that can quickly invert computationally-demanding processes once trained. Currently, two ML retrieval algorithms, HELA (Márquez-Neila *et al.* 2018) and ExoGAN (Zingales and Waldmann 2018), constitute the state-of-the-art, but they are limited in scope as they only apply to hot Jupiters with fewer than a

handful of molecular species.

To date, characterization of rocky/terrestrial exoplanetary atmospheres has not been possible due to the sensitivity limits of existing telescopes. This class of exoplanets is particularly intriguing as it offers the best opportunity to remotely detect life via biosignatures, combinations of molecules indicative of life. The detection of such biosignatures are driving future telescope designs such as the Large UltraViolet, Optical, InfraRed telescope (LUVOIR) and the Habitable Exoplanet Explorer (HabEx) to enable the inference of whether or not biological processes are necessary to explain the deduced atmospheric characteristics. Here we present INARA (Intelligent exoplaNet Atmospheric RetrievAl), the first ML atmospheric retrieval algorithm for rocky/terrestrial exoplanets using a convolutional neural network trained on 100,000 synthetic planets covering a range of 28 stellar and planetary parameters. We will expand this to a data set totaling over 3,000,000 planets in the near future.

2 Background

The study of exoplanets is based on two approaches: direct and indirect detection. Direct methods rely on the use of sensors observing a signal emitted by the planet surface and atmosphere, while indirect methods rely on the effect that the planet's presence exerts on its host star. Among indirect observations, the transit method is one of the most promising; it consists of detecting stellar signal variations due to the transit of the planet as it passes in front of the star as viewed by the observer. This allows for direct measurement of the apparent planetary radius as a function of wavelength. Information related to the atmospheric structure and composition of the planet can be deduced from this apparent radius variability as the molecular species present in the atmosphere at the day-night terminator absorb differing amounts of stellar flux at particular wavelengths (Crossfield 2015).

Another observation technique uses a coronagraph to suppress the stellar emission such that direct imaging of the exoplanet is possible (Fujii *et al.* 2018). While the transit method directly measures planetary radius to infer atmospheric composition, coronagraphic observations directly measure the emission from the planet, which is a combination of reflection of host star emission and emission from either the surface of the planet (rocky bodies) or the deep atmosphere (gaseous bodies). This measured emission is due to the atmospheric composition and, to a greater extent for hot, gaseous bodies, the temperature structure. With enough measurements across a broad wavelength range, the atmospheric composition and temperature structure can be determined with some degree of uncertainty; this process is known as atmospheric retrieval (Madhusudhan 2018). By determining the atmospheric composition of a rocky/terrestrial exoplanet, we are then able to begin the process of deducing whether or not life may exist on an exoplanet. This is because the atmospheric composition (and ultimately the planetary spectrum we observe) of an exoplanet may be influenced by gas fluxes from biological sources (as outlined in

Figure 1).

To date, the bulk of retrieval methods consist of assuming an atmosphere with a set of molecules at a particular pressure–temperature profile, generating the spectrum resulting from this atmosphere (forward model), binning the spectrum according to the instruments and filters used to observe the exoplanet, and comparing the result to the measured data. This is repeated thousands to millions of times over some large parameter space, typically using a Bayesian method such as Markov chain Monte Carlo (MCMC; ter Braak and Vrugt 2008) or nested sampling (Skilling 2004) to yield a posterior distribution. While these methods can find constraints on molecular abundances or the pressure–temperature profile, they are time-consuming and require significant computational resources to do so.

In previous studies, random forests (Márquez-Neila *et al.* 2018) and generative adversarial networks (GANs; Zingales and Waldmann 2018) have been used to perform atmospheric retrievals with ML. These two approaches, however, were limited to atmospheric retrievals of hot Jupiters with four or fewer molecular species considered. Furthermore, they only consider isothermal temperature profiles (although in reality these planets have non-isothermal temperature structures), and their spectra are constrained to low resolutions. Thus, there are significant improvements that can be made for ML retrieval models.

3 Methods

3.1 Tools, Compute and Software Environment

We used `Python` as our main programming language to develop INARA in combination with `PyTorch`, an optimized ML library for deep learning utilizing both GPUs and CPUs. In addition, we also used `TensorFlow, Keras`, and `Scikit-Learn` via `Jupyter` notebooks to evaluate specific models and their performance. The NASA Goddard Planetary Spectrum Generator (PSG; Villanueva *et al.* 2018) was the core of our spectrum generation coupled with `pypsg`, a `Python` package to generate input parameters for PSG. INARA utilizes `pypsg` to send models to PSG for spectra generation.

INARA can be run for data generation, ML model training and model inference. The code runs locally on a system depending on installed software requirements[1] or in form of a `Docker` container. The latter option requires no installation of software dependencies or ML frameworks as the container can be deployed to run on all `Docker`-supported operating systems.

A version of NASA's PSG was transferred to `Google Cloud` and is now available to be instantiated as a Virtual Machine (VM) on the cloud platform. INARA can

[1] An environment YAML Ain't Markup Language (YML) file is available at the INARA GitLab repository

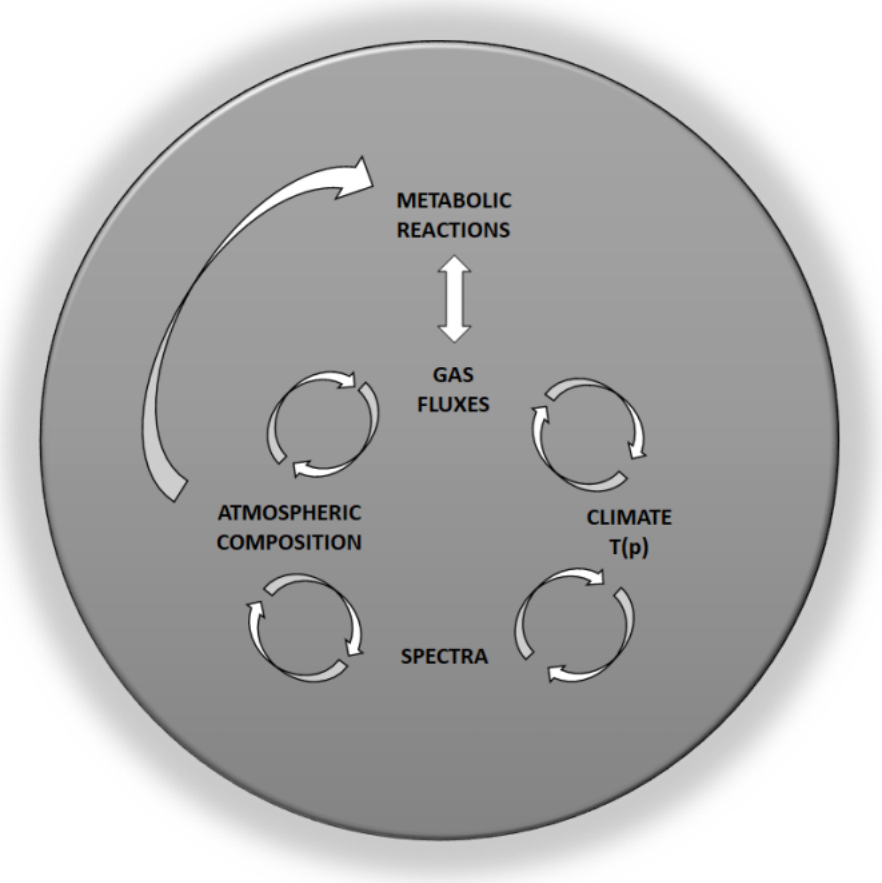

Figure 1: Schematic overview of the problem. Planetary spectra are a product of the pressure–temperature profile and atmospheric composition of an exoplanet. The atmospheric composition is influenced by gas fluxes resulting from geological processes, biological processes, or a combination of the two. Biological activity may be able to be determined based on the inferred atmospheric composition, after potential geologic contributions have been ruled out. If biology is present, we can potentially determine the metabolisms occurring on the exoplanet.

instantiate multiple such PSG instances via a single `Python` command[2]. Figure 2 displays the described software landscape.

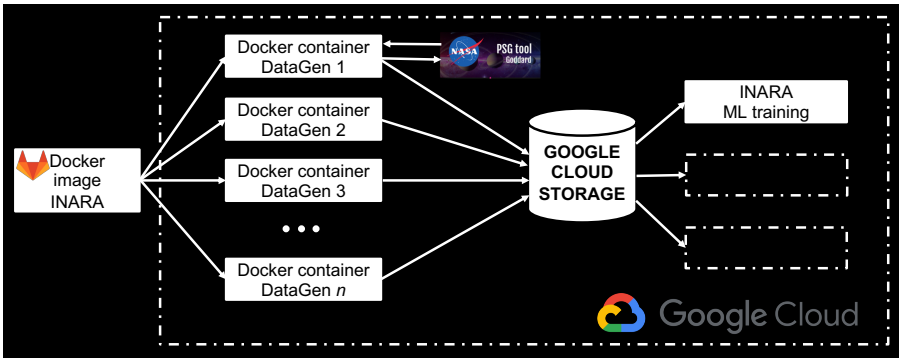

Figure 2: Overview of the entire implementation landscape. Once instantiated as a Docker container in the cloud, INARA communicates with a designated PSG VM to generate the data and store it in the cloud. INARA Docker containers are then able to read the data back in for ML training and inference purposes.

`Google Cloud` also served as the main computational platform for data generation, model training and inference. Via `Google Cloud`, we were able to instantiate ∼2000 VMs (groups of 16 INARA instances connected to one PSG node) for data generation. Figure 3 presents an overview of the data generation architecture. Although currently configured for `Google Cloud`, INARA can be used in combination with any cloud computing system.

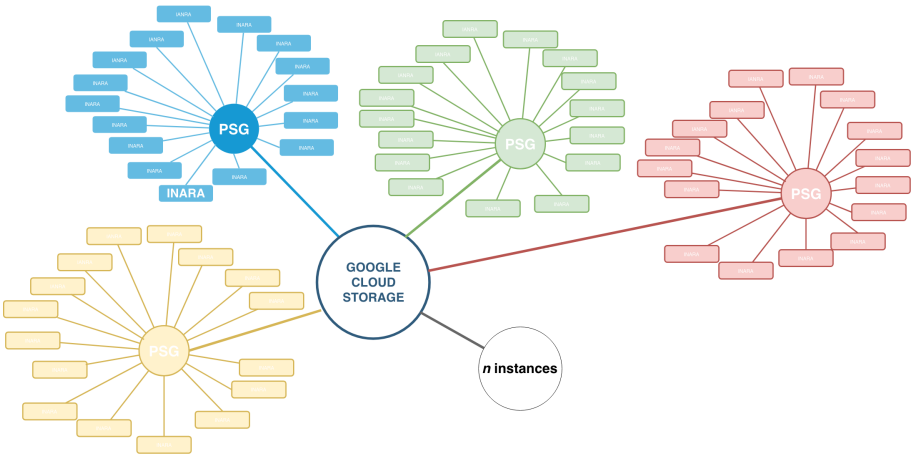

Figure 3: Structure of the instantiated INARA docker container and PSG VMs that act as nodes connected to the cloud storage.

INARA stores the generated data in blob form on `Google Cloud` and reads the data on-the-fly back into INARA for ML purposes. Trained ML models and predictions are also saved automatically to the cloud.

[2]See the documentation at https://gitlab.com/frontierdevelopmentlab/astrobiology/inara

3.2 Generation of Planetary Spectra via NASA Goddard's Planetary Spectrum Generator

In order to train our ML models, we first had to generate a data set encompassing a large parameter space. We utilized PSG to generate spectra based on a given planetary system model. In this section, we describe our choices for the parameter space considered by our model.

3.2.1 System Parameters

We consider F, G, K, and M main sequence stars. Kurucz stellar models are used for G, K, and M types, while F types are simulated as a blackbody as PSG lacks an F-type model at the time of data generation. Stellar radii and temperatures are randomly selected from uniform ranges based on Boyajian *et al.* (2012) and Boyajian *et al.* (2013). The semi-major axis of the planet is randomly selected from that of an optimistically-habitable Venus to an optimistically-habitable Mars scaled according to the host star's radius and temperature, as defined in Kopparapu *et al.* (2013). The distance of the system is randomly selected from 1.3 to 15 pc for coronagraphic observations (F, G, and K types) and from 5 to 25 pc for transit observations (M types). These ranges were chosen for observational reasons. For the minima, the closest exoplanet is Proxima Centauri b at 1.3 pc (coronagraphic), and detectors will saturate quickly at 5 pc if staring at the target (transit). For the maxima, noise necessitates many hours of observations, reducing the likelihood that rocky worlds will be studied at these distances given there are likely better targets at closer distances.

Observations are simulated using a 15 m space telescope with a resolution of 1900 covering 0.2 to 2 µm with low read noise (1 e^-/pixel) and dark current (0.001 e^-/s). The coronagraph's inner working angle is 2 λ/D. These are based on the LUVOIR-A design concept but with a much higher resolution to allow the use of both high- and low-resolution data for experimentation.

Simulated observations have similar noise characteristics; the nominal observations are 8 hours on Earth at 5 pc when it is at its greatest separation from the Sun for coronagraphic observations, and 8 hours on TRAPPIST-1e, which is at a distance of 12.1 pc, for transit observations. The observing times are modified by the squared distance ratio, such that a planet at twice the distance of the nominal case will be observed for four times longer. We also simulate the spectra out to 640 µm to provide a ground-truth spectrum across a wide wavelength range which may be used for other investigations beyond what is considered here.

3.2.2 Planetary Parameters

Planetary radii are randomly selected from a uniform range spanning 0.5 to 1.6 $R_\oplus$ due to planets with radii >1.6 $R_\oplus$ containing a significant gas mass fraction (Rogers 2015). Planetary masses are determined using the mass-radius relation of Sotin *et al.* (2007) with a random uniform $\pm$ 2% factor to account for model inaccuracies. Using

a rough approximation of the 'cosmic shoreline' model of Zahnle and Catling (2017), we throw out any planet that is likely unable to hold onto an atmosphere. We draw the surface pressure from a truncated normal distribution bound by 0.1 and 90 bars with a mean of 1 bar and a standard deviation of 2.5 bars. The pressure–temperature profile is determined from the parameterized formulation of Line *et al.* (2013). Four of these parameters, which govern the shape of the profile, are selected from a random log-uniform distribution, while the fifth, which shifts the profile in temperature, is selected from a truncated normal distribution. The ranges for these parameters were chosen such that a planet in an Earth-like orbit can vary in temperature by a few hundred Kelvin. As a result, both Earth-like planets and frozen Mars-like planets are within our parameter space. Our hottest case results in temperatures slightly below that of Venus. We allow the tropopause to vary between 0.023 and 0.23 bars due to the finding of Robinson and Catling (2014) that the planets in the solar system all have a tropopause around 0.1 bars.

We consider 12 molecules: H_2O, CO_2, O_2, N_2, CH_4, N_2O , CO, O_3, SO_2, NH_3, C_2H_6, and NO_2. Table 1 summarizes the upper bound on the random uniform distribution used to draw a value for each molecule. Following the selection of these values, they are normalized such that they sum to 1; this yields the molar mixing ratio of each gas. As a result, it is possible for some of the trace species to exceed their "upper limit" if a high value is generated for the trace gas and low values are generated for the main constituents (CO_2, O_2, N_2). Our upper limits (some of which may seem unrealistic) were selected to encapsulate as many atmospheres as possible while also restricting the parameter space to cases that are more likely for terrestrial planets. CO_2 and N_2 have been observed in excess of 90% in atmospheres in the solar system, hence their upper limits. O_2 is unlikely to exist at such high abundances but does exist at a substantial amount in Earth's atmosphere; thus, we assume a large amount to allow for a wide range of possibilities. On Earth, H_2O content in the air can be a few percent in tropical regions; we allow for up to 10% to allow for conditions slightly more extreme than those found on Earth. O_3, a photochemical product of O_2, has been shown to be at most $\sim$1% the amount of O_2 in prebiotic Earth-like atmospheres (Domagal-Goldman *et al.* 2014). We impose a similar limit for C_2H_6 as it is a photochemical product of CH_4. N_2O can act as a strong biosignature in certain situations, such as that of Earth, but must be considered in the context of the host star and the absence of gases indicative of abiotic N_2O production (Schwieterman *et al.* 2018). The upper limits on other trace gases are arbitrarily determined to cover many cases for thin atmospheres.

All gases begin with a uniform vertical abundance profile; H_2O and NH_3 are modified by calculating the saturation vapor pressure (SVP) at each layer in the atmospheric model and assuming that all excess vapor pressure condenses out into clouds. The SVPs are calculated using the Antoine equation using the available NIST data; H_2O covers a range of 255.9 to 573 K, and NH_3 covers a range of 164 to 371.5 K. For an Earth-like planet, this method leads to the abundance of H_2O decreasing as altitude increases up to some cold trap, above which clouds no longer form. While the cloud mixing ratios are calculated, clouds are ignored in our simulations due

Table 1: Upper limits on random uniform distribution draw for each molecule

Molecule	H_2O	CO_2	O_2	N_2	CH_4	N_2O	CO	O_3	SO_2	NH_3	C_2H_6	NO_2
Upper limit	0.1	1.0	1.0	1.0	0.1	0.02	0.02	0.01*O_2	0.02	0.01	0.01*CH_4	2e-5

to the computational burden, as even poor modeling efforts increase computational time by a factor of ∼50.

3.3 Data Set INARA DS1

The generated data set (INARA DS1) has three million planetary spectra that consist of a 1-dimensional numerical data vector per planetary spectrum with a total length of 15,346. Table 2 shows the content of each index in a vector. The data files (CSV) encompass four vectors per file (∼750,000 files in total). The data set is also provided in *numpy* standard binary file format (NPY) for faster access.

Table 2: Structure of the data vector and its contents

Position	Content
0	Stellar Type
1	Stellar Temperature
2	Stellar Radius
3	Distance from Earth to the planetary system
4	Semi-major axis of the exoplanet
5	Radius of the exoplanet
6	Density of the exoplanet
7	Surface pressure of exoplanet
8 – 12	Parameters to describe the pressure–temperature profile
13	Surface temperature of the planet
14 – 25	Molar mixing ratio of each molecular species (H_2O, CO_2, O_2, N_2, CH_4, N_2O , CO, O_3, SO_2, NH_3, C_2H_6, and NO_2)
26	Average molecular weight of the atmosphere
27	Surface albedo of the exoplanet
28 – 15,373 (15,346 in total)	Wavelengths of spectral data (µm)
15,374 – 30,719 (15,346 in total)	Total (star + planet) observed spectrum (erg/s/cm^2)
30,720 – 46065 (15,346 in total)	Noise model
46,066 – 61411 (15,346 in total)	Stellar spectrum
61,412 – 76757 (15,346 in total)	Planetary spectrum

3.4 Machine Learning

Machine learning is a subdomain of artificial intelligence (AI) and the science of building algorithms that allow computer systems to act autonomously without being explicitly programmed. During the last decade, a new wave of interest in machine learning in the form of deep learning (Goodfellow *et al.* 2016) and advances in computer vision (Krizhevsky *et al.* 2012) and natural language processing (Graves

et al. 2013) have provided us with accurate search algorithms, advances in self-driving cars, speech recognition and synthesis, and new ways of detecting diseases. *Supervised learning*, an area within ML, aims to minimise a loss function or in other words the error E between the known values $x = (x_1, x_2, x_3, ..., x_n)$ and the predicted values $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, ..., \bar{x}_n)$ as displayed in:

$$E_{Tr}(\bar{x}_i, x_i) \tag{1}$$

INARA is positioned in the *supervised learning* domain as we know our input parameters x_i that we used to generate the planetary spectra via PSG. The spectra then contain the atmospheric abundances o_i that are used for training the ML models and which combined form our data set $N_{Tr+V+Ts} = \{x_i, o_i\}$. N consists the subsets Tr (100,000 data points) used for training, the validation set V (10,000 data points) and the test set Ts (7,710 data points).

We define our retrieval model, f_θ, as function that can accurately infer the parameters x_i for a given observation o_i:

$$x_i = f_\theta(o_i) \tag{2}$$

This function is a deep neural network containing many parameters θ, which are learned by defining the training error E_{Tr}, which we minimise by optimising the parameters through backpropogation:

$$E_{Tr} = \sum_{i=1}^{Tr} E(\bar{x}_i = f_\theta(o_i), x_i) \tag{3}$$

To avoid overfitting and limiting the generalizability of our model, we used validation steps during training to monitor model performance. We employed *early stopping* with a buffer of 15 hits before our training would automatically stop. This means each validation step checks if the training loss decreases. If it increases 15 times in row our training would stop.

$$E_V = \sum_{i=1}^{V} E(\bar{x}_i = f_\theta(o_i), x_i) \tag{4}$$

3.4.1 Model evaluation & experiments

Once we generated the data, we performed a model grid search in order to determine the best performing ML model architecture. We started by applying a linear regression model to see initial model performance before moving to feed-forward networks in several configurations (i.e., different number of neurons in layers) and convolutional neural networks (CNNs) (e.g. Figure 4). A CNN utilizes layers and filters that compress the data on each convolution to local features (LeCun *et al.* 1998). In addition the model grid search also included modifications on different activation functions as well as some hyper parameter tuning.
We tested over 68 combinations of different architectures, learning rates (from 0.0001 to 0.01), activation functions (Tanh, Softmax ReLU, ELU, Linear), and optimization

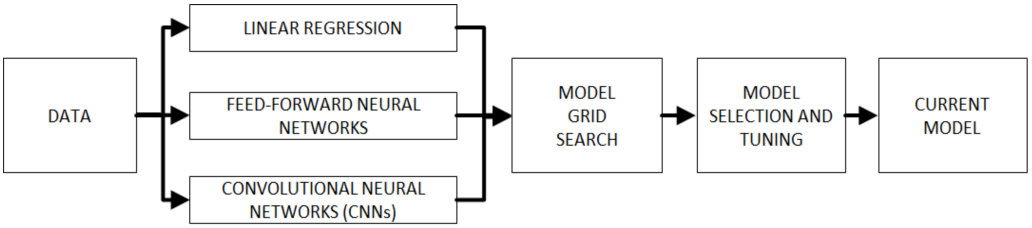

Figure 4: Evaluated ML architectures. We started with a simple linear regression (single layer), a standard feed forward and a convolutional neural network. Model grid search was led evaluating learning rates of 0.0001, 0.001 and 0.01, ADAM, SGD, ADAdelta and RMSProp optimizers, and as activation functions Tanh, Softmax, ReLU and linear. Once we found a reasonably performing model we selected the model used for further steps.

algorithms (ADAM, SGD, ADAdelta, RMSProp) toward the selection of a model with good performance (see Section 4). We evaluated model performance by considering the loss value at the end of training. Each model used 100,000 planets for learning (Tr), 10,000 for validation (V), and 7,710 for testing (Ts). The model training was set to 64 epochs (how many times the ML algorithm has seen the entire data set for training) for all evaluated training runs.

3.4.2 Producing predictive distributions over abundances by using the Monte Carlo dropout approximation

Dropout is a common regularization technique in neural networks to prevent overfitting and allow for a more generalizable model (Hinton *et al.* 2012). However, it has recently been shown that applying dropout at both training and test time is equivalent to making a variational approximation to the posterior distribution over the network weights (Gal and Ghahramani 2016). Each dropout mask removes a certain proportion, p, of NN weight connections by setting them to zero during a forward pass. Therefore, multiple forward passes with different dropout masks for the same input gives a set of predictive samples that build a predictive distribution. Through implementing dropout both at training and test time, we are effectively sampling from the posterior over weights of the network. This distribution over the weights enables us to approximate a predictive distribution over the abundances for each planet, which we represent as the output samples of the network for a given input. Figure 6 shows the mean prediction compared to the true value for each planet in the test set, and Figure 7 show a predictive distribution resulting from dropout for a specific test planet.

4 Results

As mentioned above, we trained a ML retrieval model on 117,710 model planets, with 100,000 for training (Tr), 10,000 for validation (V), and 7,710 for testing (Ts). We limited our data generation to coronagraphic observations in the interest of our available resources and time. We explored a variety of model architectures ranging in

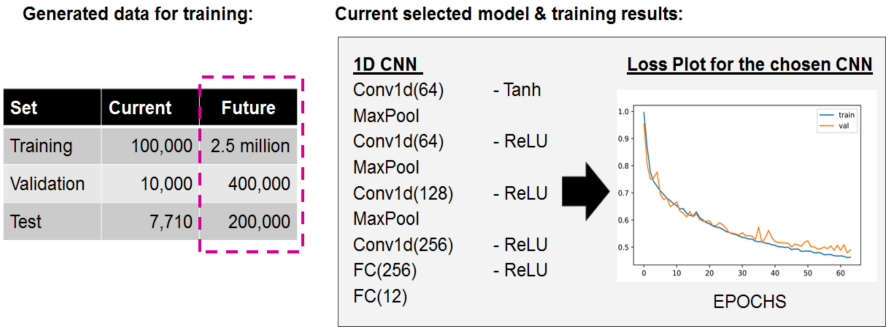

Figure 5: Available and used data in ML phases (left) and model architecture (on the right). Loss plot over the 64 epochs, with details on the relation between train and validation loss function.

complexity from linear regression and feed-forward neural networks to convolutional neural networks (CNNs). We present results from the best performing model, a 1D CNN with the following configuration: Conv1d(64) - Tanh - MaxPool - Conv1d(64) - ReLU - MaxPool - Conv1d(128) - ReLU - MaxPool - Conv1d(256) - ReLU - FC(256) - ReLU - FC(12) - training loss (0.42) - validation loss (0.49) - 64 epochs. The validation loss (orange line in Figure 5) compared to the training loss allows to ensure that the model is not trained to overfit to the limited data set at hand and thus maintain a generalized solution.

4.1 INARA Performance

We use the model described in the previous section to predict the parameters of 1,000 planetary spectra. The results of these predictions for H_2O, CO_2, O_2, N_2 and CH_4 are shown in Figure 6, where each dot represents the average of 600 runs of our model with dropout, for each of the 1000 predicted planetary parameters.

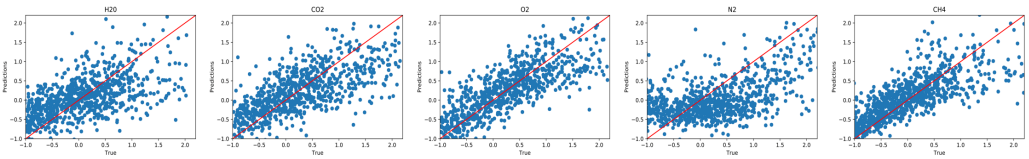

Figure 6: INARA prediction performance for the logarithm of the normalized abundance of H_2O, CO_2, O_2, N_2 and CH_4 across 1000 test planets. We normalize values by subtracting the mean and dividing by the standard deviation across all planets. Each dot is the average of 600 runs with dropout for a single planet. Predicted vs. true values are plotted, with the diagonal line indicating perfect correspondence.

A perfect correlation is represented by the red diagonal line, and the spread of the predictions about this line indicate the accuracy. Note, however, that dropout produces a predictive distribution, so the true value may still fall within the predicted distribution despite disagreement between the mean prediction and the true value. Errors are summarized in Table 3. Low mean squared error (MSE) values indicate better predictions of the true value.

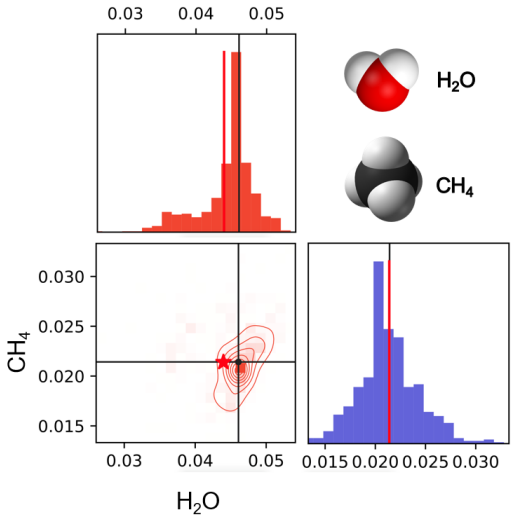

Figure 7: Detailed results of INARA in predicting one random planet's values for H_2O and CH_4. The distribution is obtained by 600 predictions with dropout where nodes were randomly removed from the network. The black lines represent INARA's median prediction, and the red lines and star represent, respectively, the real value in the dimension of H_2O and CH_4, and in the two-dimensional space with one molecule per axis.

Table 3: Reported error for five retrieved molecules

Error	H_2O	CO_2	O_2	N_2	CH_4
MSE	3.43e-4	1.02e-2	7.00e-3	2.05e-2	1.93e-4
$\pm 2\sigma$	2.28e-3	3.53e-2	2.59e-2	5.21e-2	1.07e-3

A more detailed representation of INARA's results is presented in Figure 7, which shows the predictive distributions for CH_4 and H_2O for a random planet among those simulated. Note that the true value, indicated by the red star in the bottom-left plot and the red line in the top-left and bottom-right plots, falls within the predictive distribution for both parameters. For results for more than two molecules, see Figures A8, A9, and A10.

4.2 Links to Code Repositories

The entire developed code palette is available open source at the *Frontier Development Lab* GitLab repository[3] in the astrobiology section. This includes but is not limited to:

- **pypsg** *Python interface for PSG*
  https://gitlab.com/frontierdevelopmentlab/astrobiology/pypsg
- **INARA** *Intelligent exoplaNet Atmospheric RetrievAl*
  https://gitlab.com/frontierdevelopmentlab/astrobiology/inara

[3]https://gitlab.com/frontierdevelopmentlab

- **Data set INARA DS1** *3 million planetary models in numpy (NPY) and comma-separated values (CSV) format.*
  Currently hosted on Google Cloud; public release scheduled for November 2018

The INARA repository also provides *Jupyter Notebooks* for data exploration, model training and predictions. This includes the notebook that was used to produce the results presented in this document.

5 Discussion

INARA accurately performs an atmospheric retrieval from an observed planetary spectrum in a matter of seconds. This outperforms the traditional approaches by several orders of magnitude. Other existing ML approaches provide comparable performance, but they are limited to a much narrower set of parameter and atmospheric molecules (see Table 4).

Table 4: Comparison of atmospheric retrieval methods

Method	CPU Time	# of Molecules Retrieved
Traditional	Hundreds of hours	User-specified
ExoGAN[1]	Minutes	4
HELA[2]	Seconds	3
INARA	Seconds	12

[1] Zingales and Waldmann (2018)
[2] Márquez-Neila *et al.* (2018)

Our present data set is the largest collection of rocky planet spectra to date. The analytical temperature–pressure profile we adopted (see Section 3.2) allows the prediction of those parameters as well. While the simulated observed spectra span 0.2 to 2 µm, our model still managed to predict the abundance of N_2 remarkably well despite it having no substantial features in that range, highlighting the versatility and the power of ML to deduce relationships from data.

The adoption of Monte Carlo dropout (Gal and Ghahramani 2016) in our machine learning model provides a predictive distribution (see Fig. 5), which is comparable to the posterior distributions yielded by traditional, Bayesian approaches. This is the first time this technique has been applied to atmospheric retrievals. Further investigation is necessary to determine how this predictive distribution compares to the posterior distributions of traditional methods.

While we obtained good results with our CNN model, chosen because of the high dimensionality and interrelatedness of our data set, our search for the best model is incomplete. CNNs appear to be more efficient at learning features related to molecules than other neural network architectures, however a thorough exploration of different neural network architectures is desirable.

Considering the MSE values in Table 3, the standard deviation is rather high for molecules with convoluted features in the 0.2 to 2 µm range considered. We attribute this to the limited data set used for the present results, and we except the standard deviation to minimize once the model is trained on the complete data set.

Our model is a proof-of-concept approach for our data set. A more detailed data set (i.e., in terms of wavelength, self-consistency, and the presence of clouds/hazes) could be used with INARA to generate more reliable and informative models.

6 Future Work

There are several opportunities for future work, outlined below.

In the near future it is planned to (1) train, validate, and test the ML model on the entire data set of 3 million planets; (2) release the data set through a browsable website; and (3) release the software pipeline with a user-friendly interface.

In the medium-term it is planned to (1) evaluate the generated atmospheres for self-consistency; (2) determine atmospheric gas fluxes from the concentrations retrieved using INARA; (3) determine sources of gas fluxes and the possibility of life within the generated planetary spectra; (4) generate an additional data set of planetary spectra which include the effects of clouds and hazes; and (5) train, validate and test ML model on the generated cloudy/hazy spectra. Our architecture could also be applied to other classes of planets such as hot Jupiters given a sufficient data set.

The generated data set can also be used for planning and designing future telescopes in the search of extraterrestrial life (e.g., determining the lowest resolution of spectral observations needed at various S/N ratios to still be able to deconvolve spectral components). Additionally, it is possible to host a Kaggle[4] competition using our data set to see the best ML model that the community can come up with to perform atmospheric retrievals; we are currently exploring this possibility with `Google Cloud`.

A further opportunity to validate our model is to retrieve on the Virtual Planetary Laboratory spectra of solar system planets and compare the results to the known atmospheric compositions. We will consider these cases once we have trained the model on the complete data set.

We have also generated a spectrum of a cloud-free Earth-like exoplanet (Earth's pressure–temperature and vertical abundance profiles) using PSG. This will be used as a test case for our model trained on our full data set to explore how the model performs on cases that are similar to the generated data set but have differences in the temperature structure and abundance profiles.

[4]http://www.kaggle.com

7 Conclusions

Here we have shown that ML can expedite atmospheric retrievals and perform well for rocky, terrestrial exoplanets when considering many molecules. Our ML retrieval model for rocky planets is the first of its kind, and it is the first neural network retrieval model that generates predictive distributions to mimic the traditional, Bayesian approaches to this problem.

We have provided INARA a ML training and testing framework which is able to make use of and write data in the cloud as well as read/write generated models. The software architecture is modular and flexible, making it a good resource for various ML approaches. With an alternative data generation module, INARA can be easily applied to other scientific topics.

Our work here is a proof of concept to highlight the advances that ML can enable in the physical sciences. The techniques employed here can be extended to numerous other applications where there is some time-consuming modeling process that has only one set of outputs for a given set of inputs. The application of ML to these processes stands to revolutionize how scientists approach these problems.

References

Boyajian, T. S., K. Von Braun, G. Van Belle, H. A. McAlister, A. Theo, S. R. Kane, P. S. Muirhead, J. Jones, R. White, G. Schaefer, *et al.*, Stellar diameters and temperatures. ii. main-sequence k-and m-stars, *The Astrophysical Journal* **757**, 2, 112, 2012.

Boyajian, T. S., K. von Braun, G. van Belle, C. Farrington, G. Schaefer, J. Jones, R. White, H. A. McAlister, A. Theo, S. Ridgway, *et al.*, Stellar diameters and temperatures. iii. main-sequence a, f, g, and k stars: additional high-precision measurements and empirical relations, *The Astrophysical Journal* **771**, 1, 40, 2013.

Crossfield, I. J. M., Observations of Exoplanet Atmospheres, **127**, 941, 2015. 1507.03966.

Domagal-Goldman, S. D., A. Segura, M. W. Claire, T. D. Robinson, and V. S. Meadows, Abiotic ozone and oxygen in atmospheres similar to prebiotic earth, *The Astrophysical Journal* **792**, 2, 90, 2014.

Feroz, F. and M. 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, 2008.

Fujii, Y., D. Angerhausen, R. Deitrick, S. Domagal-Goldman, J. L. Grenfell, Y. Hori, S. R. Kane, E. Pallé, H. Rauer, N. Siegler, K. Stapelfeldt, and K. B. Stevenson, Exoplanet Biosignatures: Observational Prospects, *Astrobiology* **18**, 739–778, 2018. 1705.07098.

Gal, Y. and Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in *international conference on machine learning*, pp. 1050–1059, 2016.

Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016. http://www.deeplearningbook.org.

Graves, A., A.-r. Mohamed, and G. Hinton, Speech recognition with deep recurrent neural networks, in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649, IEEE, 2013.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint arXiv:1207.0580* , 2012.

Kopparapu, R. K., R. Ramirez, J. F. Kasting, V. Eymet, T. D. Robinson, S. Mahadevan, R. C. Terrien, S. Domagal-Goldman, V. Meadows, and R. Deshpande, Erratum:habitable zones around main-sequence stars: New estimates(2013, apj, 765, 131), *The Astrophysical Journal* **770**, 1, 82, 2013.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**, 11, 2278–2324, 1998.

Line, M. R., A. S. Wolf, X. Zhang, H. Knutson, J. A. Kammer, E. Ellison, P. Deroo, D. Crisp, and Y. L. Yung, A Systematic Retrieval Analysis of Secondary Eclipse Spectra. I. A Comparison of Atmospheric Retrieval Techniques, **775**, 137, 2013. 1304.5561.

Madhusudhan, N., Atmospheric retrieval of exoplanets, *Handbook of Exoplanets* pp. 1–30, 2018.

Márquez-Neila, P., C. Fisher, R. Sznitman, and K. Heng, Supervised machine learning for analysing spectra of exoplanetary atmospheres, *Nature astronomy* **2**, 9, 719, 2018.

Robinson, T. D. and D. C. Catling, Common 0.1 bar tropopause in thick atmospheres set by pressure-dependent infrared transparency, *Nature Geoscience* **7**, 1, 12, 2014.

Rogers, L. A., Most 1.6 earth-radius planets are not rocky, *The Astrophysical Journal* **801**, 1, 41, 2015.

Schwieterman, E. W., N. Y. Kiang, M. N. Parenteau, C. E. Harman, S. DasSarma, T. M. Fisher, G. N. Arney, H. E. Hartnett, C. T. Reinhard, S. L. Olson, V. S. Meadows, C. S. Cockell, S. I. Walker, J. L. Grenfell, S. Hegde, S. Rugheimer, R. Hu, and T. W. Lyons, Exoplanet Biosignatures: A Review of Remotely Detectable Signs of Life, *Astrobiology* **18**, 663–708, 2018. 1705.05791.

Skilling, J., Nested sampling, in *AIP Conference Proceedings*, vol. 735, pp. 395–405, AIP, 2004.

Sotin, C., O. Grasset, and A. Mocquet, Mass–radius curve for extrasolar earth-like planets and ocean planets, *Icarus* **191**, 1, 337–351, 2007.

ter Braak, C. J. and J. A. Vrugt, Differential evolution markov chain with snooker updater and fewer chains, *Statistics and Computing* **18**, 4, 435–446, 2008.

Villanueva, G. L., M. D. Smith, S. Protopapa, S. Faggi, and A. M. Mandell, Planetary spectrum generator: an accurate online radiative transfer suite for atmospheres, comets, small bodies and exoplanets, *Journal of Quantitative Spectroscopy and Radiative Transfer* , 2018.

Zahnle, K. J. and D. C. Catling, The cosmic shoreline: The evidence that escape determines which planets have atmospheres, and what this may mean for proxima centauri b, *The Astrophysical Journal* **843**, 2, 122, 2017.

Zingales, T. and I. P. Waldmann, Exogan: Retrieving exoplanetary atmospheres using deep convolutional generative adversarial networks, *arXiv preprint arXiv:1806.02906* , 2018.

Appendix A

INARA Results

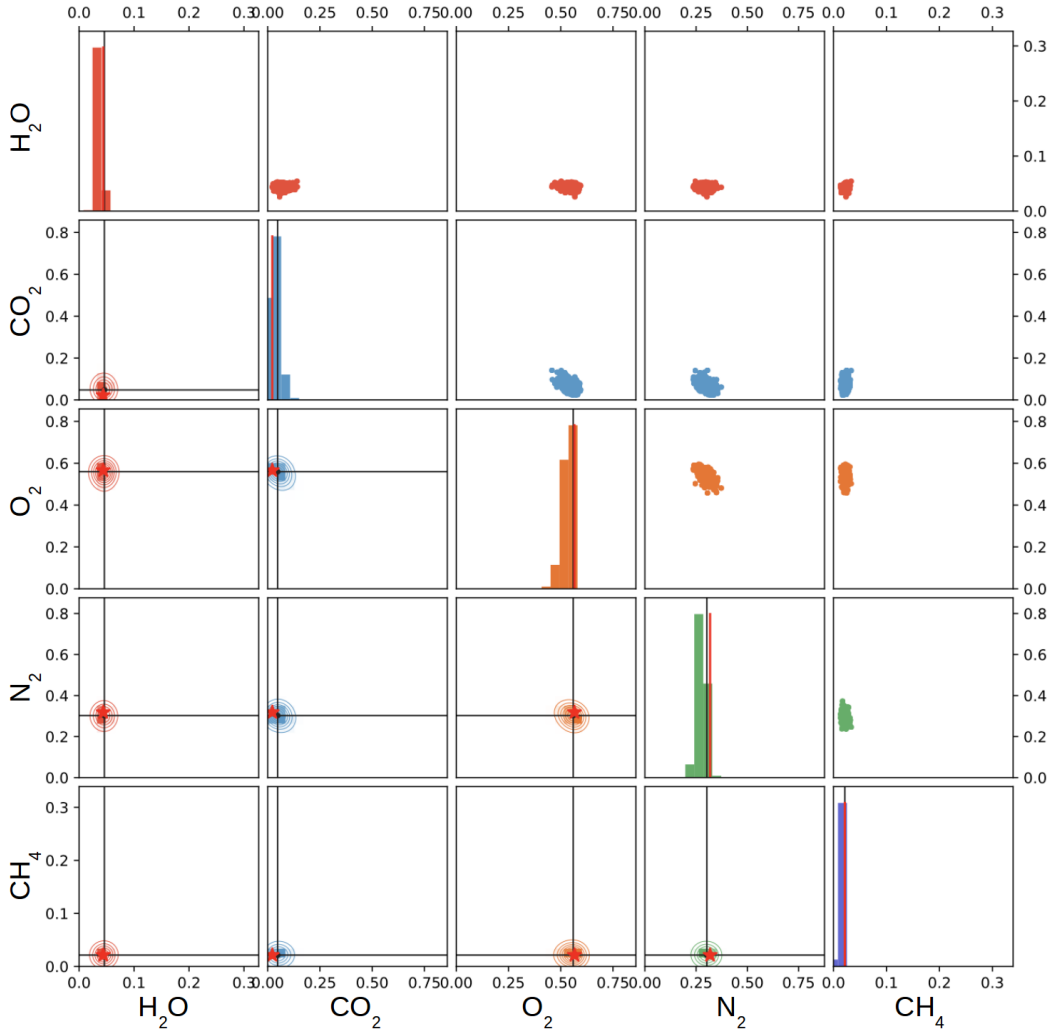

Figure A8: Detailed results of INARA in predicting one random planet's values for H_2O, CO_2, O_2, N_2, and CH_4. The distribution is obtained by 600 predictions with dropout where nodes were randomly removed from the network. Plots along the top-left to bottom-right diagonal show the histogram of the predictive distribution for each molecule. The red line represents the true value, and the black lines represent INARA's median prediction. Plots above the diagonal show the scatter plot of predictions for pairs of molecules. Plots below the diagonal show the 2-D histograms of these scatter plots for each combination of molecules. The red star represents the true value, and the black cross represents INARA's median prediction.

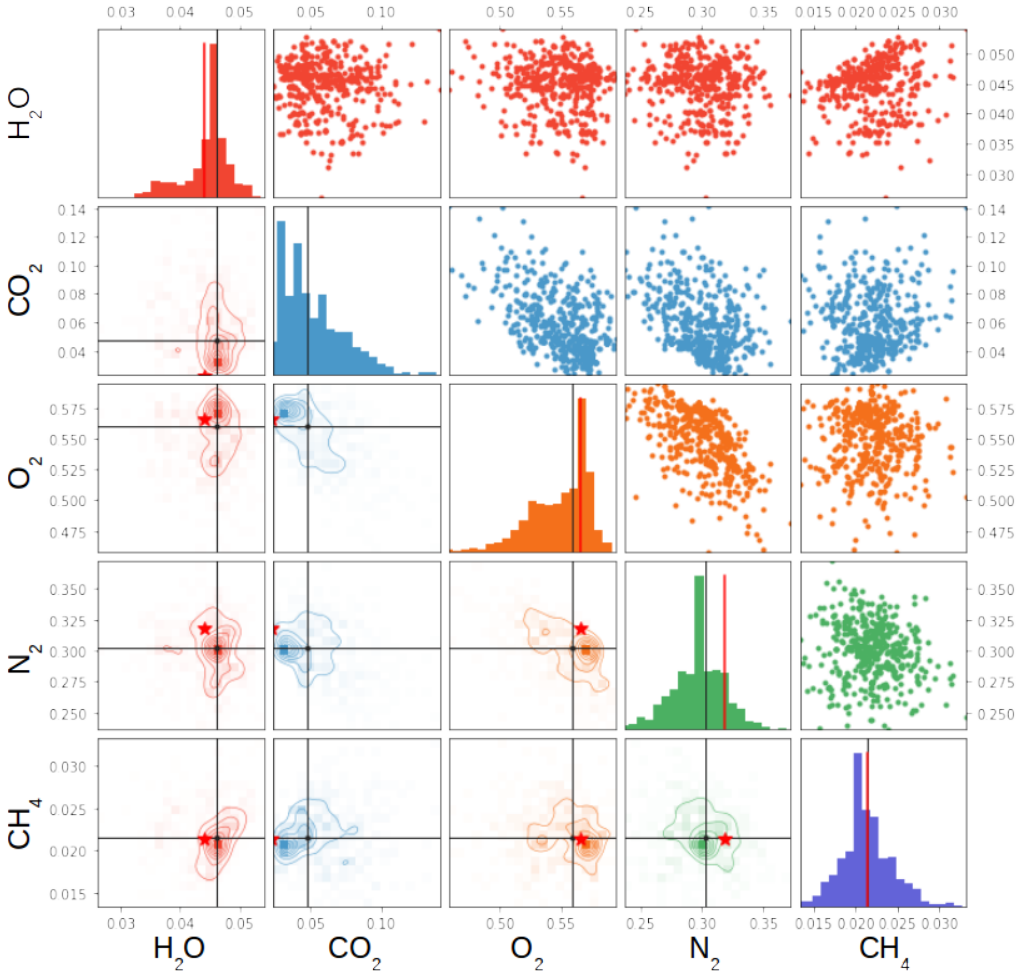

Figure A9: Same as Figure A8, but zoomed in around each distribution.

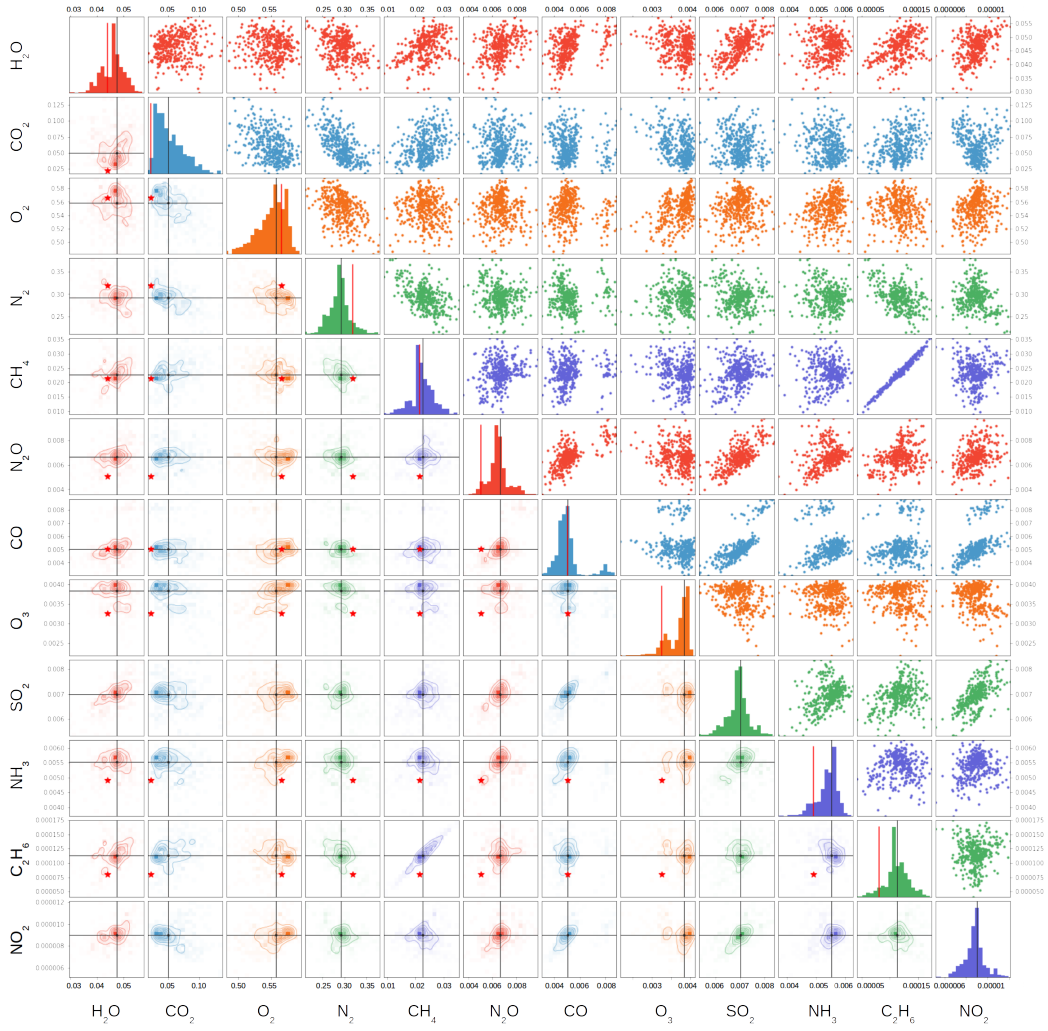

Figure A10: Same as Figure A8, but for all 12 molecules considered in the atmospheric model.

Appendix B

INARA Features

B.1 INARA command-line arguments

The following command-line arguments show the wide range of options and capabilities currently included in INARA:

- **datagen_dir** (default = train); directory for saving generated data
- **train_dir** (default = train); directory for trained models
- **valid_dir** (default = valid); directory for validation data
- **test_dir** (default = test); directory for test data
- **root**; file system directory (local file system or `Google Cloud` storage bucket
- **datagen_files** (default = 4); number of files to generate
- **datagen_planets_per_file** (default = 8); number of planets per file to generate
- **train_epochs** (default = 5); number of epochs to train
- **train_early_stop_tolerance** (default = 15); number of times the validation loss is allowed not to improve before stopping the training
- **seed** (default = 5); random number seed
- **minibatch_size** (default = 64); Size of training minibatches
- **learning_rate** (default = 0.0001); learning rate for training
- **weight_decay** (default = 1e-5); L2 regularization for training
- **mode** (default = datagen) choices = datagen, stats, train, test, createvm, createpsg; Main mode (datagen: data generation, stats: print data statistics, train: train the model, test: test the trained model with new data)
- **createvm_name** (default = None); the name of VM instance to be created
- **createvm_command** (default = None); the command to execute in the VM created
- **createpsg_name** (default = None); start a PSG server VM
- **valid_size** (default = 128); number of planets in the validation set
- **valid_iter** (default = 10); Interval (iterations) for validation and model saving
- **stats_planets** (default = 100); The number of planets to sample from the data set to compute data statistics

- **log_obs_mean** (default = -46.61); Global mean for log observation normailization, applies to training and testing
- **log_obs_stddev** (default = 2.93); Global standard deviation for log observation normalization, applies to training and testing
- **psg_api_key** (default = None); API key for PSG
- **psg_url** (default = None/api); URL for PSG
- **cuda**; Enable CUDA if available
- **train_model_type** (default = FF) choices=[LR, FF, FF2, CNN1, CNN2, CNN3, CNN4]; Type of machine learning model to train (LR: linear regression, LR: feed-forward neural net, CNN1: convolutional neural net, CNN2, CNN3, CNN4)
- **max_blobs** (default = None); Limit the maximum number of blobs that will be read from Google Cloud Storage
- **optimizer** (default = ADAM) choices=[Adam, SGD, ADA, RMS]; Select the optimisation algorithm

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-08-2018	Technical Memorandum	

4. TITLE AND SUBTITLE
NASA Frontier Development Lab: Astrobiology Team II

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Himes, Michael D. and O'Beirne, Molly D. and Soboczenski, Frank and Zorzan, Simone and Baydin, Atılım Güneş and Cobb, Adam and Angerhausen, Daniel and Arney, Giada N. and Domagal-Goldman, Shawn D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA SETI Institute
189 N. Bernardo Ave Suite 200
Mountain View, CA 94043

8. PERFORMING ORGANIZATION REPORT NUMBER
L–

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM–2018–XXXXX

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified-Unlimited
Subject Category 64
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
An electronic version can be found at http://ntrs.nasa.gov.

14. ABSTRACT

Over the past decade, the field of exoplanets has shifted from their detection to the characterization of their atmospheres. Atmospheric retrieval, the inverse modeling technique used to determine an atmosphere's temperature and composition, is both time-consuming and compute-intensive, requiring complex algorithms that generate thousands to millions of atmospheric models, compare the model to the observational data, and build a posterior distribution that gives the most probable value and uncertainty for each model parameter. For rocky, terrestrial planets, the retrieved atmospheric composition can give insight into the surface fluxes of gaseous species necessary to maintain the stability of that atmosphere, which may in turn provide insight into the geological and/or biological processes active on the planet. These atmospheres contain many molecules, some of which are biosignatures, or molecules indicative of biological activity. Runtimes of traditional retrieval models scale with the number of model parameters, so as more and more molecular species are considered, runtimes can become prohibitively long. Machine learning (ML) offers a unique way to reduce the time to perform a retrieval by orders of magnitude, given a sufficient data set to train with. Here we present the Intelligent exoplaNet Atmospheric RetrievAl (INARA) code, the first ML retrieval model for rocky, terrestrial exoplanets, and a data set of 3,000,000 spectra of synthetic rocky exoplanets generated using the NASA Planetary Spectrum Generator.

15. SUBJECT TERMS
FDL,CFD, grid

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Information Desk (help@sti.nasa.gov)
U	U	U	UU	19	19b. TELEPHONE NUMBER *(Include area code)* (757) 864-9658

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18